Partial differential equations and dynamical systems applied to problems coming from physics and biology

Maxime Breden

NNT : 2017SACLN031

THÈSE DE DOCTORAT

de

L'Université Paris-Saclay & l'Université Laval

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Ecole normale supérieure Paris Saclay

Laboratoire d'accueil : Centre de mathématiques et de leurs applications, UMR 8536 CNRS

Spécialité de doctorat : Mathématiques appliquées

Maxime BREDEN

Equations aux dérivées partielles et systèmes dynamiques appliqués à des problèmes issus de la physique et de la biologie

Date de soutenance : 10 juillet 2017

Après avis des rapporteurs :

Philippe LAURENÇOT	(Université de Toulouse)
Michael PLUM	(Karlsruhe Institute of Technology)

Jury de soutenance :

Laurent DESVILLETTES	(Université Paris Diderot)	Codirecteur de thèse
François GOLSE	(École polytechnique)	Président du jury
Robert GUÉNETTE	(Université Laval)	Examinateur
Philippe LAURENÇOT	(Université de Toulouse)	Rapporteur
Jean-Philippe LESSARD	(Univerité Laval)	Codirecteur de thèse
Michael PLUM	(Karlsruhe Institute of Technology)	Rapporteur

Remerciements

Je tiens en premier lieu à exprimer ma profonde gratitude à Laurent Desvillettes et à Jean-Philippe Lessard, tout d'abord pour m'avoir donné le goût de la recherche lors de mes stages en L3/M1, puis pour avoir accepté d'encadrer cette thèse. Merci Laurent pour tes explications toujours simples et précises, même concernant les problèmes les plus ardus, et pour ton exemplaire souci de rigueur. Merci Jean-Philippe pour ton enthousiasme communicatif, et pour avoir été une intarissable source d'idées. Merci à tous les deux pour vos conseils avisés, pour avoir été toujours disponibles tout en me laissant une grande liberté, et pour m'avoir donné autant d'opportunités de voyager et de collaborer avec de nombreux scientifiques. Vous avez été, et continuez à être, d'excellents exemples dans mon développement en tant que chercheur.

Je remercie mes rapporteurs, Philippe Laurençot et Michael Plum (vielen Dank !), pour le temps qu'ils ont consacré à la relecture de ce manuscrit. Merci également à mes examinateurs Robert Guénette et François Golse, d'avoir accepté d'évaluer cette thèse.

Une grande partie des travaux présentés dans cette thèse (ainsi que d'autres toujours en cours) sont les résultats de collaborations avec d'autres chercheurs : Jan Bouwe van den Berg, Roberto Castelli, Klemens Fellner, Maxime Murray, Francesco Salvarani, Ray Sheombarsing et Lennaert van Veen, que je remercie pour toutes ces discussions enrichissantes.

Merci également à tous les doctorants, postdoctorants et autres chercheurs que j'ai pu rencontrer lors de divers workshops ou conférences, et qui ont contribué à rendre ces moments aussi agréables qu'instructifs. Un merci tout particulier à mes frères et soeurs de thèse, et notamment à Ariane, pour ses conseils bienveillants et ses templates en tout genre.

Durant ma thèse, j'ai eu la chance d'effectuer mon monitorat dans le cadre de la préparation à l'agrégation de mathématiques de l'ENS Cachan (qui est devenue l'ENS Paris-Saclay en cours de route). Merci à mes anciens profs de m'avoir fait confiance, et à tous les intervenants de la prépa agrèg avec qui j'ai eu grand plaisir à collaborer. Mention spéciale pour mon acolyte d'analyse Arthur, merci pour toutes ces discussions, que ce soit à propos des polys, des leçons ou des écrits/oraux blancs (et pour avoir adapté le planning en fonction de mes nombreuses contraintes).

Je tiens à adresser un grand merci à Virginie, Véronique et Isabelle, pour leur aide précieuse et efficace vis à vis de mes nombreux ordres de missions (parfois imbriqués les uns dans les autres) et autres tracasseries administratives. Merci également à Agnès, notamment pour avoir fait le relai avec le DED lorsque j'étais à Québec.

Cette thèse n'est pas seulement l'aboutissement de trois années de travail, mais aussi de toute une scolarité influencée par de nombreux enseignants, que je remercie chaleureusement pour m'avoir transmis leurs goûts pour la science et pour l'enseignement. Je tiens en particulier à exprimer toute ma gratitude à Mme Chevallier, pour son implication et son soutien, pendant et après mes années au Schweitzer.

Je n'oublie pas non plus le groupe hand de l'ENS Cachan (pour tous les moments sur et en dehors des terrains), les joueurs de bad de Cachan et de l'UL, les joueurs de soccer du

département de maths de l'UL, ainsi que Jean-Philippe pour les parties de squash, qui m'ont tous permis de décompresser tout au long de ma thèse.

Je tiens également à remercier les *alsaciens* : Max, Flo, Val, Joël, Franz, Adrien... et les *cachanais* : Matthieu, Thibaud, William, Chris, Pierre, Romain, Keurcien, Mika... pour tous les bons moments passés pendant (et avant) cette thèse, je vous attends tous à Munich l'année prochaine ! Mention spéciale pour Thibaud, qui a accepté d'avoir un coloc par alternance à la place d'un chat pendant ces trois ans ;-). Merci également aux *québecois* : Ben, Thomas, Bastien et tout le monde de maths de l'Université Laval, pour m'avoir accueilli lors de mes séjours à Québec.

Enfin, je veux dire un énorme merci aux familles Tora, Breden, Salcedo et Wilhelm, auxquelles je suis très fier d'appartenir, et qui contribuent à faire de tous mes retours en Alsace des moments mémorables. Mon dernier mot sera pour mon frère et mes parents, merci pour le soutien et la confiance que vous m'avez accordés au cours des années (même lorsque j'ai du mal à vous faire comprendre ce que je fais de mes journées), c'est grâce à vous que je suis là aujourd'hui !

Contents

Part I

Introduction

Chapitre 1

Introduction en français

Cette thèse s'inscrit dans le vaste domaine des équations aux dérivées partielles (EDP) et des systèmes dynamiques, et est constituée de deux parties indépendantes (la partie II et la partie III), qui sont introduites séparément dans les sections 1.1 et 1.2.

Dans la partie II, on s'intéresse à une classe d'EDP appelées les équations de coagulation-fragmentation discrètes avec diffusion. On établit pour ces équations des propriétés générales d'existence et de régularité des solutions, ainsi que certaines propriétés qualitatives (comme la conservation de la masse).

Dans la partie III, on établit également des résultats d'existence pour certaines EDP et équations différentielles ordinaires, mais avec une approche totalement différente. En effet, au lieu d'obtenir des résultats généraux on se concentre sur des équations précises, avec des paramètres donnés, et on étudie des solutions spécifiques comme des états stationnaires ou des connexions homoclines, à l'aide de méthodes assistées par ordinateur.

1.1 Introduction à la partie II : estimations de moments pour les équations de coagulation-fragmentation discrètes avec diffusion

Dans cette section, on introduit les résultats obtenus dans la partie II de cette thèse, où on s'intéresse aux équations de coagulation-fragmentation discrètes avec diffusion. On commence par présenter brièvement les processus de coalescence et de fragmentation, ainsi que leur description mathématique dans la section 1.1.1. Dans la section 1.1.2, on décrit plus précisément le type de modèle étudié dans cette thèse. Dans la section 1.1.3, on discute ensuite des problématiques liées à la conservation de la masse (ou à sa non conservation, c'est à dire le phénomène de gélification), qui sont cruciales pour les équations de coagulation-fragmentation. Dans la section 1.1.4, on explique en quoi les estimations sur les moments peuvent être utiles pour étudier ces équations, et on présente dans la section 1.1.5 l'outil central qu'on utilise pour établir de telles estimations. On poursuit en présentant dans la section 1.1.6 les principaux résultats obtenus dans la partie II de cette thèse, et on conclut cette introduction par la section 1.1.7, où on décrit quelques problèmes et questions ouvertes qui se trouvent naturellement dans la continuité de cette thèse.

1.1.1 A propos des processus de coalescence et de fragmentation

La coalescence et la fragmentation sont des processus qui jouent un rôle très important en physique, ainsi qu'en chimie et en biologie, pour décrire des phénomènes où des particules s'agglomèrent pour former de plus grosses particules, ou se disloquent en plusieurs particules plus petites. On peut par exemple penser à l'hématologie, et plus précisément aux mécanismes de la

coagulation sanguine. Un autre exemple serait la formation de gouttelettes dans un spray. A une toute autre échelle, ces processus sont aussi utilisés pour expliquer les formations de galaxies. Pour une revue plus détaillée, on renvoie au travaux de synthèse [18, 30] et aux références mentionnées dans ces articles.

Du point de vue mathématique, les processus de coalescence et de fragmentation peuvent être étudiés à trois niveaux différents.

- La description *microscopique* considère un ensemble fini de particules qui interagissent de manière stochastique. Ce type de modèle a pour origine les travaux de Smoluchowski [42, 43]. Dans le même registre, on mentionne le processus de Marcus-Lushnikov [31, 32].
- Lorsque le nombre de particules est suffisamment élevé, on peut considérer une description *mésoscopique*, où au lieu de suivre chaque particule individuellement, on utilise une représentation statistique du système. Autrement dit, on classifie les particules, par exemple par leur taille, et on s'intéresse à l'évolution du nombre (ou de la densité) de particules de chaque taille. Cette évolution est alors décrite par des équations déterministes, comme les *équations de Smoluchowski* (également introduites dans [42, 43]), qui sont étudiées dans cette thèse.
- La troisième échelle correspond à la description *macroscopique*, où on se concentre sur les quantités physiques observables (comme la masse totale), qui sont souvent obtenues comme des moyennes de variables mésoscopiques.

On mentionne qu'il existe des liens entre ces différentes descriptions. En particulier, les modèles mésoscopiques peuvent parfois être obtenus rigoureusement à partir des modèles microscopiques, via la convergence de processus stochastiques ou des limites de champs moyens (voir par exemple [35], pour la convergence du processus de Marcus-Lushnikov vers les équations de Smoluchowski). Pour une connexion entre les descriptions mésoscopiques et macroscopiques, via des limites de réactions rapides, on renvoie à [13, 19].

1.1.2 Différents modèles mésoscopiques

Dans la partie II de cette thèse, on étudie la description mésoscopique donnée par $c = (c_i)_{i\in\mathbb{N}^*}$, où c_i représente la concentration de particules de taille i, pour chaque i dans $\mathbb{N}^*$. On suppose que le processus d'agrégation suit la *loi d'action de masse*, qui stipule que la vitesse d'une réaction chimique est proportionnelle au produit des concentrations des réactifs. Ainsi, le terme qui décrit l'agglomération d'une particule de taille i avec une particule de taille j, pour former une particule de taille $i+j$, est de la forme $a_{i,j}c_ic_j$, où $a_{i,j}$ est un coefficient de proportionnalité. Pour le processus de fragmentation, on suppose que la dislocation d'une particule n'est pas influencée par les autres particules mais dépend uniquement de sa propre taille. On considère un autre coefficient de proportionnalité B_i qui représente la probabilité de fragmentation d'une particule de taille i par unité de temps. Enfin, pour décrire le résultat d'une fragmentation, on introduit des coefficients $\beta_{i,j}$ qui représentent le nombre moyen de particules de taille $j < i$ produites lors de la dislocation d'une particule de taille i. Ces processus de coagulation et de fragmentation sont schématisés sur la figure 1.1. Il est naturel de supposer que la masse est conservée lors d'une fragmentation, que les particules de taille 1 ne se fragmentent pas davantage, et que les taux de coagulation sont symétriques, ce qui se traduit mathématiquement par les hypothèses suivantes

$$i = \sum_{j=1}^{i-1} j\beta_{i,j}, \quad B_1 = 0, \quad a_{i,j} = a_{j,i} \quad \text{et} \quad a_{i,j}, B_i, \beta_{i,j} \geq 0, \quad \forall\, i,j \in \mathbb{N}^*. \tag{1.1}$$

Enfin, on veut prendre en compte les variations des concentrations c_i en fonction de la position (représentée par la variable x). Chaque $c_i = c_i(t,x)$ est donc une fonction du temps et de l'espace, et on suppose que les particules de taille i diffusent avec un coefficient d_i.

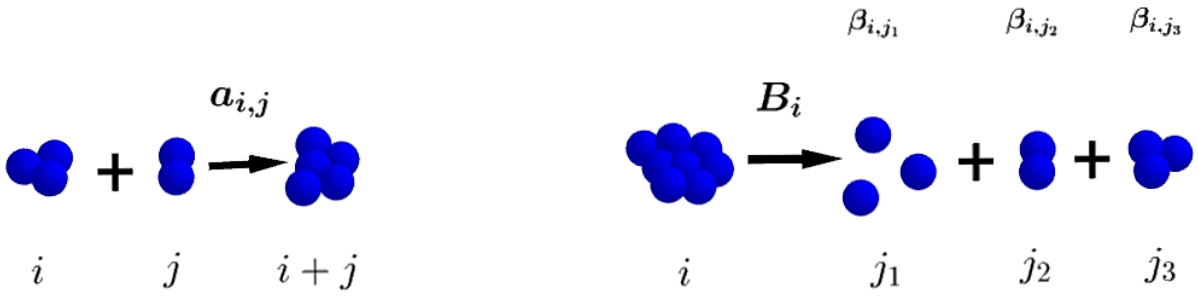


FIGURE 1.1 – Un exemple de coagulation et de fragmentation. Ici $\beta_{8,1} = 3$, $\beta_{8,2} = 1$, $\beta_{8,3} = 1$ et $\beta_{i,j} = 0$ pour $4 \leq j \leq 7$.

Considérons Ω un domaine borné de $\mathbb{R}^N$, dans lequel les particules sont supposées être confinées, et supposons données des concentrations initiales $(c_i^{in})_{i\in\mathbb{N}^*}$. Les mécanismes de coagulation-fragmentation avec diffusion décrits ci-dessus sont représentés par les équations suivantes

$$\begin{cases} \partial_t c_i - d_i \Delta_x c_i = Q_i(c) + F_i(c), & \text{on } [0,T]\times\Omega, \\ \nabla_x c_i \cdot \nu = 0 & \text{sur } [0,T]\times\partial\Omega, \qquad \forall i \in \mathbb{N}^* \\ c_i(0,\cdot) = c_i^{in} & \text{sur } \Omega, \end{cases} \tag{1.2}$$

avec $\nu(x)$ le vecteur unitaire normal *sortant* au point $x \in \partial\Omega$, et où les termes de coagulation $Q_i(c)$ [1] [2] et de fragmentation $F_i(c)$ sont donnés par :

$$Q_i(c) = Q_i^+(c) - Q_i^-(c) = \frac{1}{2}\sum_{j=1}^{i-1} a_{i-j,j} c_{i-j} c_j - \sum_{j=1}^{\infty} a_{i,j} c_i c_j, \tag{1.3}$$

$$F_i(c) = F_i^+(c) - F_i^-(c) = \sum_{j=1}^{\infty} B_{i+j}\beta_{i+j,i} c_{i+j} - B_i c_i. \tag{1.4}$$

Les conditions aux limites de Neumann homogènes sont naturelles pour modéliser qu'aucune particule n'entre ou ne sort du domaine Ω.

Avant d'aller plus loin, on passe en revue différentes variantes de ce modèle. La version qui a été la plus étudiée est sans doute celle où on on se restreint à des fragmentations binaires, c'est à dire qu'on suppose que lors d'une fragmentation, une particule se sépare en exactement deux particules plus petites. En notant $b_{i,j}$ la probabilité de fragmentation d'une particule de taille $i+j$ en une particule de taille i et une autre de taille j, le terme de fragmentation devient [3]

$$F_i(c) = F_i^+(c) - F_i^-(c) = \sum_{j=1}^{\infty} b_{i,j} c_{i+j} - \frac{1}{2}\sum_{j=1}^{i-1} b_{i-j,j} c_i.$$

On remarque qu'on peut se ramener à la formulation (1.4) en considérant des coefficients B_i et $\beta_{i,j}$ définis par

$$B_i = \frac{1}{2}\sum_{j=1}^{i-1} b_{i-j,j}, \quad \beta_{i,j} = \frac{b_{i-j,j}}{B_i}, \quad \forall\ 1 \leq j < i.$$

1. Le facteur $\frac{1}{2}$ dans Q_i^+ vient du fait que les réactions d'agglomération sont comptées deux fois, par exemple on a $1 + (i-1) \to i$, mais aussi $(i-1) + 1 \to i$. La seule exception est la réaction $\frac{i}{2} + \frac{i}{2} \to i$ (pour i pair) qui n'apparait qu'une fois. Pour que le terme (1.3) décrive fidèlement le processus d'agglomération, il faut donc prendre pour les coefficients $a_{i,i}$ deux fois leur valeur *physique*.

2. On notera également que ce terme ne prend pas en compte d'éventuelles agglomérations simultanées de trois (ou plus de trois) particules. Cette omission est justifiée par le fait que de telles agglomérations ont une probabilité beaucoup plus faible de se produire que des agglomérations entre deux particules, puisque qu'une agglomération nécessite que toutes les particules concernées soient au même endroit au même moment.

3. avec la même remarque concernant la définition des coefficients $b_{i,i}$ que pour les coefficients $a_{i,i}$.

Une situation encore plus spécifique peut être envisagée, en ne considérant que les coagulations de la forme $1 + (i-1) \to i$ et les fragmentations de la forme $i \to 1 + (i-1)$. Ce modèle a été introduit par Becker et Döring [5] pour modéliser des phénomènes de nucléation et a été amplemant étudié depuis (voir l'article de synthèse [41]).

Un autre modèle consiste à supposer que la fragmentation d'une particule n'est pas *auto-induite* mais résulte d'une collision avec une autre particule. Dans ce cas, on obtient un terme de la forme $B_{j,k}c_jc_k$ pour décrire la fragmentation d'une particule de taille j et d'une autre de taille k (due à leur collision), et on introduit le coefficient $\beta_{j,k,i}$ pour décrire le nombre moyen de particules de taille i produites par une telle fragmentation. On a alors

$$F_i(c) = F_i^+(c) - F_i^-(c) = \frac{1}{2}\sum_{j+k>i} B_{j,k}\beta_{j,k,i}c_jc_k - \sum_{j=1}^{\infty} B_{i,j}c_ic_j.$$

Dans une direction différente, on peut considérer à la place d'une description discrète $c = (c_i)_{i\in\mathbb{N}^*}$, une description continue $c = c(y)$ pour $y \in \mathbb{R}_+$. Ce modèle a été introduit par Müller [34]. Le choix d'utiliser une description discrète ou continue dépend de l'application considérée. Pour plus de détails concernant les modèles de coagulation-fragmentation continus, on renvoie encore à [18, 30]. On mentionne également que les liens entre les modèles discrets et continus on été étudiés (voir [29] et les références mentionnées dans cet article).

1.1.3 Conservation de la masse et gélification

On se concentre à présent sur les équations de coagulation-fragmentation discrètes avec diffusion (1.1)-(1.4) [4] et on s'intéresse à certaines propriétés des solutions, en particulier l'évolution de la masse totale, donnée par $\sum_{i=1}^{\infty} ic_i$ [5]. Il va être utile de considérer une formulation faible des termes de coagulation et de fragmentation : pour toute suite $(\varphi_i)_{i\in\mathbb{N}^*}$, en utilisant (1.1) on obtient (au moins formellement)

$$\sum_{i=1}^{\infty}\varphi_iQ_i(c) = \frac{1}{2}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{i,j}c_ic_j(\varphi_{i+j} - \varphi_i - \varphi_j), \tag{1.5}$$

$$\sum_{i=1}^{\infty}\varphi_iF_i(c) = -\sum_{i=2}^{\infty} B_ic_i\left(\varphi_i - \sum_{j=1}^{i-1}\beta_{i,j}\varphi_j\right). \tag{1.6}$$

En prenant $\varphi_i = i$ pour tout $i \in \mathbb{N}^*$, on remarque qu'on a

$$\sum_{i=1}^{\infty} iQ_i(c) = 0 = \sum_{i=1}^{\infty} iF_i(c).$$

En multipliant (1.2) par i et en sommant pour tout $i \in \mathbb{N}^*$, on obtient donc (toujours formellement)

$$\partial_t\left(\sum_{i=1}^{\infty} ic_i\right) - \Delta_x\left(\sum_{i=1}^{\infty} id_ic_i\right) = 0. \tag{1.7}$$

4. Cette restriction est faite principalement pour rendre le propos plus clair, dans la mesure où les résultats de cette thèses concernent presque exclusivement ce modèle précis. Néanmoins les considérations faites dans cette section s'appliquent également aux autres modèles mentionnés dans la section 1.1.2.

5. En toute rigueur, puisque que chaque c_i représente la concentration de particules de tailles i, $\sum_{i=1}^{\infty} ic_i$ désigne le nombre total de particules *élémentaires* (i.e. de taille 1) par unité de volume (une particule de taille i étant constitué de i particules de taille 1). La quantité $\sum_{i=1}^{\infty} ic_i$ n'est donc pas égale mais proportionnelle à la masse totale, le facteur de proportionnalité étant égal à la masse d'une particule élémentaire divisée par le volume total. Par abus de langage, on continuera à parler de masse pour désigner $\sum_{i=1}^{\infty} ic_i$.

Après une intégration sur Ω, et en utilisant les conditions aux limites de Neumann, il reste

$$\frac{d}{dt}\int_{\Omega}\sum_{i=1}^{\infty} ic_i = 0,$$

ce qui signifie que la masse totale devrait rester constante. Cependant, selon les valeurs des coefficients de coagulation et de fragmentation (i.e. les coefficients $a_{i,j}$, B_i et $\beta_{i,j}$), on peut avoir des situations où la masse totale devient strictement inférieure à la masse initiale après un temps fini. Ce phénomène n'est pas un artéfact mathématique, il peut être observé et expliqué d'un point de vue physique. Cette situation correspond à un changement de phase dans le système, la masse perdue étant transférée dans la phase nouvellement crée. Elle intervient notamment lors de la formation de gels colloïdaux, ce qui explique probablement pourquoi ce phénomène est maintenant désigné par le terme *gélification*. Mathématiquement, la gélification intervient lorsqu'une partie de la masse s'échappe en $i \to \infty$, ce qu'on peut interpréter comme une formation de particules de taille infinie.

Pour mieux cerner ce phénomène, on va considérer brièvement le cas particulier où $a_{i,j} = ij$ et où il n'y a pas de fragmentation ($B_i = 0$). On néglige également la diffusion ($d_i = 0$), pour pouvoir considérer des solutions dépendant uniquement de t mais plus de x. On introduit les moments

$$\rho_k = \sum_{i=1}^{\infty} i^k c_i,$$

et on utiliser la formulation faible (1.5) avec $\varphi_i = 1$, pour aboutir à

$$\frac{d}{dt}\rho_0 = \frac{d}{dt}\left(\sum_{i=1}^{\infty} c_i\right) = -\frac{1}{2}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} ijc_ic_j = -\frac{1}{2}\left(\rho_1\right)^2. \tag{1.8}$$

On remarquera que ρ_0 et ρ_1 représentent respectivement le nombre total de particules et la masse totale [6], et sont des quantités qui doivent rester positives. En intégrant (1.8) entre 0 et t, on obtient

$$\rho_0(t) + \frac{1}{2}\int_0^t \left(\rho_1(s)\right)^2 ds = \rho_0(0).$$

En particulier, on a

$$\frac{1}{2}\int_0^t \left(\rho_1(s)\right)^2 ds \le \rho_0(0), \quad \forall\ t \ge 0,$$

ce qui impose l'existence d'un temps t^* positif pour lequel la masse $\rho_1(t^*)$ est strictement inférieure à la masse initiale $\rho_1(0)$ [7]. Le premier instant pour lequel cette diminution de masse se produit est appelé le *temps de gélification*. Puisque la gélification correspond à la création de particules de taille infinie, elle devrait se traduire par une explosion des moments d'ordres supérieurs. En utilisant à nouveau la formulation faible (1.5), cette fois avec $\varphi_i = i^2$, il vient

$$\frac{d}{dt}\rho_2 = \left(\rho_2\right)^2,$$

et on a bien une explosion du second moment ρ_2 lorsque t tend vers $t_2 = \frac{1}{\rho_2(0)}$. Pour cet exemple, on peut même prouver que t_2 est le temps de gélification (pour plus de détails, et des considérations sur les *profils d'explosion*, on revoie le lecteur aux articles de synthèse [16, 30] et à leurs références).

La situation est radicalement différente si on suppose $a_{i,j} = 1$ (toujours avec $B_i = 0$, $F_i = 0$ et $d_i = 0$). Dans ce cas on a

$$\frac{d}{dt}\rho_0 = -\frac{1}{2}\left(\rho_0\right)^2,$$

6. toujours à un facteur multiplicatif près.
7. si on suppose que la masse initiale est non nulle.

ce qui, contrairement à (1.8), n'impose pas de restriction sur la mass ρ_1. On a également

$$\frac{d}{dt}\rho_2 = (\rho_1)^2 ,$$

et on n'a donc pas d'explosion pour le second moment. Pour cet exemple, on peut en fait montrer que la masse totale reste conservée pour tout $t \geq 0$.

Ces deux cas particuliers suggèrent que la gélification se produit si les particules s'agglomèrent de plus en plus vite en grossissant, c'est à dire si les coefficients de coagulation $a_{i,j}$ croissent assez rapidement avec i et j. Cette heuristique a été démontrée, le cas limite en dessous duquel la gélification ne peut pas se produire étant celui des *coefficients de coagulation linéaires*, donné par $a_{i,i} = C(i + j)$. On renvoie aux introductions des chapitres 4 et 5 pour des énoncés mathématiques précis et des références sur la question. Pour l'instant, on n'a considéré que le processus d'agglomération, mais il est plausible que même si les agglomérations s'effectuent rapidement, elles puissent être contrebalancées par le processus de fragmentation. On peut ainsi imaginer que si les grosses particules se fragmentent aussi vite qu'elles sont crées, on puisse éviter le phénomène de gélification même si la coagulation est forte. Ce principe a aussi été démontré mathématiquement (on renvoie toujours à l'introduction du Chapitre 5 pour des énoncés précis et des références), mais jusqu'à présent seulement dans le cas homogène (c'est à dire lorsqu'il n'y a pas de diffusion).

1.1.4 Pourquoi des estimations sur les moments ?

Comme le suggère le titre de la partie II, le segment de cette thèse consacré aux équations de coagulation-fragmentation est basé sur des estimations de moments. On va donner ici deux raisons pour lesquelles de telles estimations sont cruciales pour étudier les équations de coagulation-fragmentation.

La première raison est que, comme suggéré par les calculs de la section précédente, l'évolution des moments ρ_k est liée à la problématique de la gélification. Plus précisément, si on dispose d'estimations adaptées pour un moment *sur-linéaire* (i.e. un moment ρ_k avec $k > 1$) on peut prouver que la gélification ne peut pas se produire. La justification de cette affirmation est en fait reliée aux techniques utilisées pour démonter l'existence de solutions pour les équations de coagulation-fragmentation avec diffusion, on en profite donc pour décrire brièvement ces méthodes maintenant.

Il existent plusieurs approches légèrement distinctes pour prouver l'existence de solutions pour les équations de coagulation-fragmentation avec diffusion, mais elles sont toutes basées sur l'utilisation d'une version tronquée du système (1.2), où on considère uniquement les particules de tailles inférieures à un certain n. Ce système tronqué a donc un nombre fini d'équations et d'inconnues, et les termes de réaction donnés par $Q_i(c) + F_i(c)$ ne contiennent plus que des sommes finies (puisque on suppose $c_i = 0$ pour $i > n$). On peut alors utiliser des résultats classiques d'existence locale (en temps) pour les systèmes de réaction-diffusion, et les combiner avec des estimations utilisant la structure précise du système pour obtenir des solutions globales (cette procédure est détaillée dans [45]). L'étape suivante consiste à considérer une suite de solutions des systèmes tronqués avec n tendant vers l'infini. On utilise alors un argument de compacité pour extraire une sous suite convergente (c'est à ce niveau que les différentes méthodes diffèrent, en fonction des estimations *a priori* dont on dispose), et on conclut en montrant que la limite ainsi obtenue est bien une solution (faible) des équations de coagulation-fragmentation avec diffusion (1.2).

Le lien avec la gélification provient du fait que, pour une solution du système tronqué, on peut faire des calculs similaires à ceux effectuées au début de la section 1.1.3, mais cette fois rigoureusement. On peut ainsi prouver que pour une solution du système tronqué la masse reste bien égale à la masse initiale pour tout temps t, et il suffit de pouvoir passer à la limite $n \to \infty$ dans cette égalité pour montrer que la masse est conservée pour la solution du système

complet. Pour justifier ce passage à la limite, qui concerne le moment ρ_1, il suffit d'avoir une borne (indépendante du niveau de troncature n) pour un moment ρ_k d'ordre $k > 1$ dans un espace fonctionnel approprié (cette procédure est détaillée dans le chapitre 5).

La seconde raison pour laquelle les estimations sur les moments sont importantes, particulièrement pour les équations de coagulation-fragmentation *avec* diffusion, est liée à la régularité des solutions. Comme on vient de le faire remarquer, la théorie usuelle d'existence pour ce type d'équations ne fournit que des solutions faibles (dont une définition précise, suivant l'approche de Laurençot et Mischler [28], est donnée dans les chapitres 4 et 5). Néanmoins, chaque concentration c_i est solution d'une équation de type chaleur (1.2), ce qui suggère qu'on pourrait en fait obtenir des solutions régulières. La principale difficulté réside dans l'obtention d'estimations pour les second membres $Q_i(c)+F_i(c)$ de (1.2), qui contiennent des sommes infinies. Cependant, comme on l'a vu sur deux exemples dans la section 1.1.3, ces termes peuvent être exprimés en fonction des moments des solutions (ou au moins être majorés par des expressions faisant intervenir les moments, pour des coefficients de coagulation et de fragmentation plus généraux). Des estimations sur les moments pourraient donc permettre de contrôler ces termes et d'en déduire des résultats de régularité pour les équations de coagulation-fragmentation avec diffusion.

1.1.5 Le rôle des lemmes de dualité

Dans la section 1.1.3, on a obtenu des équations satisfaites par les moments ρ_0, ρ_1 et ρ_2 dans le cas de solutions homogènes (c'est à dire en supposant que les coefficients de diffusion d_i sont égaux à 0). Dans la suite on veut être capable de traiter le modèle plus réaliste qui prend en compte une possible hétérogénéité spatiale et où la diffusion joue un rôle. On doit donc adapter les calculs fait dans la section 1.1.3.

Lorsque les coefficients de diffusion d_i ne sont plus supposés nuls, on rappelle qu'on a toujours l'identité (1.7) sur la masse ρ_1. Malheureusement, à moins que les d_i soient tous égaux, (1.7) n'est pas à proprement parler une équation parabolique sur ρ_1, puisque ρ_1 n'apparait pas explicitement dans le laplacien. Cependant, on peut réécrire (1.7) sous la forme

$$\partial_t \rho_1 - \Delta_x \left(M_1 \rho_1 \right) = 0,$$

avec

$$M_1 = \frac{\sum_{i=1}^{\infty} i d_i c_i}{\sum_{i=1}^{\infty} i c_i}.$$

De manière similaire, toujours avec des coefficients de diffusion d_i non nuls mais avec $a_{i,j} = 1$ et $B_i = 0$, les calculs de la section 1.1.3 sur le second moment donnent

$$\partial_t \rho_2 - \Delta_x \left(\sum_{i=1}^{\infty} i^2 d_i c_i \right) = \left(\rho_1 \right)^2,$$

ce que l'on peut réécrire en

$$\partial_t \rho_2 - \Delta_x \left(M_2 \rho_2 \right) = \left(\rho_1 \right)^2,$$

avec

$$M_2 = \frac{\sum_{i=1}^{\infty} i^2 d_i c_i}{\sum_{i=1}^{\infty} i^2 c_i}.$$

Pour M_1 et M_2, la seule estimation naturelle disponible est une borne L^∞ de la forme

$$\inf_{i \in \mathbb{N}^*} d_i \leq M_1, M_2 \leq \sup_{i \in \mathbb{N}^*} d_i.$$

La question qui apparait alors est la suivante : quel genre d'estimations peut-on obtenir pour une fonction positive u satisfaisant

$$\begin{cases} \partial_t u - \Delta_x (M u) \leq f & \text{sur } [0,T] \times \Omega, \\ \nabla_x u \cdot n = 0 & \text{sur } [0,T] \times \partial\Omega, \end{cases} \tag{1.9}$$

où Ω est un domaine borné et régulier, f est dans $L^p([0,T]\times\Omega)$ pour un certain $p\in]1,\infty[$ et M est une fonction mesurable vérifiant $a\le M\le b$?

On peut apporter une réponse à cette question en utilisant une technique appelée *lemme de dualité* dans la litérature, qui est attribuée à Pierre et Schmitt [38], et a ensuite été généralisée par Cañizo, Desvillettes et Fellner [11]. On donne ici un aperçu de ces techniques pour en faire ressortir l'idée générale, différents versions adaptées pour des cas spécifiques sont présentées en détail dans le chapitre 4.

La stratégie consiste à considérer une sorte de *problème dual*, défini par

$$\begin{cases} \partial_t v + M\Delta_x v = -\psi & \text{sur } [0,T]\times\Omega, \\ \nabla_x v\cdot n = 0 & \text{sur } [0,T]\times\partial\Omega \\ v(T,\cdot)=0 & \text{sur } \Omega, \end{cases} \tag{1.10}$$

où ψ est une fonction test positive. On multiple l'équation de (1.9) par v (qui est positif par le principe du maximum), et on intègre, sur Ω puis sur $[0,T]$, pour obtenir

$$\begin{aligned} v\partial_t u - v\Delta_x(Mu) &\le fv \\ \partial_t(uv) - u\partial_t v - v\Delta_x(Mu) &\le fv \\ \frac{d}{dt}\int_\Omega uv - \int_\Omega \left(u\left(\partial_t v + M\Delta_x v\right)\right) &\le \int_\Omega fv \\ \int_0^T\int_\Omega u\psi &\le \int_\Omega u(0)v(0) + \int_0^T\int_\Omega fv. \end{aligned}$$

Supposons momentanément que la solution v de (1.10) vérifie une estimation de la forme

$$\|v(0)\|_{L^{p'}(\Omega)} \le C\|\psi\|_{L^{p'}([0,T]\times\Omega)} \quad \text{et} \quad \|v\|_{L^{p'}([0,T]\times\Omega)} \le C\|\psi\|_{L^{p'}([0,T]\times\Omega)}, \tag{1.11}$$

où, ici et dans la suite, p' désigne l'exposant dual de p (i.e. tel que $\frac{1}{p}+\frac{1}{p'}=1$). En utilisant l'inégalité de Hölder, on aboutit alors à

$$\int_0^T\int_\Omega u\psi \le C\left(\|u(0)\|_{L^p(\Omega)} + \|f\|_{L^p([0,T]\times\Omega)}\right)\|\psi\|_{L^{p'}([0,T]\times\Omega)},$$

et puisque cette inégalité est valable pour toute fonction test positive ψ, on a par dualité que

$$\|u\|_{L^p([0,T]\times\Omega)} \le C\left(\|u(0)\|_{L^p(\Omega)} + \|f\|_{L^p([0,T]\times\Omega)}\right).$$

Il reste à voir sous quelles conditions on peut obtenir les estimations (1.11) pour v.

Il est assez aisé de monter (en commençant par multiplier (1.10) par $\Delta_x v$ et en intégrant sur Ω), que (1.11) est vérifié pour $p=p'=2$, dès que $0<a\le b<\infty$, ce qui donne une estimation L^2 pour u. Cette estimation correspond à la version originale du lemme de dualité de Pierre et Schmitt. Cañizo, Desvillettes et Fellner ont ensuite montré que (1.11) était aussi vérifié pour $p'\in]1,\infty[$, si en plus d'avoir $0<a\le b<\infty$ on supposait que a et b étaient *suffisamment proches* l'un de l'autre. Pour caractériser précisément ce *suffisamment proche*, on commence par introduire une notation.

Definition 1.1.1. *Pour $m>0$ et $q\in]1,+\infty[$, on définit $\mathcal{K}_{m,q}>0$ comme la meilleure (i.e. la plus petite) constante indépendante de $T>0$ dans l'estimation de régularité parabolique*

$$\left(\int_0^T\int_\Omega |\partial_t v|^q + m^q\int_0^T\int_\Omega |\Delta_x v|^q\right)^{\frac{1}{q}} \le \mathcal{K}_{m,q}\left(\int_0^T\int_\Omega |f|^q\right)^{\frac{1}{q}}, \quad \forall\ f\in L^q([0,T]\times\Omega),$$

où v est l'unique solution de l'équation de la chaleur avec diffusion constante m, avec conditions aux limites de Neumann homogènes et 0 comme valeur initiale :

$$\begin{cases} \partial_t v - m\Delta_x v = f & \text{sur } [0,T]\times\Omega, \\ \nabla_x v\cdot\nu = 0 & \text{sur } [0,T]\times\partial\Omega, \\ v(0,\cdot) = 0 & \text{sur } \Omega. \end{cases}$$

L'existence d'une telle constante $\mathcal{K}_{m,q} < \infty$ indépendante du temps $T > 0$ est explictement énoncée dans [27], sous réserve que $\partial\Omega \in \mathcal{C}^{2+\alpha}$, $\alpha > 0$. Si $0 < a \le b < \infty$ et

$$\frac{b-a}{b+a}\,\mathcal{K}_{\frac{a+b}{2},p'} < 1, \tag{1.12}$$

il est alors prouvé dans [11] que (1.11) est vérifié, et on obtient donc une estimation L^p pour u. On remarquera que ce résultat contient le cas particulier $p = p' = 2$ mentionné précédemment, car on peut montrer que $\mathcal{K}_{m,2} \le 1$, d'où l'on déduit que (1.12) est toujours vérifiée pour $p' = 2$.

La technique décrite dans cette section est utilisée à maintes reprises dans cette thèse, pour obtenir des estimations L^p pour les moments ρ_k associées aux solutions des équations de coagulation-fragmentation.

1.1.6 Exposition des résultats obtenus dans les chapitres 4 et 5

Dans cette partie, on présente les principaux résultats obtenus dans la partie II de cette thèse.

Le premier résultat, auquel on a déjà fait allusion dans les sections précédentes, est une estimation L^p pour tout $p \in]1,\infty[$, pour la masse associée à une solutions des équations de coagualation-fragmentation avec diffusion (1.1)-(1.4). Cette estimation généralise un résultat de [10] où une estimation L^2 a été obtenue (en utilisant la version originale du lemme de dualité).

Proposition 1.1.2. *Soient Ω un domaine borné et régulier de $\mathbb{R}^N$ et $T > 0$. On suppose que les coefficients de coagulation et de fragmentation vérifient*

$$\lim_{j\to\infty}\frac{a_{i,j}}{j} = 0 = \lim_{j\to\infty}\frac{B_{i+j}\beta_{i+j,i}}{i+j}, \quad \forall\, i\in\mathbb{N}^*, \tag{1.13}$$

et que les coefficients de diffusion vérifient

$$0 < a := \inf_{i\ge1} d_i, \qquad \text{et} \qquad b := \sup_{i\ge1} d_i < \infty. \tag{1.14}$$

Soit $p \in]1,+\infty[$. On suppose également que les données initiales $c_i^{in} \ge 0$ sont telles que la masse initiale ρ_1^{in} est dans $L^p(\Omega)$, et que l'hypothèse (1.12) *est vérifiée.*

Alors, il existe une solution faible des équations de coagulation-fragmentation (1.1)-(1.4) *pour laquelle la masse ρ_1 est dans $L^p([0,T]\times\Omega)$.*

On note ici que l'hypothèse (1.13) est uniquement utilisée pour obtenir l'existence de solutions faibles (qui est ici garantie par les travaux de Laurençot et Mischler [28]).

Si les coefficients de coagulation sont *sous-linéaires*, on peut obtenir un résultat similaire pour les moments d'ordre supérieur.

Theorem 1.1.3. *Soient Ω un domaine borné et régulier de $\mathbb{R}^N$, $N \le 2$, et $T > 0$. On suppose que les coefficients de coagulation et de fragmentation vérifient*

$$a_{i,j} \le C_Q\left(i^\alpha j^\beta + i^\beta j^\alpha\right) \quad \text{et} \quad \lim_{j\to\infty}\frac{B_{i+j}\beta_{i+j,i}}{i+j} = 0, \quad \forall\, i\in\mathbb{N}^*,$$

où $C_Q > 0$, $\alpha, \beta \geq 0$ et $\alpha + \beta < 1$, et que les coefficients de diffusions vérifient

$$d_i > 0, \ \forall \ i \in \mathbb{N}^* \quad et \quad d_i \underset{i\to\infty}{\longrightarrow} d_\infty > 0. \tag{1.15}$$

On suppose également que les concentrations initiales $c_i^{in} \geq 0$ sont toutes dans $L^\infty(\Omega)$ et sont telles que, pour un certain $k \in \mathbb{N}^*$ le moment initial ρ_k^{in} est dans $L^p(\Omega)$ pour tout $p \in]1, \infty[$.

Alors, il existe une solution faible des équations de coagulation-fragmentation (1.1)-(1.4) *pour laquelle le moment ρ_k est dans $L^p([0,T] \times \Omega)$ pour tout $p \in]1, \infty[$.*

Remark 1.1.4. *Les différentes hypothèses de ce théorème sont commentées en détail dans les chapitres 4 et 5. On mentionne ici qu'on peut supprimer la restriction $N \leq 2$ s'il n'y a pas de fragmentation (i.e. si $B_i = 0$ pour tout $i \in \mathbb{N}$), et surtout que l'hypothèse* (1.15) *joue un rôle crucial, dans la mesure où elle permet à* (1.12) *d'être vérifiée pour tout $p \in]1, \infty[$, si on considère seulement les concentrations c_i pour i assez grand. Cette hypothèse peut sembler restrictive, mais elle n'est en fait pas beaucoup plus forte que de simplement imposer $\inf_{i\in\mathbb{N}^*} d_i > 0$* [8]*, ce qui est fait dans plusieurs autres travaux. On revient sur ce point dans la section 1.1.7.*

Comme on l'a expliqué dans la section 1.1.4, le théoreme 1.1.3 implique qu'il ne peut pas y avoir de gélification pour des coefficients de coagulation souslinéares. Ce résultat est connu depuis les travaux de Ball et Carr [3] dans le cas homogène, et a été étendu plus récemment au modèle incluant la diffusion par Cañizo, Desvillettes et Fellner dans [10]. Dans cet article, les estimations obtenues sont plus faibles que celles présentées dans le théorème 1.1.3 (on y démontre seulement une estimation L^1 pour un moment légèrement sur-linéaire), mais déjà suffisantes pour empecher la gélification. Au sujet des estimations sur les moments pour les équations de coagulation-fragmentation avec diffusion, on mentionne également les travaux de Rezakhanlou [39, 40].

On donne plus tard (voir théorème 1.1.6) une autre application du théorème 1.1.3 reliée à la régularité des solutions. Mais tout d'abord, on présente une généralisation du théorème 1.1.3 où la coagulation peut être sur-linéaire, mais compensée par une fragmentation suffisamment forte.

Theorem 1.1.5. *Soient Ω un domaine borné et régulier de $\mathbb{R}^N$, $N \leq 2$, et $T > 0$. On suppose que les coefficients de coagulation et de fragmentation vérifient*

$$a_{i,j} \leq C_Q \left(i^\alpha j^\beta + i^\beta j^\alpha \right) \quad et \quad B_i \geq C_F i^\gamma,$$

où $C_Q, C_F > 0$, $0 \leq \alpha, \beta \leq 1$, $\gamma \geq 1$ et $\alpha + \beta < \gamma$, ainsi que

$$\sup_{j\in\mathbb{N}^*} \frac{a_{i,j}}{j} < \infty \quad et \quad \sup_{j\in\mathbb{N}^*} \frac{B_{i+j}\beta_{i+j,i}}{i+j} < \infty, \quad \forall \ i \in \mathbb{N}^*. \tag{1.16}$$

On suppose également que les coefficients de diffusion vérifient (1.15)*. Enfin, on suppose que les concentrations initiales $c_i^{in} \geq 0$ sont dans $L^\infty(\Omega)$ pour tout $i \geq 1$, que pour un certain $k > 1$ vérifiant $k > 2 - (\gamma - \alpha)$ et $k > 2 - (\gamma - \beta)$, ρ_k^{in} est dans $L^1(\Omega)$, et que ρ_1^{in} est dans $L^p(\Omega)$ pour tout $p \in]1, +\infty[$.*

Alors, il existe une solution faible aux équations de coagulation-fragmentation (1.1)-(1.4)*, dont les moments vérifient*

$$\int_0^T t^{m-1} \int_\Omega \rho_{k+m(\gamma-1)}(t,x) dx dt < \infty, \quad \forall \ m \in \mathbb{N}^*. \tag{1.17}$$

8. En effet, il est raisonnable de supposer que la suite $(d_i)_{i\in\mathbb{N}^*}$ est décroissante, les particules plus grosses diffusant moins, et dans ce cas la suite $(d_i)_{i\in\mathbb{N}^*}$ converge automatiquement puisqu'elle est minorée.

On note que (1.17) avec $m = 1$ donne une borne dans $L^1([0,T] \times \Omega)$ pour un moment sur-linéaire ρ_k, et donc implique que la gélification ne peut pas se produire. L'hypothèse clef dans le théorème 1.1.5 est $\gamma > \alpha + \beta$, et traduit que la fragmentation est suffisamment forte comparée à la coagulation. Comme on l'a mentionné dans la section 1.1.3, le fait qu'une fragmentation suffisamment importante puisse empêcher la gélification même pour une coagulation sur-linéaire n'est pas surprenant, mais ce résultat (qui date des travaux de da Costa [14]) n'avait été démontré jusqu'à maintenant que dans le cadre homogène (i.e. sans diffusion).

On présente maintenant le dernier de nos résultats concernant les équations de coagulation-fragmentation, dans lequel on utilise les estimations sur les moments fournies par les théorèmes 1.1.3 et 1.1.5 pour prouver l'existence de solutions régulières.

Theorem 1.1.6. *Soient Ω un domaine borné et régulier de $\mathbb{R}^N$, $N \leq 2$, et $T > 0$. On suppose que les coefficients de coagulation et de fragmentation vérifient*

$$a_{i,j} \leq C_Q \left(i^\alpha j^\beta + i^\beta j^\alpha \right) \quad et \quad B_i \leq C_{max} i^{\gamma_{max}},$$

avec $0 \leq \alpha, \beta \leq 1$, $0 \leq \gamma_{max} < \infty$, et que les coefficients de diffusion vérifient (1.15)*. On suppose également que les concentrations initiales $c_i^{in} \geq 0$ sont de classe $\mathcal{C}^\infty(\overline{\Omega})$, compatibles avec les conditions aux limites et que pour tout $k \in \mathbb{N}^*$ les moments initiaux ρ_k^{in} sont de classe $\mathcal{C}^\infty(\overline{\Omega})$. Enfin, on suppose que l'on est dans l'une des deux situations suivantes.*

Coagulation sous-linéaire :

$$\alpha + \beta < 1 \quad et \quad \lim_{j \to \infty} \frac{B_{i+j} \beta_{i+j,i}}{i+j} = 0, \quad \forall\ i \in \mathbb{N}^*.$$

Fragmentation suffisamment forte :

Il existe $C_F > 0$ et $\gamma \geq 1$, $\gamma > \alpha + \beta$, tels que $B_i \geq C_F i^\gamma$, et (1.16) *est vérifiée.*

Alors, il existe une solution régulière aux équations de coagulation-fragmentation (1.1)-(1.4) *telle que chaque c_i est de classe $\mathcal{C}^\infty([0,T] \times \overline{\Omega})$, et telle que les moments ρ_k sont aussi de classe $\mathcal{C}^\infty([0,T] \times \overline{\Omega})$, pour tout $k \in \mathbb{N}^*$. De plus, si $\sup_{i \in \mathbb{N}^*} B_i < \infty$, la solution est unique.*

On renvoie à nouveau aux chapitres 4 et 5 pour des commentaires détaillés à propos des différentes hypothèses de ce théorème et un possible affaiblissement de certaines d'entre elles.

1.1.7 Conclusion et perspectives

Sous une hypothèse de convergence des coefficients de diffusion (1.15), on a démontré des estimations L^p pour tous les moments polynomiaux associés aux solutions des équations de coagulation-fragmentation discrètes avec diffusion (1.1)-(1.4), sous réserve que les coefficients de coagulation soient sous-linéaires. Toujours sous l'hypothèse (1.15), mais cette fois dans un cadre incluant des coefficients de coagulation sur-linéaires, on a démontré des estimations L^p similaires (avec un poids en temps) pour les moments polynomiaux ρ_k, sous réserve que la fragmentation soit suffisamment forte par rapport à la coagulation. A notre connaissance, ce résultat fournit la première preuve du fait qu'une fragmentation suffisamment forte peut empêcher la gélification, pour les équations de coagulation-fragmentation *avec* diffusion. Enfin ces estimations sur les moments permettent de démonter l'existence de solutions régulières pour les équations de coagulation-fragmentation discrètes avec diffusion (1.1)-(1.4).

On termine cette section introductive en présentant quelques extensions naturelles suggérées par les résultats de cette thèse, ainsi que quelques questions qui restent ouvertes.

- L'hypothèse (1.15), qui est utilisée dans tous les résultats princpaux exposés dans la section 1.1.6, n'est pas complètement satisfaisante du point de vue physique. En effet, plus les particules sont grosses moins elles devraient diffuser, et il est donc légitime de supposer que la suite des coefficients de diffusion d_i converge. Cependant la limite attendue est plutôt 0, et pas une valeure strictement positive. Il semble donc naturel d'essayer d'étendre les résultats de cette thèse au cas où les coefficients de diffusion d_i convergent vers 0. On mentionne que des résultats partiels dans cette direction ont été obtenus dans [9, 23].
- Un autre point intéressant concerne la gélification pour des coefficients de coagulation linéaires $a_{i,j} = C(i+j)$. On sait dans le cadre homogène (i.e. sans diffusion) que la gélification ne peut pas se produire avec de tels coefficients de coagulation, mais la question reste toujours ouverte pour le modèle incluant la diffusion.

1.2 Introduction à la partie III : *Calcul rigoureux*, théorie et applications

Dans cette section, on introduit les résultats obtenus dans la partie III de cette thèse, où on s'intéresse à des méthodes de preuves assistées par ordinateur dans le domaine des systèmes dynamiques. On commence par passer brièvement en revue différents types de méthodes dans la section 1.2.1. On décrit plus précisément dans la section 1.2.2 les techniques basées sur des *validations a posteriori*, via des théorèmes de point fixe, qui sont celles utilisées dans cette thèse. Dans les sections 1.2.3, 1.2.4 et 1.2.5, on présente une partie des contributions de cette thèse. Les travaux de ces trois chapitres permettent d'élargir le champ d'application des preuves assistées par ordinateur appliquées aux systèmes dynamiques et d'augmenter leur efficacité. Dans les sections 1.2.6 and 1.2.7, on présente le reste des contributions de cette thèse. Dans ces deux chapitres, on applique des méthodes de preuves assistées par ordinateur pour étudier des systèmes non linéaires qui résistent aux méthodes purement analytiques, ce qui permet répondre à certaines questions jusqu'alors ouvertes. Pour finir, on résume les résultats obtenus et on présente de possibles futures directions de recherche dans la section 1.2.8.

1.2.1 A propos des preuves assistées par ordinateur pour les systèmes dynamiques

Pour comprendre le comportement global d'un système non linéaire, la première étape consiste à étudier ses ensembles invariants. En effet, des solutions spécifiques comme les états stationnaires, les orbites périodiques et les connexions entre ces solutions forment les blocs de base qui organisent la dynamique globale. Bien qu'il existe de nombreuses théories mathématiques pour étudier l'existence de telles solutions, il est souvent difficile des les appliquer pour un exemple spécifique. Qui plus est, lorsqu'on se concentre sur une application précise, les propriétés qualitatives de telles solutions sont souvent aussi intéressantes que la simple preuve de leur existence. Dans ce cas, un outil puissant et fréquemment utilisé est l'analyse numérique des solutions, qui est parfaitement adaptée à l'étude d'un système explicite et peut permettre de mieux comprendre le comportement des solutions, pour un problème où les termes non linéaires empêchent l'usage de techniques purement analytiques.

On peut aller encore plus loin, et combiner des arguments théoriques avec des simulations numériques pour obtenir des preuves assistées par ordinateur. De telles techniques ont déjà été utilisées avec succès pour résoudre plusieurs problèmes célèbres, l'exemple le plus connu étant sans doute celui du théorème des quatre couleurs [189]. Pour des applications relatives aux systèmes dynamiques, on mentionne la première preuve de l'universalité de la constante de Feigenbaum [152], ainsi que la preuve de l'existence de l'attracteur étrange de Lorenz [200]. Pour une description plus exhaustive de l'histoire et des applications des preuves assistées par

ordinateur dans le cadre des systèmes dynamiques, on recommande les références suivantes [63, 173, 201, 191, 177]. On notera que l'intérêt de ces techniques pour prouver l'existence de solutions spécifiques est au moins double. D'une part, les solutions en questions peuvent avoir un intérêt propre (voir notamment les applications dans les chapitres 9 et 10), mais être difficiles voire impossibles à obtenir en utilisant des méthodes analytiques. D'autre part, l'existence de solutions spécifiques est parfois une condition suffisante pour obtenir des dynamiques complexes (dans l'esprit de *période trois implique chaos)*. Dans le domaines des équations différentielles ordinaires, un exemple bien connu est le théorème de Shilnikov [195], où l'existence d'une orbite homocline implique l'existence d'une infinité d'orbites périodiques (voir aussi [62]). Dans une direction similaire, on mentionne que les méthodes de preuves assistées par ordinateur peuvent être associées à des méthodes topologiques comme la théorie de Morse-Conley, pour obtenir des informations sur la dynamique globale d'un système [51].

Dans la partie III de cette thèse, on développe et on applique des méthodes de preuves assistées par ordinateur pour étudier des solutions spécifiques et des variétés invariantes d'équations différentielles ordinaires (EDO) et d'équations aux dérivées partielles (EDP) [9]. Avant de présenter plus en détail les contributions ce cette thèse, on donne une brève description des méthodes de preuves assistées par ordinateur existantes pour les systèmes dynamiques. Ces techniques sont souvent appelées *rigorous numerics* ou *validated numerics* dans la littérature anglophone [10]. On peut les séparer en deux catégories

- La première, parfois décrite comme une approche *géométrique*, est basée sur un encadrement rigoureux des solutions numeriques, directement dans l'espace des phases. Grosso modo, les méthodes numériques classiques comme la méthode d'Euler (ou de Runge et Kutta...) qui donnent une suite de valeurs approchées, sont remplacées par des algorithmes qui donnent des suites d'ensembles dans lesquels on est certain que la vraie solution se trouve. Une construction simpliste de tels algorithmes aboutit à des ensembles dont la taille explose très rapidement, mais il existe des techniques astucieuses (comme l'algorithme de Lohner [218]) qui permettent d'éviter les effets de *wrapping*, rendant ces techniques utilisables en pratique. On renvoie à [86, 220, 125, 217, 106] pour une exposition plus poussée de cette approche ainsi que des exemples d'application.
- La deuxième, parfois décrite comme une approche *fonctionnelle*, fournit également un encadrement rigoureux de solutions numériques, mais cette fois dans un espace fonctionnel adapté, sous la forme d'une *validation a posteriori*. Étant donné une solution numérique, la stratégie est d'établir des estimations permettant d'appliquer un théorème de point fixe, afin de démonter qu'une vraie solution existe au voisinage de la solution numérique. Les techniques développées et utilisées dans la partie III de cette thèse s'inscrivent dans cette catégorie, et on détaille donc la présentation de l'approche *fonctionnelle* dans la section suivante.

Dans les deux cas, les erreurs d'arrondi doivent être contrôlées à un certain point si on veut obtenir des énoncés mathématiques rigoureux. Ceci peut être fait en utilisant l'arithmétique des intervalles (dans cette thèse on utilise la *toolbox* INTLAB [190]).

1.2.2 Validation a posteriori via des théorèmes de point fixe

Les techniques décrites dans cette partie visent à prouver l'existence d'un zéro isolé pour une fonction F définie sur un espace de Banach $\mathcal{X}$. On supposera que F est de la forme

$$F = L + N,$$

9. Des techniques similaires peuvent aussi être utilisées pour étudier des équations différentielles avec délais, mais cette possibilité n'est pas explorée dans cette thèse.

10. On utilisera le terme *calcul rigoureux* pour désigner ces méthodes dans cette introduction en français.

où L est une application linéaire inversible ayant une inverse compacte et N est un terme non linéaire, tels que $L^{-1}N$ envoie $\mathcal{X}$ dans lui même est soit compact [11]. Étant donné un zéro approché $\bar{x}$ de F, la stratégie usuelle est de considérer un opérateur de type Newton $T : \mathcal{X} \to \mathcal{X}$ de la forme

$$T = I - AF, \tag{1.18}$$

et d'appliquer les théorèmes de point fixe de Schauder ou de Banach sur un voisinage de $\bar{x}$, pour obtenir (en supposant que A est injectif) l'existence d'un vrai zéro de F dans ce voisinage. Les différentes techniques de ce genre se distinguent par le choix de A, par les différents théorèmes de point fixe invoqués, et surtout par les estimations utilisées pour pouvoir appliquer ces théorèmes de point fixe.

La méthode principalement utilisée dans cette thèse a été introduite par Day, Lessard et Mischaikow [109]. Elle consiste à prendre pour A une inverse approchée de $DF(\bar{x})$ judicieusement choisie (plus de détails dans la section 1.2.3), et de démonter qu'on peut appliquer le théorème du point fixe de Banach à T défini comme en (1.18), dans un voisinage de $\bar{x}$. Cette technique est inspirée des travaux de Yamamoto [214], qui sont décrits plus loin. On mentionne également les travaux d'Arioli et Koch [53], qui ont développé indépendamment une technique similaire.

On note qu'appliquer le théorème du point fixe de Banach à un opérateur de type Newton comme T défini en (1.18), revient en fait à appliquer un théorème de type Newton-Kantorovich (voir par exemple [184, 188]) à F, comme le montre l'énoncé suivant.

Proposition 1.2.1. *Soient $\mathcal{X}, \mathcal{Y}$ des espaces de Banach, $F : \mathcal{X} \to \mathcal{Y}$ une fonction de classe $\mathcal{C}^1$ et $A \in \mathcal{L}(\mathcal{Y}, \mathcal{X})$ injectif. On considère $\bar{x} \in \mathcal{X}$ et on suppose qu'il existe des constantes positive Y, Z_1 et une fonction strictement positive, croissante, convexe et régulière $r \mapsto Z_2(r)$ telles que*

$$\|AF(\bar{x})\|_{\mathcal{X}} \leq Y \tag{1.19}$$

$$\|I - ADF(\bar{x})\|_{\mathcal{X}} \leq Z_1 \tag{1.20}$$

$$\|A(DF(x) - DF(\bar{x}))\|_{\mathcal{X}} \leq r Z_2(r) \quad \forall\ x \in \mathcal{B}_{\mathcal{X}}(\bar{x}, r), \tag{1.21}$$

où $\mathcal{B}_{\mathcal{X}}(\bar{x}, r)$ désigne la boule fermée de $\mathcal{X}$, centrée en $\bar{x}$ et de rayon r. On définit le polynôme de rayon [12] *P par*

$$P(r) = \frac{1}{2} Z_2(r) r^2 - (1 - Z_1) r + Y. \tag{1.22}$$

On suppose qu'il existe $r > 0$ tel que $P(r) < 0$, et on désigne par $\underline{r}$ et $\overline{r}$ les deux zéros positifs de P, avec $\underline{r} < \overline{r}$. Alors, F a un unique zéro dans $\mathcal{B}_{\mathcal{X}}(\bar{X}, r)$, pour tout r dans l'intervalle non vide $[r_{\min}, r_{\max}[$, où $r_{\min} = \underline{r}$ et $r_{\max} = \min(\overline{r}, r^*)$, r^* étant défini par

$$Z_2(r^*) r^* = 1 - Z_1.$$

Cette proposition n'est pas sans rappeler le théorème de Newton-Kantorovich, mais elle peut aussi être vue comme une conséquence relativement directe du théorème de point fixe de Banch (voir la preuve ci-dessous). On notera que l'hypothèse clef, qui en pratique peut être ou ne pas être vérifiée, est l'existence d'un r strictement positif tel que $P(r) < 0$, qui entraine alors l'existence d'un zéro de F, localement unique. On va justifier pourquoi on peut raisonnablement s'attendre à l'existence d'un tel r, en supposant pour simplifier que la fonction Z_2 est constante. P est alors simplement un polynôme quadratique, et l'existence d'un $r > 0$ tel que $P(r) < 0$ est équivalente aux conditions

$$Z_1 < 1 \quad \text{et} \quad 2Y Z_2 < (1 - Z_1)^2.$$

11. La méthode utilisée dans cette thèse ne requière pas explicitement que ces hypothèses soient satisfaites, ou du moins pas sous cette forme, mais c'est pour ce type de fonction F qu'on s'attend à ce que la méthode fonctionne (plus de détails dans la section 1.2.3).

12. En pratique, la fonction Z_2 est souvent polynomiale (soit parce que les termes non linéaires du problème sont eux-mêmes polynomiaux ; soit parce qu'on se restreint à un voisinage de $\bar{x}$, auquel cas on peut prendre Z_2 constante), ce qui explique la terminologie.

La condition $Z_1 < 1$ devrait être satisfaite si A est une *suffisamment bonne* inverse approchée de $DF(\bar{x})$. Une fois qu'un tel A est fixé, la condition $2YZ_2 < (1 - Z_1)^2$ devrait également être satisfaite si $\bar{x}$ est un *suffisamment bon* zéro approché de F, rendant Y très petit. Une partie importante de la méthode consiste donc à obtenir de bonnes approximations $\bar{x}$ et A, l'autre partie consistant à établir des estimations assez précises pour les bornes Y, Z_1 et $Z_2(r)$ afin qu'il existe un r positif tel que $P(r) < 0$.

Remark 1.2.2. *Cette technique est* assistée par ordinateur *à deux niveaux. Premièrement, l'ordinateur est très utile (et souvent absolument nécessaire) pour obtenir une bonne solution approchée $\bar{x}$ et ensuite définir A (qui dépend de $\bar{x}$). Deuxièmement, on utilise également l'ordinateur pour établir les bornes Y, Z_1 et $Z_2(r)$, qui sont en général obtenues en combinant des quantités calculées numériquement (car dépendant de $\bar{x}$) et des estimations théoriques (essentiellement pour contrôler les erreurs de troncatures, entre le sous espace de dimension finie de $\mathcal{X}$ qui doit être utilisé numériquement et l'espace $\mathcal{X}$ tout entier). Le fait que* (1.19)-(1.21) *et la condition $P(r) < 0$ soient des inégalités met en évidence l'un des avantages de la reformulation en terme de point fixe : vérifier rigoureusement des* égalités *(comme $F(x) = 0$) à l'aide d'un ordinateur est plutôt délicat, alors que vérifier rigoureusement des* inégalités *(comme celles qu'on vient juste de mentionner) peut être fait assez aisément en utilisant l'arithmétique des intervalles pour contrôler les erreurs d'arrondi.*

Remark 1.2.3. *De nombreux travaux dans le domaine du calcul rigoureux sont basés sur la proposition 1.2.1, ou sur de légères variantes de ce résultat. Cependant, dans tous ces travaux le facteur $\frac{1}{2}$ est absent dans la définition de P, ce qui rend la méthode moins optimale qu'elle ne pourrait l'être*[13]*. On mentionne que ce facteur $\frac{1}{2}$ est présent dans la version usuelle du théorème de Newton-Kantorovich.*

Preuve de la proposition 1.2.1. On note que (grâce aux hypothèses sur Z_2) P est convexe (sur $\mathbb{R}_+$) et que $P(0) = Y \geq 0$. L'existence d'un $r > 0$ tel que $P(r) < 0$ implique donc que P a exactement deux zéros positifs distincts, $\underline{r}$ et $\overline{r}$ sont donc bien définis.

On commence par montrer que l'image par T, défini en (1.18), de la boule $\mathcal{B}_{\mathcal{X}}(\bar{x}, r)$ est contenue dans cette même boule, pour tout $r \in [\underline{r}, \overline{r}]$. Pour ce faire, on considère $x \in \mathcal{B}_{\mathcal{X}}(\bar{x}, r)$ et on majore

$$\begin{aligned}
\|T(x) - \bar{x}\|_{\mathcal{X}} &\leq \|T(x) - T(\bar{x})\|_{\mathcal{X}} + \|T(\bar{x}) - \bar{x}\|_{\mathcal{X}} \\
&\leq \|x - \bar{x} - A(F(x) - F(\bar{x}))\|_{\mathcal{X}} + Y \\
&\leq \|I - ADF(\bar{x})\|_{\mathcal{X}} \|x - \bar{x}\|_{\mathcal{X}} + \|A(F(x) - F(\bar{x}) - DF(\bar{x})(x - \bar{x}))\|_{\mathcal{X}} + Y \\
&\leq Z_1 r + \frac{1}{2} Z_2(r) r^2 + Y \\
&\leq r,
\end{aligned}$$

la dernière inégalité étant exactement $P(r) \leq 0$, qui est valide pour tout $r \in [\underline{r}, \overline{r}]$.

L'étape suivante est de montrer que T est contractant sur $\mathcal{B}_{\mathcal{X}}(\bar{x}, r)$, pour tout $r \in [r_{\min}, r_{\max}[$, afin de pouvoir appliquer le théorème du point fixe de Banach qui donne alors le résultat annoncé (un point fixe de T correspond bien à un zéro de F car on a supposé A injectif). En partant de $DT(x) = I - ADF(x)$ et en introduisant $DF(\bar{x})$ comme dans le calcul précédant, on obtient

$$\|DT(x)\|_{\mathcal{X}} \leq Z_1 + Z_2(r) r, \quad \forall\, x \in \mathcal{B}_{\mathcal{X}}(\bar{x}, r).$$

Ainsi, pour que T soit contractant sur $\mathcal{B}_{\mathcal{X}}(\bar{x}, r)$, il suffit que r soit tel que $Z_1 + Z_2(r) r < 1$, ce qui équivaut à avoir $r < r^*$ (car Z_2 est une fonction croissante). On a donc bien montré que T était une contraction de la boule fermée $\mathcal{B}_{\mathcal{X}}(\bar{x}, r)$ dans elle même, pour tout $r \in [r_{\min}, r_{\max}[$.

13. Autrement dit, dans une situation où les bornes Y, Z_1 ou Z_2 sont légèrement trop grandes et où il n'existe pas de $r > 0$ tel que $\tilde{P}(r) < 0$ pour le polynôme de rayon *usuel* $\tilde{P}(r) = Z_2(r) r^2 - (1 - Z_1) r + Y$, il se peut qu'il existe quand même un $r > 0$ tel que $P(r) < 0$ avec P défini comme dans la proposition 1.2.1.

Il reste à s'assurer que $\underline{r} < r^*$, pour monrter que $[r_{\min}, r_{\max}[$ est bien non vide. Si Z_2 était une constante, r^* serait l'abscisse du sommet du polynôme quadratique P, et on aurait bien que $\underline{r} < r^*$. Pour traiter le cas général où Z_2 peut dépendre de r, on introduit le polynôme quadratique

$$\underline{P}(r) = \frac{1}{2} Z_2(\underline{r}) r^2 - (1 - Z_1) r + Y.$$

Puisque Z_2 est croissant, on a $\underline{P}(r) \geq P(r)$ pour $0 \leq r \leq \underline{r}$, et $\underline{P}(r) \leq P(r)$ pour $r \geq \underline{r}$. En particulier, $\underline{r}$ est la plus petite racine de $\underline{P}$, i.e.

$$\underline{r} = \frac{1 - Z_1 - \sqrt{(1 - Z_1)^2 - 2Y Z_2(\underline{r})}}{Z_2(\underline{r})}.$$

On en déduit que $Z_2(\underline{r})\underline{r} < 1 - Z_1$, et par croissance de Z_2 on a bien $\underline{r} < r^*$. □

Dans les chapitres 6 à 10 on applique avec succès la proposition 1.2.1 [14] pour prouver l'existence de solutions dans des contextes variés. Avant de présenter ces résultats plus en détail, on passe en revue plusieurs techniques analogues de validation a posteriori.

On commence par la technique introduite par Yamamoto [214], qui est aussi basée sur le théorème de point fixe de Banach. Dans la méthode de Yamamoto, on considère $\tilde{F} = L^{-1}F = I + L^{-1}N$ et une décomposition $\mathcal{X} = \mathcal{X}_h \oplus \mathcal{X}_\infty$, où $\mathcal{X}_h$ est un sous espace de dimension finie de $\mathcal{X}$. On note Π_h la projection sur $\mathcal{X}_h$ et $A_h = \left(\Pi_h D\tilde{F}(\bar{x})|_{\mathcal{X}_h}\right)^{-1}$. L'opérateur de point fixe considéré est alors donné par

$$\tilde{T}(x) = \left(\Pi_h x - A_h \Pi_h \tilde{F}(x)\right) + (I - \Pi_h)\left(\tilde{F}(x) - x\right). \tag{1.23}$$

On note que, si A_h est injectif, $\tilde{T}(x) = x$ implique $\tilde{F}(x) = 0$ qui implique $F(x) = 0$. Ici, la projection de dimension finie $\Pi_h \tilde{T}(x) = \left(\Pi_h x - A_h \Pi_h \tilde{F}(x)\right)$ est de type Newton, et le *reste de troncature* $(I - \Pi_h)\tilde{T}(x) = (I - \Pi_h)\left(\tilde{L}^{-1}N(x)\right)$ devrait être petit si la dimension de $\mathcal{X}_h$ est assez grande. Yamamoto utilise des conditions suffisantes similaires [15] à celles de la proposition 1.2.1 pour appliquer le théorème de point fixe à $\tilde{T}$ et ainsi obtenir l'existence de zéros de F. Les différences entre *notre* opérateur T défini en 1.18 et l'opérateur de Yamamoto défini en 1.23 sont principalement des différences de présentation (dans notre cas, la projection de dimension finie est dissimulée dans la définition de A). Elles sont liées au fait qu'en pratique notre F est défini en utilisant des méthodes spectrales, alors que les travaux de Yamamoto utilisent des méthodes d'approximation locales de type éléments finis. On trouvera davantage de détails sur l'utilisation de méthodes spectrales dans le cadre du calcul rigoureux dans la section 1.2.3 (des méthodes spectrales sont également utilisées dans les sections 1.2.4, 1.2.6 et 1.2.7), et on discutera de l'utilisation de méthodes locales, comme l'interpolation polynomiale, dans le cadre du calcul rigoureux dans la section 1.2.5.

Une autre technique que l'on se doit de mentionner est celle de Nakao (voir l'article de synthèse [177] et les références qui y sont mentionnées), a qui on doit l'introduction de l'opérateur de type Newton (1.23) [16] dans le cadre du calcul rigoureux, et qui a influencé les travaux de

14. En pratique, on utilise une légère variation de la proposition 1.2.1, qui fait intervenir une application linéaire $A^\dagger$ approchant $DF(\bar{x})$. Le terme $I - ADF(\bar{x})$ est alors séparé en $\left(I - AA^\dagger\right) + A\left(DF(\bar{x}) - A^\dagger\right)$ et les deux termes sont majorés séparément. De ce fait, la borne Z_1 de la proposition 1.2.1 est en pratique remplacée par deux bornes (notées Z_0 et Z_1).

15. Pour être précis, la présentation de la proposition 1.2.1 étant inspirée de l'article [214], ce sont plutôt nos estimations qui sont similaires à celle de Yamamoto.

16. Dans des études antérieures, notamment l'article original de Nakao [176] sur ce type de méthodes, d'autres reformulations en terme de point fixe on été utilisées, comme $T = (-L)^{-1}N$. Cependant, l'avantage de l'approche de type Newton est que les théorèmes de point fixe de Banach ou de Schauder devraient *en principe* être toujours applicables près de la solution numérique, ce qui n'était pas les cas pour ces formulations alternatives.

Yamamoto. La principale différence avec la technique de Yamamoto est que l'approche de Nakao est basée sur le théorème du point fixe de Schauder. La stratégie consiste à construire un sous ensemble fermé, borné et convexe $W = W_h \oplus W_\infty$, où $W_h \subset \mathcal{X}_h$ et $W_\infty \subset \mathcal{X}_\infty$, sont tels que

$$\Pi_h \tilde{T}(W) \subset W_h \quad \text{et} \quad (I - \Pi_h)\tilde{T}(W) \subset W_\infty.$$

Le terme de dimension finie est traité en utilisant de manière importante des inclusions d'ensembles (voir [177] pour les détails), ce qui fait que la méthode de Nakao est parfois décrite comme une approche hybride entre les méthodes géométriques et fonctionnelles.

On termine en mentionnant la méthode introduite par Plum (voir l'article de synthèse [187] et les références qui y sont mentionnées), qui au premier abord peut avoir l'air similaire à la méthode de Nakao, dans la mesure où elle est aussi basée sur le théorème du point fixe de Schauder, mais qui a également de nombreuses similarités avec notre approche. La technique de Plum est basée sur une reformulation en terme d'un opérateur de point fixe de la forme

$$\hat{T}(x) = \bar{x} - (DF(\bar{x}))^{-1} \left[F(\bar{x}) + N(x) - N(\bar{x}) - DN(\bar{x})(x - \bar{x})\right],$$

pour lequel on cherche à montrer que, pour un certain $r > 0$, T envoie la boule $\mathcal{B}_\mathcal{X}(\bar{x}, r)$ dans elle même afin d'appliquer le théorème de point fixe de Schauder (une vérification directe, en se rappelant que $F = L + N$ avec L linéaire, montre que les points fixes de T correspondent bien au zéros de F). La méthode de Plum donne des conditions suffisante [17] pour que l'inclusion $\hat{T}(\mathcal{B}_\mathcal{X}(\bar{x}, r)) \subset \mathcal{B}_\mathcal{X}(\bar{x}, r)$ soit vérifiée, qui sont très semblables à celles de la proposition 1.2.1. En effet, étant donné des bornes $\hat{Y}$, K et une fonction croissante $\hat{Z}_2$ telles que

$$\|F(\bar{x})\|_\mathcal{X} \leq \hat{Y}, \quad \|DF(\bar{x})^{-1}\|_\mathcal{X} \leq K \quad \text{et} \quad \|DN(x) - DN(\bar{x})\|_\mathcal{X} \leq r\hat{Z}_2(r) \quad \forall\, x \in \mathcal{B}_\mathcal{X}(\bar{x}, r),$$

on a $\hat{T}(\mathcal{B}_\mathcal{X}(\bar{x}, r)) \subset \mathcal{B}_\mathcal{X}(\bar{x}, r)$ dès que

$$\hat{P}(r) := \frac{1}{2} K \hat{Z}_2(r) r^2 - r + K\hat{Y} < 0.$$

La différence principale avec notre formulation est qu'au lieu d'utiliser une inverse approchée A de $DF(\bar{x})$, la méthode de Plum considère l'inverse exacte (plus précisément, une borne de la norme de l'inverse exacte). Ceci explique l'absence du terme Z_1 (qui peut être pris égal à 0 si $A = (DF(\bar{x}))^{-1}$). Hormis cette différence, les conditions sont pratiquement identiques, et ceci s'explique en remarquant que

$$\begin{aligned}
\hat{T}(x) &= \bar{x} - (DF(\bar{x}))^{-1} \left[L\bar{x} + N(\bar{x}) + N(x) - N(\bar{x}) - DN(\bar{x})(x - \bar{x})\right] \\
&= \bar{x} - (DF(\bar{x}))^{-1} \left[Lx + N(x) - (L + DN(\bar{x}))(x - \bar{x})\right] \\
&= \bar{x} - (DF(\bar{x}))^{-1} \left[F(x) - DF(\bar{x})(x - \bar{x})\right] \\
&= x - (DF(\bar{x}))^{-1} F(x).
\end{aligned}$$

Ainsi, $\hat{T}$ n'est rien d'autre qu'un opérateur de type Newton de la forme (1.18), avec $A = (DF(\bar{x}))^{-1}$. Cette différence au niveau du choix de A, qui a déjà été mentionnée plus haut, a un impact sur la manière dont on obtient les différentes bornes. Dans notre approche, on prend pour A une approximation bien choisie de $DF(\bar{x})^{-1}$, qui est en un certain sens plus simple que $DF(\bar{x})^{-1}$ (ce point est détaillé dans la section 1.2.3). Ce choix nous permet de majorer directement la norme de certains produits comme $\|AF(\bar{x})\|$ et $\|A(DF(x) - DF(\bar{x}))\|$, alors que l'approche de Plum utilise des majorations de la forme $\|A\|\|F(\bar{x})\|$ et $\|A\|\|(DF(x) - DF(\bar{x}))\|$ (dans son cas avec A égal à $DF(\bar{x})^{-1}$), qui sont génériquement plus grossières. Cependant, on paye un prix pour nos meilleures majorations, qui est qu'on doit construire l'inverse approchée

17. on en a légèrement reformulé certaines pour faciliter la comparaison.

A, ce qui en pratique requiert l'inversion d'une matrice représentant une projection en dimension finie de $DF(\bar{x})$ (cette inverse est en fait similaire au A_h des méthodes de Nakao et Yamamoto, en revoie à nouveau vers la section 1.2.3 pour plus de détails). Dans la méthode de Plum, on n'a pas besoin d'inverser $DF(\bar{x})$ mais seulement de borner la norme de l'inverse. Pour obtenir une telle borne, Plum utilise des méthodes pour d'encadrement de valeurs propres (de $DF(\bar{x})$ ou de $DF(\bar{x})^* DF(\bar{x})$). Ces méthodes sont elles mêmes plus ou moins difficiles à appliquer, en fonction du problème qu'on considère (en particulier il faut que l'espace de Banach $\mathcal{X}$ ait une structure hilbertienne).

Dans les cinq sections suivantes, on décrit les contributions de cette thèse relatives aux méthodes de calcul rigoureux. Les travaux présentés dans les sections 1.2.3, 1.2.4 et 1.2.5 sont principalement ciblés sur l'élargissement du cadre d'application de ces techniques et sur l'amélioration de leur efficacité (ce qui agrandit aussi leur cadre d'application, d'un point du vue pratique). Les travaux présentés dans les sections 1.2.6 et 1.2.7 sont plus axés sur les applications, et fournissent de parfaits exemples où les techniques de calcul rigoureux permettent de résoudre des problèmes qui n'avaient pas pu être traités par des techniques purement analytiques

1.2.3 Exposition des résultats obtenus dans le chapitre 6

Dans cette section, on présente les résultats principaux du chapitre 6 de cette thèse. Pour mettre ces résultats en perspective, on commence par donner un exemple simple d'équation où notre technique de calcul rigoureux peut être appliquée, et on détaille pour cet exemple la construction de l'inverse approchée A.

On considère l'équation

$$\begin{cases} -u'' + u + u^2 = g, \\ u'(0) = 0 = u'(\pi), \end{cases} \tag{1.24}$$

où g est une fonction régulière, paire et 2π périodique. Cet exemple pourrait bien sûr être étudié sans avoir recours aux méthodes de calcul rigoureux, par exemple via une approche variationnelle. Notre but ici est seulement de présenter les idées principales de notre méthode, sans avoir à s'encombrer de trop de détails techniques.

On considère le développement en série de Fourier de u [18] et g :

$$u(t) = x_0 + 2\sum_{k=1}^{\infty} x_k \cos(kx) \quad \text{et} \quad g(t) = g_0 + 2\sum_{k=1}^{\infty} g_k \cos(kx),$$

et on insère ces formules dans (1.24), pour aboutir au système (infini) d'équations algébriques suivant

$$k^2 x_k + x_k + (x * x)_k = g_k, \quad \forall\ k \in \mathbb{N},$$

où $(x * y)$ désigne le produit de convolution des suites x et y, défini par

$$(x * y)_k = \sum_{l \in \mathbb{Z}} x_{|k-l|} y_{|l|}, \quad \forall\ k \in \mathbb{N}.$$

En considérant $x = (x_k)_{k\in\mathbb{N}}$ la suite des coefficients de Fourier de u, ce système peut se réécrire sous la forme $F(x) = 0$, avec $F = (F_k)_{k\in\mathbb{N}}$ défini par

$$F_k(x) = k^2 x_k + x_k + (x * x)_k - g_k, \quad \forall\ k \in \mathbb{N}.$$

C'est exactement ce genre de fonction $F : \mathcal{X} \to \mathcal{Y}$ qu'on avait à l'esprit dans la section 1.2.2, avec L défini comme [19] $L_k(x) = k^2 x_k$ pour tout $k \in \mathbb{N}$. Dans ce cadre, les espaces de Banach $\mathcal{X}$,

18. pour l'instant de manière formelle, on verra que la décroissance des coefficients obtenus permettra de justifier a posteriori tous les calculs.

19. Le terme d'*ordre zero* x_k pourrait aussi être incorporé dans L, on considérerait alors $L_k x = (k^2 + 1)x_k$, mais ceci importe peu ici.

dans lequel on veut prouver l'existence d'un zéro de F, et $\mathcal{Y}$ sont typiquement des espaces ℓ^1 ou ℓ^∞ à poids (des espaces ℓ^∞ à poids sont utilisés dans le chapitre 6, et des espaces ℓ^1 à poids sont utilisés dans les chapitres 7, 9 et 10).

Etant donné un entier $m > 0$, on suppose maintenant qu'on a calculé numériquement un zéro approché de F, en considérant une projection de Galerkin de dimension m. Plus précisément, on introduit $\mathcal{X}^{[m]}$ le sous-espace de $\mathcal{X}$ formé des suites x telles $x_k = 0$ pour tout $k \geq m$, ainsi que le projecteur $\Pi^{[m]} : \mathcal{X} \to \mathcal{X}^{[m]}$ associé, défini par

$$\Pi_k^{[m]}(x) = \begin{cases} x_k & \forall\ k < m, \\ 0 & \forall\ k \geq m. \end{cases}$$

On considère alors $F^{[m]} = \Pi^{[m]} F_{|\mathcal{X}^{[m]}}$, qui peut être vue comme une fonction de $\mathbb{R}^m$ dans lui-même, pour laquelle on peut numériquement calculer un zéro (par exemple en utilisant la méthode de Newton). On suppose qu'on a calculé un tel zéro approché, et on le note $\bar{x}$.

L'étape suivante consiste à définir une inverse approchée de $DF(\bar{x})$. Pour ce faire, on commence par approcher $DF(\bar{x})$ lui même. On a, pour tout $x \in \mathcal{X}$,

$$DF_k(\bar{x})x = k^2 x_k + x_k + 2(\bar{x} * x)_k, \quad \forall\ k \in \mathbb{N}.$$

On note que le terme asymptotiquement dominant est donné par $L_k x = k^2 x_k$. De ce fait, en identifiant les applications linéaires entre $\mathcal{X}$ et $\mathcal{Y}$ avec des matrices infinies, on définit une approximation de $DF(\bar{x})$ par

$$A^\dagger = \left(\begin{array}{c|ccc} DF^{[m]}(\bar{x}) & & (0) & \\ \hline & m^2 & & (0) \\ (0) & & (m+1)^2 & \\ & (0) & & \ddots \end{array}\right),$$

qui est constitué du bloc de dimension finie $DF^{[m]}(\bar{x})$ de taille $m \times m$, et d'un bloc (infini) diagonal où l'on a seulement conservé les termes venant de L. On considère alors $A^{[m]}$ une inverse, calculée numériquement, de $DF^{[m]}(\bar{x})$, et on définit notre inverse approchée A par

$$A = \left(\begin{array}{c|ccc} A^{[m]} & & (0) & \\ \hline & \frac{1}{m^2} & & (0) \\ (0) & & \frac{1}{(m+1)^2} & \\ & (0) & & \ddots \end{array}\right).$$

C'est ce type d'inverse approchée A de $DF(\bar{x})$ auquel on a fait allusion dans la section 1.2.2, et qu'on utilise pour la proposition 1.2.2. La *qualité* de cette approximation est contrôlée par la borne Z_1 de la proposition 1.2.1. Pour que notre méthode s'applique, il faut qu'on puisse obtenir une borne Z_1 strictement inférieure à 1, ce qui est la cas ici avec un tel A, si m est suffisamment grand [20]. Le fait que A ait une structure simple (i.e. un bloc de taille fini et un *terme de queue* diagonal) est utilisé de manière crucial dans notre approche pour établir les bornes requises dans la proposition 1.2.1.

On insiste sur le fait qu'on peut obtenir une telle structure simple pour $A^\dagger$, et donc pour A, uniquement parce que le terme asymptotiquement dominant, donné par L, est diagonal. En

20. en supposant que $A^{[m]}$ est une suffisamment bonne inverse de $DF^{[m]}(\bar{x})$.

effet, on ne peut utiliser une inversion numérique que pour un bloc de taille finie, et on utilise le fait que la queue (infinie) de $A^\dagger$ est diagonale pour pouvoir l'inverser de manière analytique.

Cet exemple fournit une situation typique où notre méthode est applicable. Pour présenter les choses de manières plus générale, on utilise le fait que $\mathcal{X}$ a une base de Schauder sur laquelle L est diagonal, et que L est asymptotiquement dominant dans $DF(\bar{x}) = L + DN(\bar{x})$, qui est une reformulation dans ce contexte de l'hypothèse $L^{-1}DN(\bar{x})$ compact.

L'objectif des travaux du chapitre 6 est d'être capable de traiter un plus large champ d'application, incluant des cas où L n'est plus diagonal. Un exemple typique où cette situation intervient naturellement s'obtient en considérant la généralisation suivante de (1.24)

$$\begin{cases} -(2+\cos(t))u''(t) + u(t) + u^2(t) = g(t), \\ u'(0) = 0 = u'(\pi). \end{cases} \tag{1.25}$$

Toujours en réécrivant u et g sous la forme de séries de Fourier, on aboutit cette fois ci à une fonction F définie par

$$F_k(x) = \frac{1}{2}(k-1)^2 x_{k-1} + 2k^2 x_k + \frac{1}{2}(k+1)^2 x_{k+1} + x_k + (x*x)_k - g_k, \quad \forall\ k \geq 1,$$

et

$$F_0(x) = x_1 + x_0 + (x*x)_0 - g_0.$$

Le terme dominant est cette fois tridiagonal et donné par

$$L_k x = \frac{1}{2}(k-1)^2 x_{k-1} + 2k^2 x_k + \frac{1}{2}(k+1)^2 x_{k+1},$$

et une approximation *simple* de $DF(\bar{x})$ sera alors de la forme

$$A^\dagger = \left(\begin{array}{ccc|ccc} & & & & (0) & \\ & DF^{[m]}(\bar{x}) & & & \beta_{m-1} & \\ \hline & & \lambda_m & \nu_m & \beta_m & (0) \\ & (0) & & \lambda_{m+1} & \nu_{m+1} & \ddots \\ & & & (0) & \ddots & \ddots \end{array}\right), \tag{1.26}$$

avec pour notre exemple $\lambda_k = \frac{1}{2}(k-1)^2$, $\nu_k = 2k^2$ et $\beta_k = \frac{1}{2}(k+1)^2$. Pour pouvoir appliquer notre méthode de calcul rigoureux dans cette situation, il nous faut :

- trouver comment inverser de manière approchée $A^\dagger$ défini en (1.26), pour pouvoir définir A,
- adapter les estimations *usuelles* pour ce nouveau A, qui n'aura plus une simple queue diagonale.

On répond à ces questions dans le chapitre 6, sous l'hypothèse qu'il existe $k_0 \in \mathbb{N}$ et $\delta < \frac{1}{2}$ tels que

$$\left|\frac{\lambda_k}{\nu_k}\right|, \left|\frac{\beta_k}{\nu_k}\right| \leq \delta, \quad \forall\ k \geq k_0.$$

Pour définir A, on commence par inverse (de manière analytique) la queue tridiagonale (donnée par le bloc en bas à droite dans (1.26)), en utilisant une généralisation de la décomposition LU pour des matrices infinies. Après avoir également inversé le bloc fini $DF^{[m]}(\bar{x})$ (cette fois numériquement), on détermine comment ces deux inverses doivent être couplées pour prendre en compte les deux termes additionnels β_{m-1} et λ_m, pour finalement obtenir une bonne inverse

approchée A de $A^\dagger$. Du fait de la structure plus complexe de ce A (qu'on peut maintenant voir comme une matrice infinie pleine), l'obtention des bornes de la proposition 1.2.1 devient beaucoup plus technique, et on ne donne pas davantage de détails ici.

A titre d'exemple, on a calculé numériquement plusieurs solutions d'une version de l'équation (1.25) dépendant d'un paramètre, et on a validé ces solutions en utilisant les techniques qu'on vient de décrire (voir Figure 1.2).

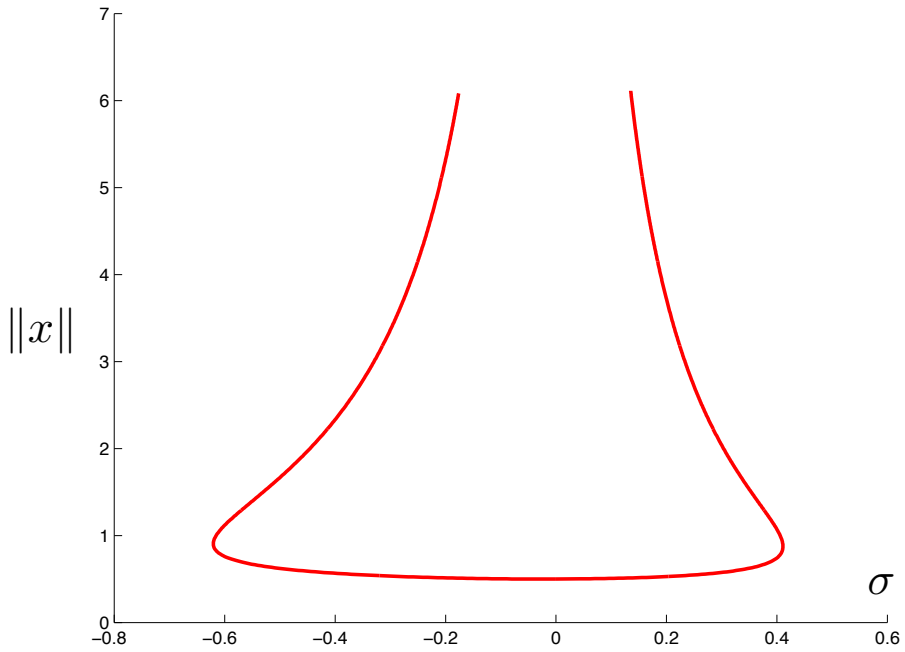


FIGURE 1.2 – Solutions validées de l'équation $-(2+\cos(t))u''(t) + u(t) + \sigma u^2(t) = g(t)$, avec conditions au limites de Neumann, et $g(t) = \frac{1}{2} + 3\cos(t) + \frac{1}{2}\cos(2t)$, pour différentes valeurs de σ.

1.2.4 Exposition des résultats obtenus dans le chapitre 7

Dans cette section, on présente les principaux résultats du chapitre 7 de cette thèse, qui traite de variétés stables et instables.

Soit $g : \mathbb{R}^n \to \mathbb{R}^n$ un champ de vecteurs $\mathcal{C}^1$. On considère l'EDO

$$y' = g(y). \tag{1.27}$$

Soit $p \in \mathbb{R}^n$ un point d'équilibre de cette équation, i.e. $g(p) = 0$. On suppose que la matrice jacobienne $Dg(p)$ a k valeurs $\lambda_1 \leq \ldots \leq \lambda_k$ de parties réelles strictement négatives. Pour simplifier la présentation on suppose que ces valeurs propres sont réelles et simples. Le cas des valeurs propres complexes est traité dans le chapitre 7, pour une généralisation incluant le cas des valeurs propres multiples on réfère à [66]. On note $V_1, \ldots, V_k$ les vecteurs propores associés et E^s le sous-espace vectoriel de dimension k de $\mathbb{R}^n$ engendré par ces vecteurs propres. On rappelle (voir par exemple [186]), qu'il existe une sous variété de dimension k, notée $W^s_{loc}(p)$, tangente au sous espace stable E^s en p, qui est stable sous l'action du flot ϕ généré par g, et qui vérifie $\phi_t(x) \underset{t\to\infty}{\longrightarrow} p$ pour tout $x \in W^s_{loc}(p)$. $W^s_{loc}(p)$ est appelée la *variété stable locale* de p. De manière similaire, il existe une *variété instable locale* associée aux valeurs propres ayant une partie réelle strictement positives.

Les variétés stables et instables locales jouent un rôle crucial dans la compréhension de la dynamique d'une solution de (1.27) près du point d'équilibre p. Elles sont aussi très utiles pour

étudier les connexions entre points d'équilibre (i.e. les solutions y telles que $y(t)$ converge vers un point d'équilibre quand $t \to \pm\infty$). En effet, si on peut prouver l'existence d'une solution qui est dans la variété instable (d'un point d'équilibre p) à un instant t_0, et dans la variété stable (d'un point d'équilibre q) à un instant t_1, alors on a terminé, car par définition des variétés stables et instables la solution doit alors converger vers p en $-\infty$ et vers q en $+\infty$. L'utilisation des variétés stables et instables permet donc de réduire le problème de trouver une connexion (définie sur $\mathbb{R}$) entre deux états d'équilibre, et de *seulement* devoir trouver une orbite (définie sur un intervalle de temps fini) qui relie les variétés. Cette réduction d'un intervalle de temps infini à fini peut s'avérer déterminante, surtout dans l'optique d'une preuve basée sur une résolution numérique préalable. Cependant, pour que cette réduction soit effective, on doit d'abord être capable d'obtenir les variétés en question.

Il existe de nombreuses techniques pour calculer de telles variétés, celle qu'on utilise dans le chapitre 7 a été introduite par Cabré, Fontich et de la Llave [82, 83, 84] et est appelé la *méthode de la paramétrisation*. Comme son nom l'indique, cette méthode est basée sur l'obtention d'une paramétrisation spécifique de la variété, qui semi-conjugue la dynamique du système avec celle du linéarisé autour du point d'équilibre. Plus précisément, en notant Λ la matrice diagonale contenant les valeurs propres stables $\lambda_1, \ldots, \lambda_k$, et $\mathcal{B}_\nu$ la boule fermée de rayon ν dans $\mathbb{R}^k$, on cherche une paramétrisation $f : \mathcal{B}_\nu \to \mathbb{R}^n$, telle que $f(0) = p$, et vérifiant

$$f(e^{t\Lambda}\theta) = \phi_t(f(\theta)), \quad \forall\ \theta \in \mathcal{B}_\nu,\ \forall\ t \geq 0. \tag{1.28}$$

On remarque que si une telle fonction f existe, alors $f(\mathcal{B}_\nu)$ est bien une variété stable locale de p [21]. Cependant cette formulation n'est pas très pratique car elle fait intervenir le flot ϕ. Pour s'en débarrasser, on dérive (1.28) par rapport à t, puis on évalue en $t = 0$, pour obtenir ce qu'on appelle *l'équation d'invariance*

$$Df(\theta)\Lambda\theta = g(f(\theta)), \tag{1.29}$$

qui est en fait équivalente à (1.28). En effet, si on suppose que f satisfait l'équation d'invariance (1.29), en considérant $h(t) = f(e^{t\Lambda}\theta)$ on a $h(0) = f(\theta)$ et $h'(t) = g(f(e^{t\Lambda}\theta)) = g(h(t))$, donc $h(t) = \phi_t(f(\theta))$ et f est bien solution de (1.28).

On suppose maintenant que le champ de vecteur g est analytique. Il est démontré dans [82] que l'équation d'invariance (1.29) admet alors une solution analytique f, sous réserve qu'une condition de *non résonance* entre les valeurs propres (détaillée dans le chapitre 7) soit satisfaite. En supposant cette condition satisfaite, on peut donc chercher f sous la forme d'une série entière

$$f(\theta) = \sum_{|\alpha|\geq 0} a_\alpha \theta^\alpha, \quad \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix} \in \mathbb{R}^k,\ a_\alpha = \begin{pmatrix} a_\alpha^{(1)} \\ \vdots \\ a_\alpha^{(n)} \end{pmatrix} \in \mathbb{R}^n, \tag{1.30}$$

avec les notations multi-indices classiques $|\alpha| = \alpha_1 + \ldots + \alpha_k$ et $\theta^\alpha = \theta_1^{\alpha_1} \ldots \theta_{n_s}^{\alpha_k}$. En utilisant cet ansatz dans (1.29), on obtient un système (infini) d'équations algébriques dont les inconnues sont les coefficients $(a_\alpha)_{|\alpha|\geq 0}$, et auquel on peut appliquer notre technique de calcul rigoureux. Cette association de la méthode de la paramétrisation avec des techniques de preuves assistées par ordinateur a pour origine l'article de van den Berg, Lessard, Mireles James et Mischaikow [65], où des paramétrisations de variétés stables locales sont calculées et validées, puis utilisées pour démontrer l'existence d'orbites homoclines pour l'équation de Gray-Scott.

Pour qu'on puisse appliquer à l'équation d'invariance (1.29) notre technique de validation, basée sur le théorème du point fixe de Banach, il faut que les solutions de cette équation soient isolées. En plus de la condition $f(0) = p$, on ajoute donc une condition pour $Df(0)$, le choix naturel étant

$$Df(0) = \begin{pmatrix} V_1 & \ldots & V_k \end{pmatrix}, \tag{1.31}$$

21. et est de dimension k si $Df(0)$ est de rang k.

ou une condition similaire avec des vecteurs propres renormalisés. Par conséquent, en plus des paramètres inhérents à notre méthode de validation comme la dimension de la projection utilisée pour définir A, lorsqu'on s'intéresse à la validation de variétés on a affaire à un ensemble de paramètres libres supplémentaires, en l'occurrence ν (qui fixe la taille du domaine de définition $\mathcal{B}_\nu$ de la paramétrisation f), et la norme des différents vecteurs propres. Il est important d'ajuster la valeur de ces paramètres de manière appropriée, dans la mesure où ils influencent la variété (i.e. l'objet géométrique $f(\mathcal{B}_\nu)$) qui est obtenue. Selon la situation dans laquelle on est, on peut vouloir une variété qui soit la plus grande possible, or alors qui aille le plus loin possible dans une direction donnée (ceci peut être particulière pertinent pour des systèmes lent-rapide).

Avant les travaux du chapitre 7, ces paramètres étaient choisis à la suite de longs et fastidieux tâtonnements basés sur des expérimentations numériques. Dans le chapitre 7 on présente une sorte de procédure automatisée pour sélectionner les valeurs de ces paramètres, qui peut être aisément adaptée en fonction du problème considéré. Cette technique est basée sur l'observation, simple mais cruciale suivante. Il est équivalent (au moins d'un point de vue théorique) de multiplier le rayon ν par un facteur $\gamma > 0$, ou de renormaliser les vecteurs propres par γ, c'est à dire de remplacer la condition (1.31) par

$$Df(0) = \begin{pmatrix} \gamma V_1 & \dots & \gamma V_k \end{pmatrix}.$$

Évidemment, si on veut obtenir la variété la plus grande possible il est naturel de prendre ν le plus grand possible, pour augmenter le domaine de définition de la paramétrisation. Cependant, en pratique il vaut mieux garder ν proche de 1, pour des raisons de stabilité numérique. Par conséquent, on fixe $\nu = 1$ et on change seulement la normalisation des vecteurs propres. Pour promouvoir une certaine direction dans la variété, on peut aussi considérer une renormalisation *anisotrope* $\gamma = (\gamma_1, \dots \gamma_k)$, associée à la condition

$$Df(0) = \begin{pmatrix} \gamma_1 V_1 & \dots & \gamma_k V_k \end{pmatrix}. \tag{1.32}$$

Dans le chapitre 7, on développe d'abord une méthode pour choisir cette renormalisation qui est uniquement basée sur un calcul numérique d'erreur, puis une autre (légèrement plus couteuse) qui peut être utilisée pour des validations a posteriori. Plus précisément, on établit des bornes Y, Z_1 et Z_2 comme définies dans la proposition 1.2.1 (pour la fonction F obtenue en ajoutant la condition (1.32) au système en les coefficients (a_α) représentant l'équation d'invariance (1.29)), de telle sorte que leur dépendance par rapport à $\gamma = (\gamma_1, \dots \gamma_k)$ soit explicite. En conséquence, dès que les bornes ont été calculées une fois, on peut les recalculer rapidement pour n'importe quelle renormalisation $\gamma = (\gamma_1, \dots \gamma_k)$.

A titre d'exemple, on applique cette méthode pour calculer des variétés stables et instables locales dans les systèmes de Lorenz et de FitzHugh-Nagumo, ainsi que pour l'équation du pont suspendu. Ce dernier exemple est étudié plus en détail dans le chapitre 9, où on utilise notre technique pour calculer et valider (uniformément en un paramètre β de l'équation) des variétés stables, qui sont ensuite utilisées pour démontrer l'existence d'orbites homoclines, pour tout β dans un certain intervalle (plus de détails dans la section 1.2.6).

1.2.5 Exposition des résultats obtenus dans le chapitre 8

Dans cette section, on présente les principaux résultats du chapitre 8 de cette thèse, qui traite de l'interpolation polynomiale dans le cadre du calcul rigoureux.

Dans l'article original de Yamamoto [214], ainsi que dans des travaux plus récents (voir par exemple [65, 61]), on utilise l'interpolation linéaire par morceaux pour résoudre des problèmes aux limites ou pour prouver l'existence de connexions homoclines/hétéroclines à l'aide des méthodes de validation a posteriori. L'interpolation linéaire par morceau, et plus généralement les méthodes d'éléments finis sont souvent plus souples d'utilisation que les méthodes spectrales,

dans la mesure où elles permettent de traiter des problèmes ayant des géométries complexes, tandis que les méthodes spectrales sont difficilement utilisables en dehors de domaines rectangulaires. De plus, l'interpolation polynomiale est mieux adaptée pour étudier des systèmes avec des termes non linéaires compliqués (non polynomiaux) [22]. Cependant, dans des situations où les deux types de techniques peuvent être utilisées, les méthodes spectrales sont substantiellement plus efficaces pour obtenir des validations a posteriori.

Plus précisément, supposons qu'on veuille utiliser des techniques de calcul rigoureux pour valider la solution d'un problème de Cauchy [23] pour une EDO

$$\begin{cases} u' = \phi(u) \quad \text{sur } [0,\tau] \\ u(0) = u_0, \end{cases} \tag{1.33}$$

avec un champ de vecteur polynomial ϕ. Avec les méthodes dont nous avons connaissance, l'orbite la plus longue pouvant être validée avec une méthode spectrale comme les séries de Tchebychev, est nettement plus longue que celle qu'on pourrait valider en utilisant une interpolation linéaire par morceaux. On pourrait immédiatement objecter que cette comparaison n'est pas équitable, dans la mesure où une interpolation linéaire par morceaux est bien plus grossière qu'une approximation à l'aide d'une série de Tchebychev. Cependant, notre discussion ne porte pas sur la qualité de l'approximation, mais seulement sur la possibilité de valider rigoureusement la solution numérique ainsi obtenue. Et bien que les deux soient liées, on fera clairement apparaître sur des exemples dans le chapitre 8 que la qualité de l'approximation n'est pas le facteur limitant dans ce cas, et qu'une simple augmentation de l'ordre d'approximation (par exemple l'utilisation d'interpolation quadratique ou cubique par morceaux) ne permet pas de valider de plus longues orbites.

Dans le chapitre 8, on présente une nouvelle technique, que l'on nomme *bootstrap a priori*, qui nous permet de mettre à profit des approximations polynomiales d'ordre plus élevées pour valider des orbites bien plus longues. Avant de décrire cette technique, on rappelle brièvement comment l'interpolation linéaire par morceaux était utilisée jusqu'alors dans le cadre du calcul rigoureux. Par souci de simplicité, on se restreint à l'étude du problème (1.33). Avec les notations utilisées dans la section 1.2.2, ce problème est du type $F = L + N = 0$ avec $Lu = u'$ and $N(u) = -\phi(u)$. Cependant, comme on l'a mentionné en présentant la méthode de Yamamoto, il est plus opportun de considérer la formulation $\tilde{I} + L^{-1}N = 0$, ce qui correspond à la reformulation intégrale de l'EDO

$$u(t) = u_0 + \int_0^t \phi(u(s))ds, \quad \forall\ t \in [0,\tau]. \tag{1.34}$$

On définit donc

$$\tilde{F}(u)(t) = u(t) - u_0 - \int_0^t \phi(u(s))ds, \quad \forall\ t \in [0,\tau],$$

et on cherche à prouver l'existence d'un zéro de $\tilde{F}$. La méthode de validation se base sur l'espace $\mathcal{X} = \mathcal{C}^0([0,\tau])$ des fonctions continues sur $[0,\tau]$ muni de la norme uniforme, et utilise la décomposition $\mathcal{X} = \mathcal{X}_h \oplus \mathcal{X}_\infty$, avec $\mathcal{X}_h$ le sous-espace des fonctions linéaires par morceaux associé à un maillage de pas h sur $[0,\tau]$. Autrement dit, $\mathcal{X}_h$ est le sous-espace constitué de toutes les fonctions u telles que $u_{|]jh,(j+1)h[}$ est linéaire, pour tout $0 \leq j \leq \frac{\tau}{h} - 1$ [24]. L'opérateur de point fixe pour lequel on veut utiliser le théorème du point fixe de Banach est alors exactement l'opérateur $\tilde{T}$ défini en (1.23). En fait, on lui applique une sorte de théorème de type Newton-Kantorovich, très similaire à la proposition 1.2.1, et on doit donc majorer (entre autres) des

22. On mentionne tout de même, que dans le cadre du calcul rigoureux avec des méthodes spectrales, des techniques inspirée de la différenciation automatique permettent parfois de traiter des nonlinéarités non polynomiales (voir [154] ainsi que le chapitre 10).

23. La comparaison reste valable pour la solution d'un problème aux limites, ou pour une orbite périodique.

24. où h est supposé choisi de telle sorte que $\frac{\tau}{h}$ soit un entier.

erreurs d'interpolation comme

$$(I - \Pi_h)(F(u) - u) = (I - \Pi_h)\left(t \mapsto \int_0^t \phi(u(s))ds\right).$$

C'est ce genre de terme qui constitue le facteur limitant dans les estimations, et il est *seulement* d'ordre h, ce qui provient du fait que pour une fonction continue u, la fonction $t \mapsto \int_0^t \phi(u(s))ds$ est seulement de classe $\mathcal{C}^1$ (des estimations précises et des explications détaillées sont données dans le chapitre 8). On voit déjà ici qu'une simple augmentation de l'ordre d'interpolation (par exemple en considérant pour $\mathcal{X}_h$ le sous-espace des fonctions quadratiques par morceaux, avec Π_h la projection associée de $\mathcal{X}$ sur $\mathcal{X}_h$), n'améliore pas la situation. En effet, pour une fonction de classe $\mathcal{C}^1$ et un pas h fixé, doubler le degré de l'interpolation polynomiale divise grosso modo par deux l'erreur d'interpolation (on réfère toujours au chapitre 8 pour des estimations précises), ce qui n'est pas plus efficace que de simplement prendre h deux fois plus petit sans changer le degré de l'interpolation polynomiale.

Le but de notre procédure de bootsrap a priori est de contourner cette limitation. Pour ce faire, on retourne à l'équation (1.33) qu'on dérive, pour obtenir un problème d'ordre deux équivalent (où on a remplacé u' par $\phi(u)$)

$$\begin{cases} u'' = D\phi(u)\phi(u) \quad \text{sur } [0,\tau] \\ u(0) = u_0, \\ u'(0) = \phi(u_0). \end{cases} \tag{1.35}$$

qu'on réintègre ensuite (deux fois), pour obtenir une autre reformulation intégrale [25]

$$u(t) = u_0 + t\phi(u_0) + \int_0^t (t-s)D\phi(u(s))\phi(u(s))ds, \quad \forall\ t \in [0,\tau].$$

Cette formulation est certes (légèrement) plus compliquée que (1.34), mais la différence cruciale est que pour une fonction continue u, la fonction $t \mapsto \int_0^t (t-s)D\phi(u(s))\phi(u(s))ds$ est maintenant de classe $\mathcal{C}^2$, ce qui permet de meilleures erreurs d'interpolation. On pourrait réitérer ce processus pour obtenir une fonction de classe $\mathcal{C}^3$, et ainsi de suite. D'une certaine manière, notre technique de bootstrap a priori permet d'incorporer dans la formulation intégrale une partie de la régularité qu'on sait a priori que la solution va avoir. C'est seulement après cette reformulation qu'il devient utile d'utiliser de l'interpolation d'ordre supérieur. En effet, une fois qu'on a suffisamment diminué l'ancien facteur limitant (qui est lié à la borne Z_1), la qualité de l'interpolation (i.e. la borne Y) commence à avoir de l'importance.

Dans le chapitre 8 on établit des estimations générales, pour utiliser les méthodes de validation a posteriori avec un nombre arbitraire p de bootstrap a priori et un ordre k d'interpolation quelconque (aux points de Tchebychev de seconde espèce). Cette technique de bootsrap a priori peut sembler naive, mais les améliorations quelle permet dans le cadre du calcul rigoureux sont considérables. Une étude comparative détaillée est menée dans le chapitre 8, dont on se contente de mentionner un exemple ici, pour le système de Lorenz. Pour un u_0 donné et une dimension fixée du sous-espace X_h, le temps τ maximal pour lequel on peut valider la solution est de 0.7 sans bootstrap a priori, mais passe à 5.6 en utilisant une fois la reformulation du bootstrap a priori, et à 8.1 en l'utilisant une deuxième fois. En particulier, en combinant cette méthode avec la technique de maximisation des variétés locales développée dans le chapitre 7, on a été capable de valider une connexion hétérocline pour le système de Lorenz avec les paramètres standard, qui n'avait pas pu être validée sans bootstrap a priori dans une étude antérieure [155].

25. qu'on aurait également pu obtenir à partir de (1.34) en utilisant une intégration par partie adéquate.

On est également parvenu à valider des solutions spécifiques (appelées orbites balistiques en spirales) pour les flots de type ABC [55, 112, 126], où le champ de vecteur est donné par

$$\phi_{A,B,C}(x,y,z) = \begin{pmatrix} A\sin(z) + C\cos(y) \\ B\sin(x) + A\cos(z) \\ C\sin(y) + B\cos(x) \end{pmatrix}, \quad A, B, C \in \mathbb{R}.$$

Ce système est un exemple typique où l'utilisation de méthodes spectrales est délicate dans le cadre des techniques de validation a priori, à cause des termes non polynomiaux, mais où on peut utiliser l'interpolation polynomiale combinée au bootstrap a priori.

1.2.6 Exposition des résultats obtenus dans le chapitre 9

Dans cette section, on présente les principaux résultats du chapitre 9 de cette thèse, qui traite de l'existence d'ondes progressives pour l'équation du pont suspendu.

Motivés par un ancien rapport d'observation du pont du Golden Gate durant une tempête en 1938, Chen et McKenna ont initié l'étude mathématique de l'existence d'ondes progressives dans un pont suspendu [97]. Ils ont considéré l'équation

$$\partial_{tt}U + \partial_{xxxx}U + (1+U)^+ - 1 = 0, \tag{1.36}$$

où $U^+ := \max(U, 0)$. Ici $U(t,x)$ désigne la hauteur de la chaussée du pont au temps t et à la position x (x étant une variable unidimensionnelle décrivant la direction du trafic), la hauteur $U = 0$ décrivant le pont au repos. Chen et McKenna ont démontré l'existence d'ondes progressives pour cette équation, en les calculant explicitement. Plus précisément, ils s'intéressent à des solutions de la forme

$$U(t,x) = u(x - ct), \tag{1.37}$$

et telles que le profil u tend exponentiellement vers 0 lorsque la variable indépendante $\xi = x - ct$ tend vers $\pm\infty$. En utilisant l'ansatz (1.37) dans l'équation du pont (1.36), on doit en fait prouver l'existence de connexions homoclines pour l'EDO

$$u'''' + \beta u'' + (1+u)^+ - 1 = 0, \tag{1.38}$$

où $\beta = c^2$. En tirant parti du fait que cette équation est linéaire par morceaux, Chen et McKenna ont construit explicitement des connections homoclines pour tout $\beta \in [\beta_1, \beta_2]$, où $0 < \beta_1 < \beta_2 < 2$. Dans une étude ultérieure [98], ils développent une méthode basée sur le théorème du col qui leur permet de démontrer l'existence de connexions homoclines pour (1.38) pour tout β dans $]0, 2[$. Cependant, des observations numériques leur suggèrent qu'il serait plus pertinent de considérer un modèle légèrement plus sophistiqué (avec un terme non linéaire plus régulier), donné par

$$\partial_{tt}U + \partial_{xxxx}U + e^U - 1 = 0. \tag{1.39}$$

Toujours en utilisant l'ansatz (1.37), l'étude d'ondes progressives pour cette nouvelle équation conduit à l'EDO

$$u'''' + \beta u'' + e^u - 1 = 0. \tag{1.40}$$

La méthode développée par Chen et McKenna s'applique pour certaines equations de la forme

$$u'''' + \beta u'' + g(u) - 1 = 0,$$

avec un terme non linéaire g pouvant être plus général que $g(u) = (1+u)^+$, mais elle ne permet pas de traiter (1.40). Toutefois, Chen et McKenna ont étudié numériquement l'existence d'orbites homoclines pour (1.40), et les résultats obtenus les ont amené à conjecturer l'existence d'orbites homoclines pour (1.40) pour tout $\beta \in]0, 2[$.

Cette conjecture a été partiellement résolue par van den Berg et Smets [68], qui ont démontré l'existence d'orbites homoclines pour (1.40), pour *presque tout* β dans $]0,2[$, à l'aide de méthodes variationnelles. Il a ensuite été démontré [193], qu'une orbite homocline existait pour tout β dans $]0,\beta^*]$, où $\beta^* \simeq 0.5516$.

Dans le chapitre 9, on utilise les techniques du calcul rigoureux pour compléter (partiellement) ces résultats en démontrant le théorème suivant.

Theorem 1.2.4. *Pour tout β dans $[0.5, 1.9]$, il existe une orbite homocline pour* (1.40).

Avant de donner quelques détails sur la manière dont on prouve ce résultat, on mentionne que des méthodes de preuves assistées par ordinateur ont déjà été utilisées pour étudier l'équation du pont suspendu : la méthode de Plum a été appliquée [80] pour prouver l'existence de 36 orbites homoclines différentes pour une valeur de β donnée (en l'occurrence $\beta = 1.69$), répondant ainsi positivement à une autre quesion soulevée par Chen et McKenna dans l'article [98].

Notre preuve est partiellement basée sur les travaux du chapitre 7, et on utilise les méthodes qui y sont développées pour calculer et valider de manière efficiace la variété stable locale de l'origine [26] pour (1.40). La partie finie de l'orbite arrivant jusqu'à la variété stable est ensuite calculée et validée en utilisant des séries de Tchebychev. On mentionne également que la démonstration de notre théorème requiert une généralisation de la proposition 1.2.1. En effet, telle qu'elle est énoncée dans la section 1.2.2, la proposition 1.2.1 peut être appliquée (pour valider la variété stable locale et la partie finie de l'orbite) seulement pour une valeur de β à la fois. Pour pouvoir traiter des intervalles de valeurs de β, on utilise dans le chapitre 9 une autre version de la proposition 1.2.1, obtenue en utilisant une version à paramètre du théorème du point fixe de Banach. Cette généralisation n'est pas nouvelle dans le contexte du calcul rigoureux, voir [64, 79, 53].

1.2.7 Exposition des résultats obtenus dans le chapitre 10

Dans cette section, on présente les principaux résultats du chapitre 10 de cette thèse. Dans ce dernier chapitre, on étudie les états d'équilibre d'un système de diffusion croisée utilisé en dynamiques des populations, qui a été introduit par Shigesada, Kawasaki and Teramoto [194]

$$\begin{cases} \partial_t u = \partial_{xx}\left((d_1 + d_{12}v)u\right) + (r_1 - a_1 u - b_1 v)u & \text{sur } \mathbb{R}_+ \times (0,1), \\ \partial_t v = d_2 \partial_{xx} v + (r_2 - b_2 u - a_2 v)v & \text{sur } \mathbb{R}_+ \times (0,1), \\ \partial_x u(t,0) = 0 = \partial_x u(t,1) & \text{sur } \mathbb{R}_+, \\ \partial_x v(t,0) = 0 = \partial_x v(t,1) & \text{sur } \mathbb{R}_+. \end{cases} \tag{1.41}$$

Ici $u = u(t,x)$ et $v = v(t,x)$ représentent la densité de population de deux espèces, au temps t et à la position x. Les termes de réactions sont ceux du traditionnel modèle de Lotka-Volterra, les coefficients positifs r_i, a_i et b_i $(i = 1,2)$ décrivant respectivement le taux de croissance intrinsèque, le taux de compétition à l'intérieur même d'une espèce et le taux de compétition entre espèces différentes. Le terme qui distingue ce modèle d'un système de réaction diffusion classique est $d_{12}\partial_{xx}(vu)$, qui modélise que le taux de diffusion de l'espèce u est affecté par la présence de l'espèce v.

Ce terme a été ajouté aux équations de Lotka-Volterra avec diffusion par Shigesada, Kawasaki et Teramoto pour prendre en compte le phénomène de *ségrégation spatiale*. En effet, des observations dans la nature suggèrent que des espèces en compétition peuvent coexister en se répartissant de manière à s'éviter, formant des configurations (*pattern*) spécifiques. D'un point de vue mathématique, cette situation se traduit par l'existence d'états stationnaires non

26. A l'aide d'un argument de symétrie, on se ramène à l'étude d'une demi-orbite, i.e. pour $t \geq 0$, et la variété instable n'est donc pas nécessaire.

homogènes, et stables. Cependant, de telles solutions ne peuvent pas exister pour les équations de Lotka-Volterra avec diffusion (i.e. le système (1.41) avec $d_{12} = 0$, plus de détails dans le chapitre 10), ce qui justifie l'introduction d'un modèle plus complexe.

Depuis l'article original de Shigesada, Kawasaki et Teramoto en 1979, les états d'équilibre du système (1.41) ont été l'objet de nombreuses études, tant numériques que théoriques (voir par exemple [137] et les références qui y sont mentionnées, ainsi que l'introduction du chapitre 10). Les études numériques suggèrent qu'en fonction des valeurs des différents paramètres, le système (1.41) possède une grande variété d'états d'équilibre non homogènes (voir par exemple la figure 1.3). Cependant, les résultats théoriques existants ne permettent d'obtenir qu'une fraction des ces solutions observées numériquement (principalement celles qui sont proches d'une bifurcation depuis un état d'équilibre homogène, ou pour des valeurs asymptotiques de paramètres). Il en est de même pour les questions relatives à la stabilité de ces états stationnaires.

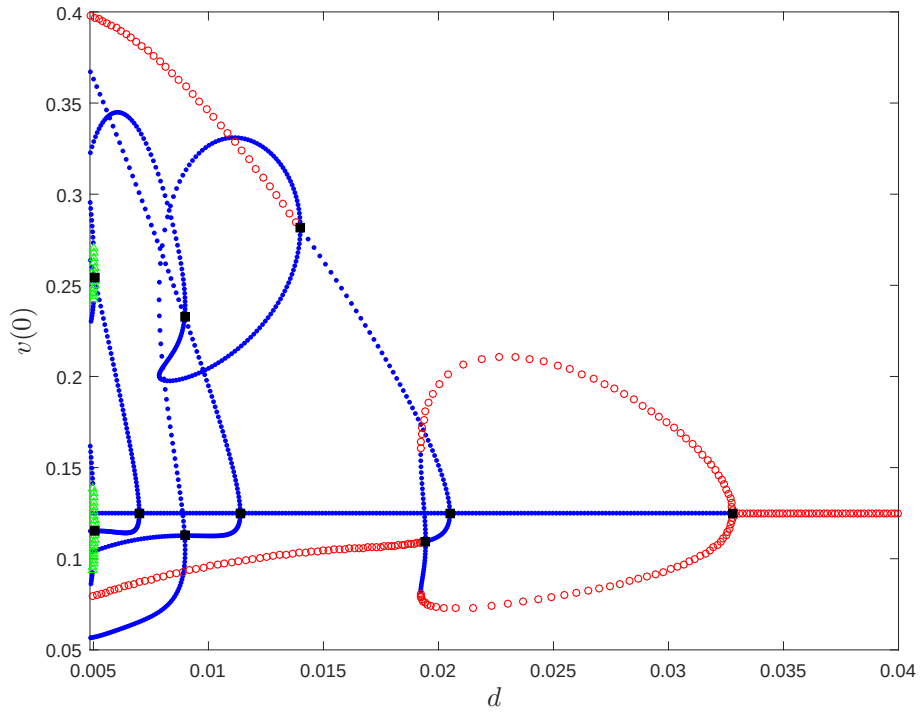


FIGURE 1.3 – Un diagramme de bifurcation d'états d'équilibre validés pour (1.41), avec $r_1 = 5$, $r_2 = 2$, $a_1 = 3$, $a_2 = 3$, $b_1 = 1$, $b_2 = 1$, $d_{12} = 3$ et $d_{21} = 0$, $d_1 = d_2 = d$ étant le paramètre de bifurcation. Chaque point bleu correspond à un état d'équilibre validé, dont l'instabilité a été prouvée. Chaque triangle vert correspond à un état d'équilibre validé, qui semble instable mais pour lequel la preuve de l'instabilité n'a pas pu être effectuée. Chaque cercle rouge correspond à un état d'équilibre validé, qui semble stable [28].

Dans le chapitre 10, on utilise les techniques de validation a posteriori pour étudier les états stationnaires du système (1.41). Les conditions aux limites homogènes de Neumann permettent d'utiliser une expansion en séries de Fourier. Cependant, à cause du terme de diffusion croisée, on n'obtient pas directement un terme dominant linear diagonal (ou même tridiagonal) comme dans la section 1.2.3. Par conséquent, on commence par utiliser un changement de variable $w = (d_1 + d_{12}v)u$ avant d'appliquer notre technique de validation a posteriori. Une fois les états d'équilibre validés, on s'intéresse à leur stabilité linéaire. On est capable de prouver qu'un grand nombre de ces états d'équilibre sont en fait instables, en calculant puis en validant une

28. On note que la droite de solutions pour $v(0) = 0.125$ correspond à des états d'équilibre homogènes, pour lesquels l'étude de la stabilité pourrait s'effectuer de manière analytique.

valeur propre de partie réelle strictement positive pour le système linéarisé autour de l'état d'équilibre (non homogène). Malheureusement, pour ce problème on n'est pas encore capable d'étendre cette technique pour prouver la stabilité de certains états d'équilibre (mais les calculs numériques suggèrent que certains de états d'équilibre obtenus sont bien stables). Les résultats obtenus sont résumé dans le théorème qui suit.

Theorem 1.2.5. *Dans la figure 1.3, chaque point bleu, triangle vert et cercle rouge correspond à un état d'équilibre validé du système de diffusion croisée* (1.41)*, avec* $r_1 = 5$*,* $r_2 = 2$*,* $a_1 = 3$*,* $a_2 = 3$*,* $b_1 = 1$*,* $b_2 = 1$*,* $d_{12} = 3$*,* $d_{21} = 0$ *et* $d_1 = d_2 = d$*. De plus, chaque solution représentée par un point bleue est linéairement instable.*

En particulier, pour $d = 0.005$ *il existe* 13 *états d'équilibre différents, dont* 11 *sont instables.*

1.2.8 Conclusion et perspectives

Dans le chapitre 6, on a élargi le champ d'application des méthodes spectrales pour la validation a posteriori, qui inclut désormais des situations avec un *terme dominant linéaire tridiagonal*. Dans le chapitre 7, on a facilité l'utilisation de la méthode de la paramétrisation pour valider a posteriori des variétés invariantes, en obtenant des estimations qui permettent d'adapter la renormalisation des vecteurs propres pour optimiser la variété obtenue. Dans le chapitre 8, on a amélioré de manière significative l'efficacité des méthodes de validation a posteriori basées sur l'interpolation polynomiale, en introduisant une technique de *bootstrap a priori*. Dans le chapitre 9 on a combiné la méthode introduite au chapitre 7 avec l'utilisation de séries de Tchebychev dans le contexte du calcul rigoureux, pour prouver partiellement une conjecture de Chen et McKenna à propos de l'existence d'ondes progressives pour l'équation du pont suspendu. Enfin, dans le chapitre 10, on a appliqué les techniques de validation a posteriori au système de diffusion croisée SKT, pour prouver l'existence de nombreux états d'équilibre non homogènes ainsi que l'instabilité d'un certain nombre d'entre eux.

On termine cette section introductive en présentant quelques extensions naturelles suggérées par les résultats de cette thèse, ainsi que quelques questions qui restent ouvertes.

- Une des possibles éxtensions des résultat du chapitre 8, qui est en fait la motivation principale derrière ce projet, serait de parvenir à combiner l'interpolation polynomiale (en temps) avec une méthode spectrale comme les séries de Fourier (en espace, pour des conditions aux limites appropriées), pour développer des techniques de validation a posteriori pour les EDP paraboliques. Les seuls résultats existants (avec les techniques de validation a posteriori utilisées dans cette thèse) pour les EDP paraboliques concernent l'existence de solutions périodiques pour lesquelles (avec des conditions aux limites adaptées) on peut utiliser des expansions en séries de Fourier pour toutes les variables (voir par exemple [122]). Cependant, pour certains problèmes aux limites en temps, par exemple pour obtenir des connexions homoclines, les séries de Fourier ne sont plus utilisables. D'un point de vue théorique, il semble qu'on pourrait utiliser l'interpolation polynomiale en temps dans ces situations, mais l'efficacité jusque là médiocre de cette technique dans le cadre des méthodes de validation a posteriori a découragé les tentatives dans cette direction. Avec les améliorations apportées par le bootstrap a priori, cette piste parait désormais prometteuse.
- Pour prouver complètement la conjecture de Chen et McKenna à propos des ondes progressives pour l'équation du pont suspendu, il reste à démontrer l'existence de connexions homoclines pour l'équation (1.40) pour tout $\beta \in]1.9, 2[$. La technique utilisée dans le chapitre 9 pourrait fournir l'existence de connexions légèrement au delà de $\beta = 1.9$, mais une technique différente est requise pour atteindre 2. En effet, lorsque β tend vers 2, la partie réelle des valeurs propres tend vers 0 (on a affaire à une bifurcation de Hamilton-Hopf), et notre formulation $F = 0$ pour trouver une paramétrisation de la variété stable devient singulière. Un projet intéressant serait d'étudier en détail cette bifurcation afin de démonter l'existence

d'un voisinage à gauche (explicite) de $\beta = 2$ pour lequel une connexion homocline existe. On pourrait alors tenter d'atteindre ce voisinage avec la technique du chapitre 9, pour ainsi compléter la preuve de la conjecture.

- Dans le chapitre 10, on a prouvé l'existence de nombreux états d'équilibre non homogènes, dont certains semblent linéairement stables, mais on n'a pas été capable de démontrer cette stabilité. Cependant, cette propriété est cruciale du point de vue biologique, dans la mesure où les solutions stables sont celles qu'on s'attend à observer en pratique. Par conséquent, une étude plus poussée des propriétés de stabilité de ces états d'équilibre serait particulièrement intéressante. On mentionne que des techniques de validation a posteriori ont déjà été appliquées pour obtenir des résultats de stabilité (voir par exemple [202, 96]), mais elles sont plus délicates à utiliser ici à cause du terme de diffusion croisée. Le changement de variable $w = (d_1 + d_{12}v)u$ a permis de contourner ces difficultés pour l'étude des états stationnaires (i.e. lorsque $\partial_t u = 0 = \partial_t v$), mais il ne semble pas simplifier la situation lorsque qu'on considère l'équation dépendant du temps.

Chapter 2

Introduction in english

The subject of this thesis falls within the broad framework of partial differential equations (PDE) and dynamical systems, and is composed of two independent parts (Part II and Part III), which are introduced separately in Section 2.1 and Section 2.2.

In Part II, general existence and regularity results as well as some qualitative properties (such as mass conservation) are established for a class of PDEs called the discrete coagulation-fragmentation equations with diffusion.

In Part III, we also establish existence results for some PDEs and ordinary differential equations, but the results have a different flavor. Indeed, we do not get general result but focus on specific equations, with given parameter values, and very precise results for some specific solutions, such as steady states or connecting orbits, by computer-assisted methods.

2.1 Introduction of Part II: Moments estimates for the discrete coagulation-fragmentation equations with diffusion

In this section we introduce the results obtained in Part II of this thesis, which is about the discrete coagulation-fragmentation equations with diffusion. We begin by giving a brief presentation of coalescence and fragmentation processes and of their possible mathematical descriptions in Section 2.1.1. In Section 2.1.2, we then describe more precisely the kind of model that is studied in this thesis. Next, we discuss in Section 2.1.3 the issue of mass conservation (or absence thereof, that is appearance of gelation) which is a crucial matter for coagulation-fragmentation equations. In Section 2.1.4 we explain why moments estimates can be very useful for the study of these equations and we expose in Section 2.1.5 the central tool that we use to obtain such estimates. We then present in Section 2.1.6 the main results obtained in Part II of this thesis, and finally we conclude this introduction in Section 2.1.7 by describing some further problems and open questions that lie in the continuation of this thesis.

2.1.1 About coalescence and fragmentation processes

Coalescence and fragmentation processes play a very important role in physics, chemistry and biology, to describe phenomena where particles or clusters aggregate together to form bigger clusters, or break apart into smaller ones. For instance, one could think of hematology, and more precisely of the mechanism of blood coagulation. Another example would be the break-up of droplets in a spray. At a very different scale, these processes are also used to try and explain the formation of galaxies. For a more exhaustive list see the surveys [18, 30] and the references therein.

From a mathematical perspective, coalescence and fragmentation processes can be studied at three different levels.

- The *microscopic* description considers a finite set of particles that interact with each other in a stochastic way. This kind of models originate from the work of Smoluchowski [42, 43]. In the same vein, we also mention the Marcus-Lushnikov process [31, 32].
- When the number of particles is sufficiently high, the *mesoscopic* description can be considered, where instead of tracking each particle individually, a statistical representation of the system is used. This means that the particles are classified, for instance by size, and that the focus is on the evolution of the number (or density) of particle of each size. This leads to deterministic equations, such as the so called *Smoluchowski equations* (also introduced in [42, 43]), which are studied in this thesis.
- The third level is the *macroscopic* description, which focuses on physically observable quantities such as the total mass, that usually are average of variables of the mesoscopic description.

We point out that these three descriptions are not unrelated. In particular, a rigorous derivation of the mesoscopic description from the microscopic one can sometimes be obtained through convergence of stochastic processes or mean field limits (see for instance [35], for the convergence of the Marcus-Lushnikov process to the Smoluchowski equations). For some connection between the mesoscopic and the macroscopic descriptions, through fast-reaction limits, we refer to [13, 19].

2.1.2 Different mesoscopic models

In Part II of this thesis, we study a particular mesoscopic description given by $c = (c_i)_{i\in\mathbb{N}^*}$, where c_i denotes the concentration of clusters of size i, for all i in $\mathbb{N}^*$. We assume that the aggregation process follows the *law of mass action*, which states that the rate of a chemical reaction is proportional to the product of the concentrations of the reactants. Therefore, the rate of occurrence of the reaction in which a cluster of size i and a cluster of size j merge to form a cluster of size $i+j$ is of the form $a_{i,j}c_ic_j$, where $a_{i,j}$ is a proportionality coefficient. Concerning the fragmentation process, we assume that the break up of a cluster is self induced and depends only on the size of the cluster. Therefore the rate of occurrence of a fragmentation of a cluster of size i is of the form B_ic_i, where B_i is another proportionality coefficient. To describe the result of such a fragmentation, we introduce coefficients $\beta_{i,j}$ denoting the average number of clusters of size $j < i$ produced when a cluster of size i breaks up. The coagulation and fragmentation processes are schematized in Figure 2.1. The fragmentation of one cluster into smaller pieces

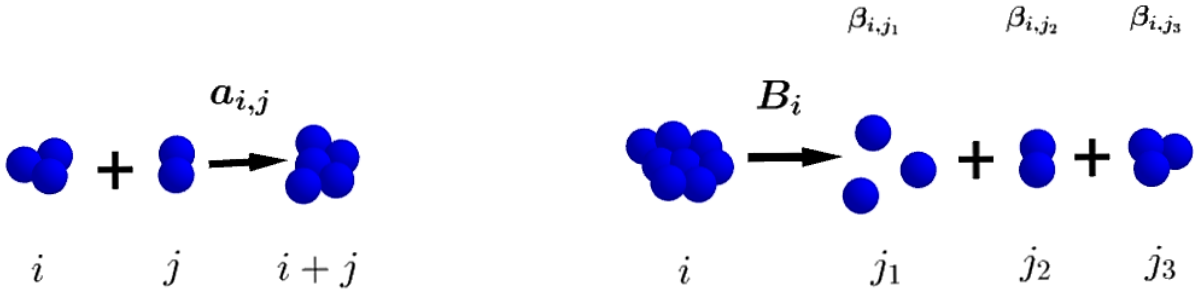


Figure 2.1 – An example of coagulation and fragmentation process. Here $\beta_{8,1} = 3$, $\beta_{8,2} = 1$, $\beta_{8,3} = 1$ and $\beta_{i,j} = 0$ for $4 \le j \le 7$.

should conserve mass, clusters of size 1 should not break up further and the coagulation rates should be symmetric, leading to the following set of natural hypothesis

$$i = \sum_{j=1}^{i-1} j\beta_{i,j}, \quad B_1 = 0, \quad a_{i,j} = a_{j,i} \quad \text{and} \quad a_{i,j}, B_i, \beta_{i,j} \ge 0, \quad \forall\, i,j \in \mathbb{N}^*. \tag{2.1}$$

Finally, we do not assume that the concentrations c_i have to be spatially uniform, but we allow them to vary depending on the position x. Therefore, each $c_i = c_i(t,x)$ is a function of time and space, and we assume that clusters of size i diffuse spatially with a coefficient d_i.

Let Ω be a smooth bounded domain of $\mathbb{R}^N$, in which the clusters are supposed to be confined, and $(c_i^{in})_{i\in\mathbb{N}^*}$ be the initial concentrations. The equations corresponding to the above described diffusive coagulation-fragmentation process are the following

$$\begin{cases} \partial_t c_i - d_i \Delta_x c_i = Q_i(c) + F_i(c), & \text{on } [0,T]\times\Omega, \\ \nabla_x c_i \cdot \nu = 0 & \text{on } [0,T]\times\partial\Omega, \qquad \forall i \in \mathbb{N}^* \\ c_i(0,\cdot) = c_i^{in} & \text{on } \Omega, \end{cases} \tag{2.2}$$

where $\nu(x)$ denotes the outward unit normal vector at point $x \in \partial\Omega$, and the coagulation term $Q_i(c)$ [1] [2] and the fragmentation term $F_i(c)$ respectively write:

$$Q_i(c) = Q_i^+(c) - Q_i^-(c) = \frac{1}{2}\sum_{j=1}^{i-1} a_{i-j,j} c_{i-j} c_j - \sum_{j=1}^{\infty} a_{i,j} c_i c_j, \tag{2.3}$$

$$F_i(c) = F_i^+(c) - F_i^-(c) = \sum_{j=1}^{\infty} B_{i+j}\beta_{i+j,i} c_{i+j} - B_i c_i. \tag{2.4}$$

The homogeneous Neumann boundary conditions are natural to model that no cluster enters nor leaves the domain Ω.

Before proceeding further, let us mention possible variations of this model. The most studied one is probably the model where only binary fragmentation is considered, that is any cluster can only break up into exactly two smaller clusters, and the rate of fragmentation of a cluster of size $i+j$ into a cluster of size i and a cluster of size j is then of the form $b_{i,j}c_{i+j}$. In that case, the fragmentation term is given by [3]

$$F_i(c) = F_i^+(c) - F_i^-(c) = \sum_{j=1}^{\infty} b_{i,j} c_{i+j} - \frac{1}{2}\sum_{j=1}^{i-1} b_{i-j,j} c_i.$$

Notice that we can recover the formulation (2.4) by considering coefficients B_i and $\beta_{i,j}$ defined as

$$B_i = \frac{1}{2}\sum_{j=1}^{i-1} b_{i-j,j}, \quad \beta_{i,j} = \frac{b_{i-j,j}}{B_i}, \quad \forall\ 1 \le j < i.$$

An even more specific situation can be considered, where both the coagulation and the fragmentation processes can only involve a cluster and a single particle (i.e. a cluster of size 1). This model was introduced by Becker and Döring [5] to describe nucleation, and has been studied quite extensively (see for instance the survey [41]).

Another possibility is to suppose that, instead of being self induced, the fragmentation of a cluster is due to its collision with another cluster. The occurrence rate of a fragmentation produced by the collision of a cluster of size j with a cluster of size k is then of the form

1. The $\frac{1}{2}$ factor in Q_i^+ comes from the fact that aggregation reactions are counted twice, for instance $1+(i-1)\to i$, and then $(i-1)+1\to i$. However, when i is even the reaction $\frac{i}{2}+\frac{i}{2}\to i$ does not appear twice. Therefore the coefficients $a_{i,i}$ should be taken as twice their *physical* value for (2.3) to faithfully describe the aggregation process.

2. We also point out that this term does not take into account the possible simultaneous aggregation of three or more clusters into a larger one. This is justified by the fact that this kind of aggregation is way less likely to happen than the aggregation of two clusters, because it requires all the concerned clusters to be at the same place at the same time.

3. with the same caveat for the definition of the coefficients $b_{i,i}$ as for the coefficients $a_{i,i}$.

$B_{j,k}c_jc_k$, and produces on average a number that we denote $\beta_{j,k,i}$ of clusters of size i. The fragmentation term is then given by

$$F_i(c) = F_i^+(c) - F_i^-(c) = \frac{1}{2}\sum_{j+k>i} B_{j,k}\beta_{j,k,i}c_jc_k - \sum_{j=1}^{\infty} B_{i,j}c_ic_j.$$

In a different direction, instead of considering a discrete description of the system given by $c = (c_i)_{i\in\mathbb{N}^*}$, we can consider a continuous description $c = c(y)$ for $y \in \mathbb{R}_+$. This model was first introduced by Müller [34]. The relevance of the discrete or the continuous description depends on the application at hand. For more details about the continuous model we again refer to the surveys [18, 30]. We also mention that the connections between the discrete and the continuous model have been studied (see [29] and the references therein).

2.1.3 Mass concervation and gelation

We now focus on the discrete coagulation-fragmentation equations (2.1)-(2.4) [4] and discuss some of the properties of the solution, especially the evolution of the total mass, given by $\sum_{i=1}^{\infty} ic_i$ [5]. The following weak formulation of the coagulation and fragmentation terms is going to be insightful. For any sequence $(\varphi_i)_{i\in\mathbb{N}^*}$, using (2.1) we have (at least formally)

$$\sum_{i=1}^{\infty}\varphi_iQ_i(c) = \frac{1}{2}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{i,j}c_ic_j(\varphi_{i+j} - \varphi_i - \varphi_j), \tag{2.5}$$

$$\sum_{i=1}^{\infty}\varphi_iF_i(c) = -\sum_{i=2}^{\infty} B_ic_i\left(\varphi_i - \sum_{j=1}^{i-1}\beta_{i,j}\varphi_j\right). \tag{2.6}$$

Taking $\varphi_i = i$ for all $i \in \mathbb{N}^*$, we notice that

$$\sum_{i=1}^{\infty} iQ_i(c) = 0 = \sum_{i=1}^{\infty} iF_i(c),$$

therefore, multiplying (2.2) by i and summing for all $i \in \mathbb{N}^*$, we get (still formally)

$$\partial_t\left(\sum_{i=1}^{\infty} ic_i\right) - \Delta_x\left(\sum_{i=1}^{\infty} id_ic_i\right) = 0. \tag{2.7}$$

After integrating over Ω and using the Neumann boundary conditions, we are left with

$$\frac{d}{dt}\int_\Omega \sum_{i=1}^{\infty} ic_i = 0,$$

which means that the total mass should stay constant. However, depending on the coagulation and fragmentation coefficients (i.e. the coefficients $a_{i,j}$, B_i and $\beta_{i,j}$), it can happen that the total mass becomes strictly less than the initial mass after a finite time. This phenomenon is not a mathematical artifact but can be explained and observed physically. It corresponds to a phase transition of the system, the lost mass being transferred to the newly created phase.

4. This restriction is mainly for the clarity of the exposition, and because the results of this thesis are almost exclusively concerned with this precise model. However, the considerations of this section are also relevant for the other models mentioned in Section 2.1.2.

5. To be precise, since each c_i denotes the concentration of clusters of size i, $\sum_{i=1}^{\infty} ic_i$ actually represents the total number of *elementary* particles by unit of volume (a cluster of size i being made of i elementary particles). Therefore $\sum_{i=1}^{\infty} ic_i$ is proportional to the total mass, the proportionality factor being the mass of an elementary particle divided by the total volume. For brevity, we still simply refer to $\sum_{i=1}^{\infty} ic_i$ as the total mass.

One example is the formation of colloidal gels in chemistry, which leads to this loss of mass being referred to as *gelation*. Mathematically, gelation occurs when some of the mass escapes as $i \to \infty$, which can be interpreted as the formation of clusters of infinite size.

To get a better understanding of this phenomenon, let us consider the particular case where $a_{i,j} = ij$ and there is no fragmentation ($B_i = 0$). We also assume that there is no diffusion ($d_i = 0$), so that we can consider homogeneous solutions depending only on t but not on x. Introducing the moments

$$\rho_k = \sum_{i=1}^{\infty} i^k c_i,$$

and using the weak formulation (2.5) with $\varphi_i = 1$, we get

$$\frac{d}{dt}\rho_0 = \frac{d}{dt}\left(\sum_{i=1}^{\infty} c_i\right) = -\frac{1}{2}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} ijc_ic_j = -\frac{1}{2}\left(\rho_1\right)^2. \tag{2.8}$$

Notice that ρ_0 is simply the total number of clusters [6], and ρ_1 the total mass, and that those are non negative quantities. Integrating (2.8) on $[0, t]$, we end up with

$$\rho_0(t) + \frac{1}{2}\int_0^t \left(\rho_1(s)\right)^2 ds = \rho_0(0).$$

In particular

$$\frac{1}{2}\int_0^t \left(\rho_1(s)\right)^2 ds \leq \rho_0(0), \quad \forall\ t \geq 0,$$

which means that there must exist a positive time t^*, for which the mass $\rho_1(t^*)$ is strictly less than the initial mass $\rho_1(0)$ [7]. The first time for which this occurs is called *gelation time*. A look at the second moment ρ_2 can provide some insight about the gelation time. Indeed, since gelation corresponds to the formation of infinite size clusters, it should create a blow up of higher order moments. Using again the weak formulation (2.5), this time with $\varphi_i = i^2$, we get

$$\frac{d}{dt}\rho_2 = \left(\rho_2\right)^2,$$

and thus the second moment ρ_2 does blow up, when t goes to $t_2 = \frac{1}{\rho_2(0)}$. In this specific case, it can in fact be proven that t_2 is the gelation time (for more details, as well as considerations about *blow up profiles*, we refer the reader to the surveys [16, 30] and the references therein).

On the other hand, in the case where $a_{i,j} = 1$ (still with $B_i = 0$, $F_i = 0$ and $d_i = 0$), we get

$$\frac{d}{dt}\rho_0 = -\frac{1}{2}\left(\rho_0\right)^2,$$

which, contrarily to (2.8), does not impose any restriction on the mass ρ_1, together with

$$\frac{d}{dt}\rho_2 = \left(\rho_1\right)^2,$$

so there is no blow up of the second order moment. And it can be proven that for this choice of parameters, the total mass is in fact conserved for all $t \geq 0$.

These two examples are coherent with the intuitive idea that gelation occurs if the clusters are aggregating faster and faster with each other to create larger and larger clusters, that is when the coagulation coefficients $a_{i,j}$ are growing fast enough with i and j. This has been proven to be true, the limiting case below which no gelation occurs being the one of *linear*

6. again, up to multiplication by the total volume.
7. If we assume that the initial mass is not equal to zero.

coagulation coefficients, given by $a_{i,i} = C(i + j)$. We refer to the introductions of Chapters 4 and 5 for precise mathematical statements and references. Another intuitive statement would be that even when the aggregation process is important, if the fragmentation process is also present and strong enough, then gelation can be prevented because large clusters break up as fast as they are created. This was also proven mathematically (again see the introduction of Chapters 5 for precise statements and references), but up to now only in the homogeneous case (i.e. when there is no diffusion).

2.1.4 Why moments estimates?

As suggested by the title of Part II, the segment of this thesis dedicated to the coagulation-fragmentation equations is mainly concerned about moments estimates. We now give two precise reasons why such estimates are crucial and useful.

The first one is that, as we just saw, moments ρ_k of the solutions are related to the gelation issue. More precisely, having bounds on a superlinear moment (i.e. a moment with $k > 1$) allows to prove that gelation cannot occur. The justification of this statement provides us with the perfect opportunity to make a slight detour through the theory of existence of solutions for the coagulation fragmentation equations with diffusion, which we have been purposely avoiding until now.

There are several marginally different approaches to prove existence of solutions of the coagulation fragmentation equation with diffusion, but they all rely on first considering a truncated version of (2.2), where only clusters of size less than n are taken into account. This truncated system only has a finite number of equations/unknown, and the r.h.s. given by $Q_i(c) + F_i(c)$ only contain finite sums (since c_i is assumed to be zero for $i > n$). Therefore standard existence theory for reaction diffusion systems can be used to get local (in time) existence of solutions, and some bounds using the precise structure of the equations can then be derived to prevent blow up and get global solutions (this is done in details in [45]). The next step is to consider a sequence of such solutions, when n goes to infinity, to use some compactness argument to extract a converging subsequence (this is where the methods can differ from one another, depending on the *a priori* estimates available), and finally to show that the obtained limit is a (weak) solution of the full coagulation fragmentation system (2.2).

The link with the gelation issue is that, for the solution of the truncated system, computations analogous to the ones made at the beginning of Section 2.1.3 can be done, but this time rigorously. Thus we can prove that the mass of the solution of the truncated system is equal to the initial truncated mass for all time t, and it suffices to be able to pass to the limit $n \to \infty$ in this equality to show that the mass is conserved for the solution of the full system. To justify this last step, which concerns the moment ρ_1, it is enough to have a bound (independent of the truncation number n) on a moment ρ_k of order $k > 1$ in a suitable functional space (this whole procedure is detailed in Chapter 5).

The second reason for which moment estimates are relevant, especially for coagulation-fragmentation equations *with* diffusion, is related to the regularity of the solutions. As we just pointed out, the usual existence theory for this kind of equations only provides weak solutions (of which a precise definition, following the approach of Laurençot and Mischler [28], can be found in Chapters 4 and 5). However, each concentration c_i solves a heat equation (2.2), which suggest that smooth strong solutions could be obtained. The main difficulty lies in getting estimates for the r.h.s. $Q_i(c) + F_i(c)$ of (2.2), which contains infinite sums. But as we saw on two specific examples in Section 2.1.3, these terms can be expressed in terms of moments of the solutions (or at least bounded by expressions involving moments, for more general coagulation and fragmentation coefficients). Therefore moments estimates may also allows to get regularity results for the solutions of the coagulation-fragmentation equations with diffusion.

2.1.5 The role of duality lemma

In Section 2.1.3, we derived some equations satisfied by moments ρ_0, ρ_1 and ρ_2, in the homogeneous case (that is where the diffusion coefficients d_i are equal to 0). However, we want to handle the more realistic model where spatial heterogeneity is taken into account and diffusion plays a role. Therefore we have to adapt the computations.

When the diffusion coefficients d_i are non zero, we recall that we still get the equality (2.7) about the mass ρ_1. Unfortunately, unless all the d_i are equal, (2.7) it is not *strictly speaking* a parabolic equation on ρ_1, since ρ_1 does not explicitly appear in the Laplacian term. However, we can rewrite (2.7) as

$$\partial_t \rho_1 - \Delta_x \left(M_1 \rho_1 \right) = 0,$$

where

$$M_1 = \frac{\sum_{i=1}^{\infty} i d_i c_i}{\sum_{i=1}^{\infty} i c_i}.$$

Similarly, still with non zero diffusion coefficients d_i but with $a_{i,j} = 1$ and $B_i = 0$, the computations of Section 2.1.3 on the second moment become

$$\partial_t \rho_2 - \Delta_x \left(\sum_{i=1}^{\infty} i^2 d_i c_i \right) = \left(\rho_1 \right)^2,$$

which we can rewrite as

$$\partial_t \rho_2 - \Delta_x \left(M_2 \rho_2 \right) = \left(\rho_1 \right)^2,$$

where

$$M_2 = \frac{\sum_{i=1}^{\infty} i^2 d_i c_i}{\sum_{i=1}^{\infty} i^2 c_i}.$$

Both for M_1 and M_2, the only natural estimates available are L^∞ bounds of the form

$$\inf_{i \in \mathbb{N}^*} d_i \leq M_1, M_2 \leq \sup_{i \in \mathbb{N}^*} d_i.$$

The question that is then raised is: what kind of estimate can we get for a nonnegative function u satisfying

$$\begin{cases} \partial_t u - \Delta_x (Mu) \leq f & \text{on } [0,T] \times \Omega, \\ \nabla_x u \cdot n = 0 & \text{on } [0,T] \times \partial\Omega, \end{cases} \tag{2.9}$$

where Ω is a smooth bounded domain, f lies in $L^p([0,T] \times \Omega)$ for some $p \in (1, \infty)$ and M is a measurable function satisfying $a \leq M \leq b$?

The answer is provided by a tool called *duality lemma* in the literature, which is attributed to Pierre and Schmitt [38], and was then further generalized by Cañizo, Desvillettes and Fellner [11]. We sketch here the general technique contained in this duality lemma, several variations tailored for specific cases are presented in details in Chapter 4.

The strategy is to consider a kind of *dual problem* associated to (2.9), defined as

$$\begin{cases} \partial_t v + M \Delta_x v = -\psi & \text{on } [0,T] \times \Omega, \\ \nabla_x v \cdot n = 0 & \text{on } [0,T] \times \partial\Omega \\ v(T, \cdot) = 0 & \text{on } \Omega, \end{cases} \tag{2.10}$$

where ψ is a non negative test function. Multiplying the equation from (2.9) by v (which is non negative by the maximal principle), and integrating over Ω and then over $[0,T]$, we get

$$\begin{aligned} v \partial_t u - v \Delta_x (Mu) &\leq fv \\ \partial_t (uv) - u \partial_t v - v \Delta_x (Mu) &\leq fv \\ \frac{d}{dt} \int_\Omega uv - \int_\Omega \left(u \left(\partial_t v + M \Delta_x v \right) \right) &\leq \int_\Omega fv \\ \int_0^T \int_\Omega u\psi &\leq \int_\Omega u(0) v(0) + \int_0^T \int_\Omega fv. \end{aligned}$$

Assume momentarily that the solution v of (2.10) satisfies

$$\|v(0)\|_{L^{p'}(\Omega)} \leq C \|\psi\|_{L^{p'}([0,T]\times\Omega)} \quad \text{and} \quad \|v\|_{L^{p'}([0,T]\times\Omega)} \leq C \|\psi\|_{L^{p'}([0,T]\times\Omega)}, \tag{2.11}$$

where, here and in the sequel, p' denotes the dual exponent of p (i.e. such that $\frac{1}{p} + \frac{1}{p'} = 1$). Then using Hölder's inequality we end up with

$$\int_0^T \int_\Omega u\psi \leq C \left(\|u(0)\|_{L^p(\Omega)} + \|f\|_{L^p([0,T]\times\Omega)} \right) \|\psi\|_{L^{p'}([0,T]\times\Omega)},$$

and since this holds for any non negative test function ψ, we get by duality that

$$\|u\|_{L^p([0,T]\times\Omega)} \leq C \left(\|u(0)\|_{L^p(\Omega)} + \|f\|_{L^p([0,T]\times\Omega)} \right).$$

It remains to be seen under which conditions (2.11) can be obtained.

It is rather easy to show (starting by multiplying (2.10) by $\Delta_x v$ and integrating over Ω), that (2.11) holds for $p = p' = 2$ as soon as $0 < a \leq b < \infty$, yielding an estimate in L^2 for u. This is the original result from Pierre and Schmitt. Cañizo, Desvillettes and Fellner then showed that (2.11) also holds for $p' \in (1, \infty)$, if besides $0 < a \leq b < \infty$ we assume that a and b are *close enough* to one another. To precise this closeness condition, we first have to introduce some notation.

Definition 2.1.1. *For $m > 0$ and $q \in]1, +\infty[$, we define $\mathcal{K}_{m,q} > 0$ as the best (i.e. the smallest) constant independent of $T > 0$ in the parabolic regularity estimate*

$$\left(\int_0^T \int_\Omega |\partial_t v|^q + m^q \int_0^T \int_\Omega |\Delta_x v|^q \right)^{\frac{1}{q}} \leq \mathcal{K}_{m,q} \left(\int_0^T \int_\Omega |f|^q \right)^{\frac{1}{q}}, \quad \forall\ f \in L^q([0,T]\times\Omega),$$

where v is the unique solution of the heat equation with constant diffusion coefficient m, homogeneous Neumann boundary conditions and zero initial data:

$$\begin{cases} \partial_t v - m\Delta_x v = f & \text{on } [0,T]\times\Omega, \\ \nabla_x v \cdot \nu = 0 & \text{on } [0,T]\times\partial\Omega, \\ v(0,\cdot) = 0 & \text{on } \Omega. \end{cases}$$

The existence of a such a constant $\mathcal{K}_{m,q} < \infty$ independent of the time $T > 0$ is explicitly stated in [27] provided that $\partial\Omega \in \mathcal{C}^{2+\alpha}$, $\alpha > 0$. If $0 < a \leq b < \infty$ and

$$\frac{b-a}{b+a} \mathcal{K}_{\frac{a+b}{2},p'} < 1, \tag{2.12}$$

then it is proved in [11] that (2.11) holds, and we get an estimate in L^p for u. We point out that this contains the *easier* case $p = p' = 2$ mentioned above, because one can show that $\mathcal{K}_{m,2} \leq 1$, and therefore (2.12) is always satisfied for $p' = 2$.

The technique described in this section is used extensively in this thesis to get L^p estimates for the moments ρ_k of the solution.

2.1.6 Exposition of the results of Chapters 4 and 5

In this section we present the main results obtained in Part II of this thesis.

The first one, which has been alluded to in the previous section, is an L^p estimate, for any $p \in (1, \infty)$, for the mass ρ_1 associated to solutions of the diffusive coagulation-fragmentation equations (2.1)-(2.4). It generalizes a result from [10] where an L^2 estimate was already obtained (using the original version of the duality lemma).

Proposition 2.1.2. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$ and $T > 0$. Assume that the coagulation and fragmentation coefficients satisfy*

$$\lim_{j\to\infty} \frac{a_{i,j}}{j} = 0 = \lim_{j\to\infty} \frac{B_{i+j}\beta_{i+j,i}}{i+j}, \quad \forall\ i \in \mathbb{N}^*, \tag{2.13}$$

and that the diffusion coefficients satisfy

$$0 < a := \inf_{i\geq 1} d_i, \qquad \text{and} \qquad b := \sup_{i\geq 1} d_i < \infty. \tag{2.14}$$

Let $p \in]1, +\infty[$ and assume also that the nonnegative initial data $c_i^{in} \geq 0$ have an initial mass ρ_1^{in} which lies in $L^p(\Omega)$, and that (2.12) *holds.*

Then, there exists a weak solution of the coagulation-fragmentation equations (2.1)-(2.4) *such that the mass ρ_1 lies in $L^p([0,T]\times\Omega)$.*

We point out that assumption (2.13) is only needed to get the existence of weak solutions (which in that case is guaranteed by the work of Laurençot and Mischler [28]).

If the coagulation coefficients are *sublinear*, we can then get a similar result for moments of higher order.

Theorem 2.1.3. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$, $N \leq 2$, and $T > 0$. Assume that the coagulation and fragmentation coefficients satisfy*

$$a_{i,j} \leq C_Q\left(i^\alpha j^\beta + i^\beta j^\alpha\right) \quad \text{and} \quad \lim_{j\to\infty} \frac{B_{i+j}\beta_{i+j,i}}{i+j} = 0, \quad \forall\ i \in \mathbb{N}^*,$$

where $C_Q > 0$, $\alpha, \beta \geq 0$ and $\alpha + \beta < 1$ and that the diffusion coefficients satisfy

$$d_i > 0,\ \forall\ i \in \mathbb{N}^* \quad \text{and} \quad d_i \underset{i\to\infty}{\longrightarrow} d_\infty > 0. \tag{2.15}$$

Assume also that the initial concentrations $c_i^{in} \geq 0$ each lies in $L^\infty(\Omega)$ and are such that, for some $k \in \mathbb{N}^*$ the initial moment ρ_k^{in} lies in $L^p(\Omega)$ for all $p \in (1,\infty)$.

Then, there exists a weak solution of the coagulation-fragmentation equations (2.1)-(2.4) *for which the moment ρ_k lies in $L^p([0,T]\times\Omega)$ for all $p \in (1,\infty)$.*

Remark 2.1.4. *The various hypothesis of this theorem are commented in detail in Chapters 4 and 5. We only point out here that the restriction $N \leq 2$ can be dropped if there is no fragmentation (i.e. $B_i = 0$ for all $i \in \mathbb{N}$), and more importantly that assumption* (2.15) *plays a crucial role, as it allows for* (2.12) *to be satisfied for any $p \in (1,\infty)$ if we only consider the concentrations c_i for i large enough. While this assumption might seem rather restrictive, it is in fact not much more stringent than simply assuming $\inf_{i\in\mathbb{N}^*} d_i > 0$* [8]*, which is an hypothesis that as been used in several previous studies. We come back to this point in Section 2.1.7.*

As explained in Section 2.1.4, Theorem 2.1.3 implies that no gelation can occur for sublinear coagulation coefficients. This fact has been known since the work of Ball and Carr [3] in the homogeneous case, and has been extended more recently by Cañizo, Desvillettes and Fellner in [10] to the case including diffusion. They obtained moments estimates weaker than Theorem 2.1.3 (only an L^1 estimate for a slightly superlinear moment), but that were already sufficient to prevent gelation. Concerning moments estimates for coagulation-fragmentation equations with diffusion, we also mention the work of Rezakhanlou [39, 40].

We give later (see Theorem 2.1.6) another application of Theorem 2.1.3 related to the smoothness of the solutions. But first, we consider a generalization of Theorem 2.1.3 where the coagulation can be superlinear, but counterbalanced by strong enough fragmentation.

8. Indeed, it is reasonnable to expect the sequence $(d_i)_{i\in\mathbb{N}^*}$ to be decreasing, since larger clusters should diffuse less. In that case the sequence $(d_i)_{i\in\mathbb{N}^*}$ is then automatically convergent since it is bounded.

Theorem 2.1.5. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$, $N \leq 2$, and $T > 0$. Assume also that the coagulation and fragmentation coefficients satisfy*

$$a_{i,j} \leq C_Q \left(i^\alpha j^\beta + i^\beta j^\alpha \right) \quad \text{and} \quad B_i \geq C_F i^\gamma,$$

where $C_Q, C_F > 0$, $0 \leq \alpha, \beta \leq 1$, $\gamma \geq 1$ and $\alpha + \beta < \gamma$, together with

$$\sup_{j \in \mathbb{N}^*} \frac{a_{i,j}}{j} < \infty \quad \text{and} \quad \sup_{j \in \mathbb{N}^*} \frac{B_{i+j}\beta_{i+j,i}}{i+j} < \infty, \quad \forall\ i \in \mathbb{N}^*. \tag{2.16}$$

Assume also that the diffusion coefficients satisfy (2.15)*. Finally, assume that c_i^{in} lies in $L^\infty(\Omega)$ for all $i \geq 1$, that for some $k > 1$ also satisfying $k > 2 - (\gamma - \alpha)$ and $k > 2 - (\gamma - \beta)$, ρ_k^{in} lies in $L^1(\Omega)$, and that ρ_1^{in} lies in $L^p(\Omega)$ for all $p < \infty$.*

Then, there exists a weak solution of the coagulation-fragmentation equations (2.1)-(2.4)*, whose moments satisfy*

$$\int_0^T t^{m-1} \int_\Omega \rho_{k+m(\gamma-1)}(t,x) dx dt < \infty, \quad \forall\ m \in \mathbb{N}^*. \tag{2.17}$$

We point out that (2.17) with $m = 1$ yields a bound in $L^1([0,T] \times \Omega)$ for a superlinear moment ρ_k, and thus implies that gelation cannot occur. The key assumption in Theorem 2.1.5 is $\gamma > \alpha + \beta$, which can be interpreted as fragmentation being strong enough compared to coagulation. As mentioned in Section 2.1.3, the fact that strong enough fragmentation can prevent gelation is not surprising, but it was only proved up to now in the homogeneous case (i.e. whitout diffusion), since the work of da Costa [14].

We now present the last of our results concerning coagulation-fragmentation equations with diffusion, where we use the moments estimates provided by Theorems 2.1.3 and 2.1.5 to prove the existence of smooth solutions.

Theorem 2.1.6. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$, $N \leq 2$, and $T > 0$. Assume that the coagulation and fragmentation coefficients satisfy*

$$a_{i,j} \leq C_Q \left(i^\alpha j^\beta + i^\beta j^\alpha \right) \quad \text{and} \quad B_i \leq C_{max} i^{\gamma_{max}},$$

where $0 \leq \alpha, \beta \leq 1$, $0 \leq \gamma_{max} < \infty$, and that the diffusion coefficients satisfy (2.15)*. Assume also that the initial data $c_i^{in} \geq 0$ are of class $\mathcal{C}^\infty(\overline{\Omega})$, compatible with the boundary conditions and that for all $k \in \mathbb{N}^*$ the initial moments ρ_k^{in} are of class $\mathcal{C}^\infty(\overline{\Omega})$. Finally, assume that we are in one of the two following cases.*

Sublinear coagulation case:

$$\alpha + \beta < 1 \quad \text{and} \quad \lim_{j \to \infty} \frac{B_{i+j}\beta_{i+j,i}}{i+j} = 0, \quad \forall\ i \in \mathbb{N}^*.$$

Strong fragmentation case:

There exists $C_F > 0$ and $\gamma \geq 1$, $\gamma > \alpha + \beta$, such that $B_i \geq C_F i^\gamma$, and (2.16) *holds.*

Then there exists a smooth solution of the coagulation-fragmentation equations (2.1)-(2.4) *such that each c_i is of class $\mathcal{C}^\infty([0,T] \times \overline{\Omega})$, and such that the moments ρ_k are also of class $\mathcal{C}^\infty([0,T] \times \overline{\Omega})$, for any $k \in \mathbb{N}^*$. Besides, if $\sup_{i \in \mathbb{N}^*} B_i < \infty$, the solution is unique.*

Again, we refer to Chapters 4 and 5 for detailed comments about the various hypotheses in this Theorem, and how they can possibly be relaxed.

2.1.7 Conclusion and perspectives

Under an assumption of convergence of the diffusion coefficients (2.15), we proved L^p estimates for all polynomial moments ρ_k associated to the solutions of the diffusive coagulation-fragmentation equations (2.1)-(2.4), provided that the coagulation coefficients are sublinear. Still under assumption (2.15), but this time in situations including superlinear coagulation coefficients, we proved similar L^p estimates (weighted in time) for polynomial moments ρ_k, provided that the fragmentation is strong enough compared to the coagulation. Up to our knowledge, this yields the first proof that strong enough fragmentation can prevent gelation for the diffusive coagulation-fragmentation equations. Finally, the above-mentioned moments estimates allow us to prove the existence of smooth solutions of the diffusive coagulation-fragmentation equations (2.1)-(2.4).

We finish this introductory section by presenting some natural extensions suggested by the results of this thesis, and remaining open questions.

- The hypothesis (2.15), that is used in all the main results presented in Section 2.1.6, in not fully satisfying from the physical point of view. Indeed, larger and larger clusters should diffuse less and less, therefore assuming that the diffusion coefficients d_i are converging seems legitimate, but the expected limit should be 0, and not a strictly positive value. It seems natural to try and extend the results of this thesis to the case where the diffusion coefficients d_i are converging to 0. We mention that some partial results in that direction were already obtained in [9, 23].
- Another interesting point concerns the possibility of gelation for the linear coagulation $a_{i,j} = C(i + j)$. It is known in the homogeneous case that gelation cannot occur in that case, but the question is still open when diffusion is taken into account.

2.2 Introduction of Part III: Validated numerics, theory and applications

In this section we introduce the results obtained in Part III of this thesis, which is about computed-assisted proofs for dynamical systems. We first give a brief review about this kind of methods in Section 2.2.1, and then describe more precisely in Section 2.2.2 the techniques based on a posteriori validation through a fixed point theorem, which are the ones used in this thesis. In Section 2.2.3, 2.2.4 and 2.2.5, we present some of the contributions of this thesis, which enlarge the framework of applicability of computer-assisted proofs for dynamical systems and improve their efficiency. In Section 2.2.6 and 2.2.7, we then present the rest of the contributions of this thesis, which this time are applications of computer-assisted proofs, to partially solve some open conjecture and study nonlinear systems that could not be completely investigated by purely analytical methods. Finally we sum up the obtained results and give some possible future directions of research in Section 2.2.8

2.2.1 About computer-assisted proofs in dynamics

To understand the global behavior of a nonlinear system, the first step is to study its invariant sets. Indeed, specific solutions like steady states, periodic orbits and connections between them are building blocks that organize the global dynamics. While there are many deep theoretical mathematical results about the existence of such solutions, it is often difficult to apply them to a specific example. Besides, when dealing with a precise application, it is often not only the existence of these solutions, but also their qualitative properties that are of interest. In that case, a powerful and widely used tool is numerical analysis, which is well adapted to the study of an explicit system and can provide insights for problems where the nonlinearities hinder the use of purely analytical techniques.

Going one step further, one can combine theoretical arguments and numerical simulations to get computer-assisted proofs. Such techniques have already been successfully used to settle several famous open problems, of which one of the most known example is probably the four color theorem [189]. For dynamical systems applications, we mention the first proof of the universality of the Feigenbaum constant [152], as well as the proof of the existence of the Lorenz strange attractor [200]. For a broader picture about the history and applications of computer-assisted techniques applied to dynamical systems, we refer to [63, 173, 201, 191, 177]. We just mention that the interest of these techniques to prove the existence of some specific solutions it at least twofold. The solutions in question could be interesting in their own right (see for instance Chapters 9 and 10) but not easily attainable via purely analytical techniques, or their existence could be a sufficient condition implying complex dynamics (in the spirit of *period 3 implies chaos*). A well known example in the context of ODEs is Shilnikov theorem [195], where the existence of an homoclinic orbit implies the existence of infinitely many periodic orbits (see also [62]). In a similar direction, we mention that computer-assisted techniques can be combined with topological methods such as the Morse-Conley theory, to obtain information on global dynamics [51].

In Part III of this thesis, we develop and apply computer-assisted techniques to study specific solutions and invariant manifolds of ordinary differential equations (ODEs) and partial differential equations (PDEs) [9]. Before going into the details of our own contribution, we give a brief description of the existing computer-assisted techniques in dynamics, sometimes referred to as *rigorous numerics* or *validated numerics* in the literature. We separate them in two categories.

- The first one, sometimes described as the *geometrical* approach, is based on rigorous enclosure of numerical solutions directly in phase space. Roughly speaking, the standard Euler method (or Runge-Kutta...) which gives a sequence of approximate values, is replaced by a an algorithm which gives a sequence of sets that are guaranteed to contain the true value of the solution at the associated time. A naive construction of such algorithm can result in sets whose size blows up very quickly, but there are astute techniques (like the Lohner algorithm [218]) allowing to avoid *wrapping effects* and making this techniques applicable. We refer to [86, 220, 125, 217, 106] for a more in depth exposition of this approach and examples of applications.
- The second one, sometimes called *functional analysis* approach, also provides a rigorous enclosure of solution, but this time in a carefully chosen function space and as the result of an *a posteriori* validation. Given a numerical solution, the strategy is to derive estimates allowing to use a fixed point theorem, to prove that a true solution exists in a neighborhood of the numerical one. We detail this approach in the next section, as the techniques developed and used in Part III of this thesis fall into this category.

In both cases, round off errors have to be controlled at some point to get rigorous mathematical statements. This can be done using an interval arithmetic package (in our case Intlab [190]).

2.2.2 A posteriori validation through fixed point theorems

The techniques described in this section aim at proving the existence of an isolated zero of a function F defined on a Banach space $\mathcal{X}$. We are going to suppose that F is of the form

$$F = L + N,$$

where L is a linear invertible operator having bounded and compact inverse and N is a nonlinear term, such that $L^{-1}N$ maps $\mathcal{X}$ into itself and is compact [10]. Given a numerical zero $\bar{x}$, the usual

9. Similar techniques can also be used for delay differential equations, but we do not explore this possibility in this thesis.

10. The precise method used in this thesis does not explicitly requires these hypothesis on L and N to be satisfied, at least not in this form, but these are the kind of functions F for which the method is expected to be successful (more details in Section 2.2.3).

strategy is to consider a Newton-like operator $T : \mathcal{X} \to \mathcal{X}$ of the form

$$T = I - AF, \tag{2.18}$$

and to apply Schauder's or Banach's fixed point theorem in a neighborhood of $\bar{x}$, to get (assuming A is injective) the existence of a true zero of F in this neighborhood. What then separates the different techniques is the choice of A, the fixed point theorem that is used, and the estimates that are actually used to apply that fixed point theorem.

In this thesis, the method that we mainly use was introduced by Day, Lessard and Mischaikow [109]. It consists in taking for A a well chosen approximate inverse of $DF(\bar{x})$ (more details in Section 2.2.3), and then to prove that Banach's fixed point theorem can be applied to T defined as in (2.18), in a neighborhood of $\bar{x}$. This technique is inspired by an earlier work of Yamamoto [214], which will be described in a moment. We also mention the work of Arioli and Koch [53], who independently developed a similar technique.

We point out that applying Banach's fixed point theorem to a Newton-like operator T defined as in (2.18) is in fact equivalent to applying a Newton-Kantorovich (see for instance [184, 188]) type theorem to F as is highlighted by the following statement.

Proposition 2.2.1. *Let $\mathcal{X}, \mathcal{Y}$ be Banach spaces, $F : \mathcal{X} \to \mathcal{Y}$ a $\mathcal{C}^1$ function and $A \in \mathcal{L}(\mathcal{Y}, \mathcal{X})$ injective. Let $\bar{x} \in \mathcal{X}$ and assume there exist non negative constants Y, Z_1 and a positive, non-decreasing, convexe and smooth function $r \mapsto Z_2(r)$ such that*

$$\|AF(\bar{x})\|_{\mathcal{X}} \leq Y \tag{2.19}$$

$$\|I - ADF(\bar{x})\|_{\mathcal{X}} \leq Z_1 \tag{2.20}$$

$$\|A(DF(x) - DF(\bar{x}))\|_{\mathcal{X}} \leq r Z_2(r) \quad \forall\ x \in \mathcal{B}_{\mathcal{X}}(\bar{x}, r), \tag{2.21}$$

where $\mathcal{B}_{\mathcal{X}}(\bar{x}, r)$ is the closed ball of $\mathcal{X}$, centered at $\bar{x}$ and of radius r. Define the radii polynomial [11] *P as*

$$P(r) = \frac{1}{2} Z_2(r) r^2 - (1 - Z_1) r + Y. \tag{2.22}$$

Assume there exists $r > 0$ such that $P(r) < 0$, and denote $\underline{r}$ and $\overline{r}$ the two non negative zeros of P, with $\underline{r} < \overline{r}$. Then F has a unique zero in $\mathcal{B}_{\mathcal{X}}(\bar{x}, r)$ for all r in the non empty interval $[r_{\min}, r_{\max})$, where $r_{\min} = \underline{r}$ and $r_{\max} = \min(\overline{r}, r^*)$, r^* being defined by

$$Z_2(r^*) r^* = 1 - Z_1.$$

This proposition is very reminiscent of the Newton-Kantorovich theorem, but can also be seen as a rather straightforward consequence of Banach's fixed point theorem (see the proof below). We point out that the key assumption, which in practice may or may not hold, is the existence of a positive r such that $P(r) < 0$. Let us briefly justify why it is reasonable to expect that such an r does exists, which then yields the existence of a locally unique zero of F. Assume for simplicity that Z_2 is constant. Then P is a quadratic polynomial, and the existence of a positive r such that $P(r) < 0$ is equivalent to having

$$Z_1 < 1 \quad \text{and} \quad 2 Y Z_2 < (1 - Z_1)^2.$$

The condition $Z_1 < 1$ should be satisfied if A is a *good enough* approximate inverse of $DF(\bar{x})$ and once such an A is fixed, the condition $2YZ_2 < (1 - Z_1)^2$ should also be satisfied if $\bar{x}$ is a *good enough* approximate zero of F, making Y really small. An important part of the method is thus to obtain good $\bar{x}$ and A, the other part being to actually derive tight enough bounds Y, Z_1 and $Z_2(r)$ to try and find a positive r such that $P(r) < 0$.

11. In practice, the function Z_2 is often polynomial (either because the non linear terms of the problem are themselves polynomials, or because we restrict ourselves to a neighborhood of $\bar{x}$, in which case Z_2 can be taken as a constant), which explains the terminology.

Remark 2.2.2. *This technique is computer-assisted at two different levels. First, the computer is of course very useful (and often absolutely necessary) to get a good approximate zero $\bar{x}$, and then to define A (which depends on $\bar{x}$). Second, the computer is also used to obtain the bounds Y, Z_1 and $Z_2(r)$, which are usually a combination of numerical quantities (depending on $\bar{x}$) and of pen-and-paper estimates (basically to control truncation errors, between the finite dimensional projection of $\mathcal{X}$ that must be used by the computer and $\mathcal{X}$ itself). The fact that* (2.19)-(2.21) *as well as the condition $P(r) < 0$ are inequalities highlights one of the advantages of the fixed point reformulation: rigorously checking* equalities *(such as $F(x) = 0$) on a computer is hard, whereas rigorously checking* inequalities *(such as the ones just mentioned) can be done rather easily by using interval arithmetic to control round-off errors.*

Remark 2.2.3. *There are now many works in the field of validated numerics that are based on Proposition 2.2.1, or on slight variations thereof. However, the factor $\frac{1}{2}$ in front of the Z_2 term is always lacking, which makes the result slightly less sharp than it could be*[12]*. We point out that this $\frac{1}{2}$ factor is in fact present in the usual Newton-Kantorovich theorem.*

Proof of Proposition 2.2.1. First notice that (thanks to the assumptions on Z_2) P is convex (for r positive) and that $P(0) = Y \geq 0$, therefore the existence of $r > 0$ such that $P(r) < 0$ implies the existence of exactly two non negative zeros of P, and thus $\underline{r}, \overline{r}$ are well defined.

We start by showing that T defined as in (2.18) maps the ball $\mathcal{B}_{\mathcal{X}}(\bar{x}, r)$ into itself for all $r \in [\underline{r}, \overline{r}]$. To do so, we consider $x \in \mathcal{B}_{\mathcal{X}}(\bar{x}, r)$ and estimate

$$\begin{aligned}
\|T(x) - \bar{x}\|_{\mathcal{X}} &\leq \|T(x) - T(\bar{x})\|_{\mathcal{X}} + \|T(\bar{x}) - \bar{x}\|_{\mathcal{X}} \\
&\leq \|x - \bar{x} - A(F(x) - F(\bar{x}))\|_{\mathcal{X}} + Y \\
&\leq \|I - ADF(\bar{x})\|_{\mathcal{X}} \|x - \bar{x}\|_{\mathcal{X}} + \|A(F(x) - F(\bar{x}) - DF(\bar{x})(x - \bar{x}))\|_{\mathcal{X}} + Y \\
&\leq Z_1 r + \frac{1}{2} Z_2(r) r^2 + Y \\
&\leq r,
\end{aligned}$$

where the last inequality is exactly $P(r) \leq 0$ and holds for all $r \in [\underline{r}, \overline{r}]$.

The next step is to show that T is a contraction on $\mathcal{B}_{\mathcal{X}}(\bar{x}, r)$, for all $r \in [r_{\min}, r_{\max})$. The proposition then follows from Banach's fixed point theorem. Using that $DT(x) = I - ADF(x)$ and introducing $DF(\bar{x})$ as in the previous computation, we get that

$$\|DT(x)\|_{\mathcal{X}} \leq Z_1 + Z_2(r) r, \quad \forall\, x \in \mathcal{B}_{\mathcal{X}}(\bar{x}, r).$$

Thus, for T to be a contraction on $\mathcal{B}_{\mathcal{X}}(\bar{x}, r)$, in addition to the condition $r \in [\underline{r}, \overline{r}]$ we need r to be such that $Z_1 + Z_2(r) r < 1$, which is equivalent to $r < r^*$ (using again that Z_2 in non decreasing). Therefore T is indeed a contraction on $\mathcal{B}_{\mathcal{X}}(\bar{x}, r)$, for all $r \in [r_{\min}, r_{\max})$.

It only remains to be shown that $\underline{r} < r^*$, to ensure that $[r_{\min}, r_{\max})$ is not empty. Notice that if Z_2 was constant, then r^* would be the apex of the quadratic polynomial P and thus it would indeed be strictly between $\underline{r}$ and $\overline{r}$. To treat the general case of a non constant Z_2, we introduce the quadratic polynomial

$$\underline{P}(r) = \frac{1}{2} Z_2(\underline{r}) r^2 - (1 - Z_1) r + Y.$$

Since Z_2 is non decreasing we have $\underline{P}(r) \geq P(r)$ for $0 \leq r \leq \underline{r}$, and $\underline{P}(r) \leq P(r)$ for $r \geq \underline{r}$. In particular, $\underline{r}$ is the smallest root of $\underline{P}$, i.e.

$$\underline{r} = \frac{1 - Z_1 - \sqrt{(1 - Z_1)^2 - 2Y Z_2(\underline{r})}}{Z_2(\underline{r})}.$$

12. In other words, in a situation where the bounds are slightly too tight and the a priori validation fails with the *usual* radii polynomial defined as $\tilde{P}(r) = Z_2(r) r^2 - (1 - Z_1) r + Y$, the method may still be succesful with this *new* radii polynomial P defined as in (2.22).

This shows that $Z_2(\underline{r})\underline{r} < 1 - Z_1$ and thus, since $r \mapsto Z_2(r)r$ in non decreasing, we have $\underline{r} < r^*$. □

In Chapters 6 to 10, we successfully use Proposition 2.2.1 [13] to prove the existence of solutions in various contexts. Before presenting those results in more details, we give a brief review of several analogous a posteriori validation methods.

We start with the technique of Yamamoto [214], which is also based on Banach's fixed point theorem. In Yamamoto's method, one considers $\tilde{F} = L^{-1}F = I + L^{-1}N$ together with a decomposition $\mathcal{X} = \mathcal{X}_h \oplus \mathcal{X}_\infty$, where $\mathcal{X}_h$ is a finite dimensional subspace of $\mathcal{X}$. Introducing Π_h the projection onto $\mathcal{X}_h$, and $A_h = \left(\Pi_h D\tilde{F}(\bar{x})|_{\mathcal{X}_h}\right)^{-1}$, the fixed point operator of interest is then given by

$$\tilde{T}(x) = \left(\Pi_h x - A_h \Pi_h \tilde{F}(x)\right) + (I - \Pi_h)\left(\tilde{F}(x) - x\right). \tag{2.23}$$

Notice that, as long as A_h is injective, $\tilde{T}(x) = x$ implies $\tilde{F}(x) = 0$ which in turn implies $F(x) = 0$. Here the finite dimensional part $\Pi_h\tilde{T}(x) = \left(\Pi_h x - A_h\Pi_h\tilde{F}(x)\right)$ is Newton-like, while the *reminder part* $(I - \Pi_h)\tilde{T}(x) = (I - \Pi_h)\left(\tilde{L}^{-1}N(x)\right)$ should be small if the dimension of $\mathcal{X}_h$ is large enough. Yamamoto then uses estimates similar [14] to the ones of Proposition 2.2.1 to apply Banach's fixed point theorem to $\tilde{T}$ and get the existence of a zero of F. We point out that the difference between *our* operator T defined in (2.18) and Yamamoto's operator $\tilde{T}$ defined in (2.23) is mainly one of presentation (the finite dimensional projection being kind of hidden in A in our case) and is linked to the fact that, in our case F is defined using spectral methods, whereas in Yamamoto's case F is based on a local method like finite elements. More details about the use of spectral methods in this context is given in Section 2.2.3 (they are also used in Sections 2.2.4, 2.2.6 and 2.2.7), whereas local methods for validated numerics are discussed more in Section 2.2.5.

Another technique that should be mentioned is the one of Nakao (see the survey [177] and the references therein), who was in fact the one to introduce the Newton-like operator (2.23) [15] in the context of validated numerics and influenced the work of Yamamoto. The main difference with Yamamoto's technique is that Nakao's approach is based on Schauder's fixed point theorem. The strategy is to construct a closed, bounded and convex subset $W = W_h \oplus W_\infty$, where $W_h \subset \mathcal{X}_h$ and $W_\infty \subset \mathcal{X}_\infty$, satisfying

$$\Pi_h\tilde{T}(W) \subset W_h \quad \text{and} \quad (I - \Pi_h)\tilde{T}(W) \subset W_\infty.$$

The finite dimensional part is treated with intensive use of interval enclosures of sets (see [177] for the details), and thus Nakao's method could be described as an *hybrid* approach between the geometric and the functional analysis techniques.

We finish by mentioning the method of Plum (see the survey [187] and the references therein), which at first glance is somewhat analogous to Nakao's method insofar as it is also based on Schauder's fixed point theorem, but actually has very close similarities with our approach. Plum's method is based on a fixed point reformulation of the form

$$\hat{T}(x) = \bar{x} - (DF(\bar{x}))^{-1}\left[F(\bar{x}) + N(x) - N(\bar{x}) - DN(\bar{x})(x - \bar{x})\right],$$

13. In practice, we used a slightly different version of Proposition 2.2.1, where an operator $A^\dagger$ approximating $DF(\bar{x})$ is introduced. The quantity $I - ADF(\bar{x})$ is then split as $\left(I - AA^\dagger\right) + A\left(DF(\bar{x}) - A^\dagger\right)$ and both terms are estimated separately. Therefore the bound Z_1 is replaced by two separate bounds, denoted Z_0 and Z_1.

14. To be precise, the presentation of Proposition 2.2.1 is being inspired by the paper [214], it is rather our estimates that are similar to the ones of Yamamoto.

15. In some earlier studies, such as the original work of Nakao [176], different fixed point reformulations were sometimes used, such as $T = (-L)^{-1}N$. However, the advantage of the Newton-like approach is that Schauder's or Banach's fixed point theorem should *in principle* always be applicable close to the numerical solution, which was not the case with those alternative formulations.

and aims at showing that, for some positive r, T maps $\mathcal{B}_{\mathcal{X}}(\bar{x}, r)$ into itself, to apply Schauder's fixed point theorem (a straightforward verification, remembering that $F = L + N$ with L linear, shows that $\hat{T}(x) = x$ indeed implies $F(x) = 0$). Plum's method provides sufficient conditions [16] for the enclosure $\hat{T}(\mathcal{B}_{\mathcal{X}}(\bar{x}, r)) \subset \mathcal{B}_{\mathcal{X}}(\bar{x}, r)$ to hold, which are very similar to the ones of Proposition 2.2.1. Indeed, given bounds $\hat{Y}$, K and a nondecreasing function $\hat{Z}_2$ such that

$$\|F(\bar{x})\|_{\mathcal{X}} \leq \hat{Y}, \quad \|DF(\bar{x})^{-1}\|_{\mathcal{X}} \leq K \quad \text{and} \quad \|DN(x) - DN(\bar{x})\|_{\mathcal{X}} \leq r\hat{Z}_2(r) \quad \forall\, x \in \mathcal{B}_{\mathcal{X}}(\bar{x}, r),$$

we have that $\hat{T}(\mathcal{B}_{\mathcal{X}}(\bar{x}, r)) \subset \mathcal{B}_{\mathcal{X}}(\bar{x}, r)$ as soon as

$$\hat{P}(r) := \frac{1}{2}K\hat{Z}_2(r)r^2 - r + K\hat{Y} < 0.$$

The main difference with our estimates is that, instead of using an approximate inverse A of $DF(\bar{x})$, Plum's method considers the exact inverse (or more precisely, a bound on the norm of the exact inverse). This explain the disappearance of the bound Z_1 (which can be taken equal to zero if $A = (DF(\bar{x}))^{-1}$). Aside from that, the estimates are really alike, which can be explained by noticing that

$$\begin{aligned}
\hat{T}(x) &= \bar{x} - (DF(\bar{x}))^{-1}\left[L\bar{x} + N(\bar{x}) + N(x) - N(\bar{x}) - DN(\bar{x})(x - \bar{x})\right] \\
&= \bar{x} - (DF(\bar{x}))^{-1}\left[Lx + N(x) - (L + DN(\bar{x}))(x - \bar{x})\right] \\
&= \bar{x} - (DF(\bar{x}))^{-1}\left[F(x) - DF(\bar{x})(x - \bar{x})\right] \\
&= x - (DF(\bar{x}))^{-1}F(x).
\end{aligned}$$

Thus, $\hat{T}$ is nothing but a somewhat hidden Newton-like operator of the form (2.18), with $A = (DF(\bar{x}))^{-1}$. Let us make some more comments about the differences between Plum's method and our method. Since we take for A a well chosen approximation of $DF(\bar{x})^{-1}$ which is in some sense *simpler* than $DF(\bar{x})^{-1}$ (this is detailed in Section 2.2.3), we can estimate directly the norm of products such as $\|AF(\bar{x})\|$ or $\|A(DF(x) - DF(\bar{x}))\|$, whereas Plum's technique uses the generically less sharp bounds $\|A\|\|F(\bar{x})\|$ and $\|A\|\|(DF(x) - DF(\bar{x}))\|$ (in his case with A exactly equal to $DF(\bar{x})^{-1}$). However this comes at a cost, which is the actual construction of our approximate inverse A, that requires the numerical computation of the inverse of a (possible large) matrix representing a finite dimensional projection of $DF(\bar{x})$ (this is in fact similar to the operator A_h of Nakao's and Yamamoto's methods, again see Section 2.2.3 for more details). On the other hand, Plum's method does not require to actually compute the inverse of $DF(\bar{x})$, but only get a bound on its norm. However, to get this bound Plum uses eigenvalues enclosing techniques (on $DF(\bar{x})$ or on $DF(\bar{x})^* DF(\bar{x})$), which can themselves be hard to apply, depending on the specific problem at hand, especially if the Banach space $\mathcal{X}$ does not have a Hilbert structure.

In the next five subsections, we describe the contributions of this thesis related to validated numerics techniques. The works presented in Sections 2.2.3, 2.2.4 and 2.2.5 are mainly focused on broadening the framework in which these techniques are applicable and on improving their efficiency (which also results in enlarging their applicability, at a practical level). The works presented in Sections 2.2.6 and 2.2.7 are more focused on applications and give examples where validated numerics techniques can be used to solve open problems that could not be handled by purely analytical techniques.

2.2.3 Exposition of the results of Chapter 6

In this section we present the main results obtained in Chapter 6 of this thesis. To put these results in perspective, we start by giving a simple example of an equation where our validated numerics technique can be applied, and describe the construction of the approximate inverse A.

16. We reformulate some of them slightly to make the comparison easier.

Consider the equation

$$\begin{cases} -u'' + u + u^2 = g, \\ u'(0) = 0 = u'(\pi), \end{cases} \tag{2.24}$$

where g is a smooth, even, 2π periodic function. Admittedly, the existence of solutions of (2.24) could be proved without using validated numerics, for instance using a variational approach. Our aim here is to present the main ideas without having to worry to much about technicalities.

Considering the Fourier expansions of *ufor the moment this expansion is formal, but we are going to obtain coefficients with fast enough decay to justify all the computations a posteriori* and g:

$$u(t) = x_0 + 2\sum_{k=1}^{\infty} x_k \cos(kx) \quad \text{and} \quad g(t) = g_0 + 2\sum_{k=1}^{\infty} g_k \cos(kx),$$

and plugging them in (2.24), we end up with the following (infinite) system of algebraic equations

$$k^2 x_k + x_k + (x * x)_k = g_k, \quad \forall\ k \in \mathbb{N},$$

where $(x * y)$ denotes the convolution product of the sequences x and y, defined as

$$(x * y)_k = \sum_{l \in \mathbb{Z}} x_{|k-l|} y_{|l|}, \quad \forall\ k \in \mathbb{N}.$$

If we consider $x = (x_k)_{k\in\mathbb{N}}$ the sequence of Fourier coefficients of u, this system can be rewritten under the form $F(x) = 0$, where $F = (F_k)_{k\in\mathbb{N}}$ is defined as

$$F_k(x) = k^2 x_k + x_k + (x * x)_k - g_k, \quad \forall\ k \in \mathbb{N}.$$

This is exactly the kind of function $F : \mathcal{X} \to \mathcal{Y}$ we had in mind in Section 2.2.2, with L given by[17] $L_k(x) = k^2 x_k$ for all $k \in \mathbb{N}$. Typical choices for the Banach spaces $\mathcal{X}$, in which we want to prove the existence of a zero of F, and $\mathcal{Y}$ are weighted ℓ^1 and ℓ^∞ spaces (weighted ℓ^∞ spaces are used in Chapter 6 and weighted ℓ^1 spaces are used in Chapters 7, 9 and 10).

Fix an integer $m > 0$ and assume that we have computed numerically an approximate zero of F, by considering the Galerkin projection of size m. More precisely, consider $\mathcal{X}^{[m]}$ the subspace of $\mathcal{X}$ of sequences x such that $x_k = 0$ for all $k \geq m$, together with the associated projection $\Pi^{[m]} : \mathcal{X} \to \mathcal{X}^{[m]}$, defined as

$$\Pi_k^{[m]}(x) = \begin{cases} x_k & \forall\ k < m, \\ 0 & \forall\ k \geq m. \end{cases}$$

Then $F^{[m]} := \Pi^{[m]} F_{|\mathcal{X}^{[m]}}$ can be seen as a function from $\mathbb{R}^m$ to itself, for which we can numerically find a zero by using (for instance) Newton's method. We now assume we have computed such a zero, and denote it $\bar{x}$.

The next step is to define an approximate inverse of $DF(\bar{x})$. To do so, we start by approximating $DF(\bar{x})$ itself. We have that, for all $x \in \mathcal{X}$,

$$DF_k(\bar{x})x = k^2 x_k + x_k + 2(\bar{x} * x)_k, \quad \forall\ k \in \mathbb{N}.$$

Notice that the asymptotically dominant term is given by $L_k x = k^2 x_k$. Therefore, seeing linear operators between $\mathcal{X}$ and $\mathcal{Y}$ as infinite dimensional matrices, we define the following approximation for $DF(\bar{x})$

$$A^\dagger = \left(\begin{array}{c|ccc} DF^{[m]}(\bar{x}) & & (0) & \\ \hline & m^2 & & (0) \\ (0) & & (m+1)^2 & \\ & (0) & & \ddots \end{array}\right),$$

17. The *order zero* term x_k could also be included in L, and we would then consider $L_k x = (k^2 + 1)x_k$, but this does not change much here.

which is made of a finite dimensional bloc $DF^{[m]}(\bar{x})$ of size $m \times m$, and of a diagonal tail where we only kept the term coming from L. Considering $A^{[m]}$ a numerically computed inverse of $DF^{[m]}(\bar{x})$, we then define our approximate inverse A as

$$A = \left(\begin{array}{c|ccc} A^{[m]} & (0) & & \\ \hline & \frac{1}{m^2} & & (0) \\ (0) & & \frac{1}{(m+1)^2} & \\ & (0) & & \ddots \end{array}\right).$$

This is the kind of approximate inverse A of $DF(\bar{x})$ that we alluded to in Section 2.2.2, and that we use for Proposition 2.2.1. The *quality* of this approximate inverse is ultimately checked by the bound Z_1 of Proposition 2.2.1. Indeed, for our method to succeed, we need to be able to get a bound Z_1 strictly less than 1, which is the case here with such an A if m is large enough [18]. The fact that A has a simple structure (i.e. a finite block and a diagonal tail) is crucial in our approach to actually derive the bounds needed in Proposition 2.2.1.

We point out that we can obtain such a simple structure for $A^\dagger$ and then for A only because the asymptotically dominant term, given by L, is diagonal. Indeed, we can perform a numerical inversion for a finite bloc only, but the fact that the (infinite) tail of $A^\dagger$ is diagonal allows to invert it analytically to get A.

This example gives a typical situation in which our method is applicable. To present things in a more general setting, we used that $\mathcal{X}$ has a Schauder basis on which L is diagonal, and that L is asymptotically dominant in $DF(\bar{x}) = L + DN(\bar{x})$, in the sense that $L^{-1}DN(\bar{x})$ is compact.

The aim of the work of Chapter 6 is to be able to handle a broader range of situations, including cases where L is no longer diagonal. A typical situation where this naturally occurs, comes from the following generalization of (2.24)

$$\begin{cases} -(2+\cos(t))u''(t) + u(t) + u^2(t) = g(t), \\ u'(0) = 0 = u'(\pi). \end{cases} \tag{2.25}$$

After considering again a Fourier expansion, the function F we end up with is this time given by

$$F_k(x) = \frac{1}{2}(k-1)^2 x_{k-1} + 2k^2 x_k + \frac{1}{2}(k+1)^2 x_{k+1} + x_k + (x*x)_k - g_k, \quad \forall\ k \geq 1,$$

and

$$F_0(x) = x_1 + x_0 + (x*x)_0 - g_0.$$

The dominant linear term is then tridiagonal and given by

$$L_k x = \frac{1}{2}(k-1)^2 x_{k-1} + 2k^2 x_k + \frac{1}{2}(k+1)^2 x_{k+1},$$

therefore a natural *simple* approximation of $DF(\bar{x})$ is now of the form

$$A^\dagger = \left(\begin{array}{cc|ccc} & & & (0) & \\ DF^{[m]}(\bar{x}) & & & \beta_{m-1} & \\ \hline & \lambda_m & \nu_m & \beta_m & (0) \\ (0) & & \lambda_{m+1} & \nu_{m+1} & \ddots \\ & & (0) & \ddots & \ddots \end{array}\right), \tag{2.26}$$

18. assuming that $A^{[m]}$ is a good enough numerical inverse of $DF^{[m]}(\bar{x})$.

where, for this particular example $\lambda_k = \frac{1}{2}(k-1)^2$, $\nu_k = 2k^2$ and $\beta_k = \frac{1}{2}(k+1)^2$. To apply our validated numerics technique in this case, we need to:

- Find out how to approximately inverse $A^\dagger$ defined as in (2.26), to define A.
- Adapt the usual estimates to this new A which no longer has a simple diagonal tail.

Chapter 6 answers these two points, under the assumption that there exist $k_0 \in \mathbb{N}$ and $\delta < \frac{1}{2}$ such that

$$\left|\frac{\lambda_k}{\nu_k}\right|, \left|\frac{\beta_k}{\nu_k}\right| \leq \delta, \quad \forall\ k \geq k_0.$$

To define A, we start by inverting (analytically) the tridiagonal tail (given by the bottom right bloc in (2.26)), using some generalization of the LU decomposition to infinite matrices. After also inverting the finite bloc $DF^{[m]}(\bar{x})$ (this time numerically), we determine how these two inverses must be coupled to take into account the two additional terms β_{m-1} and λ_m, to finally get a good approximate inverse A of $A^\dagger$. Because of the more complicated structure of the A we obtain (it is now basically a full (infinite dimensional) matrix), the derivation of the bounds for Proposition 2.2.1 becomes much more technical, and we do not give more details here.

As an example, several solutions of a parameter dependent version of (2.25) are computed and validated using our technique (see Figure 2.2).

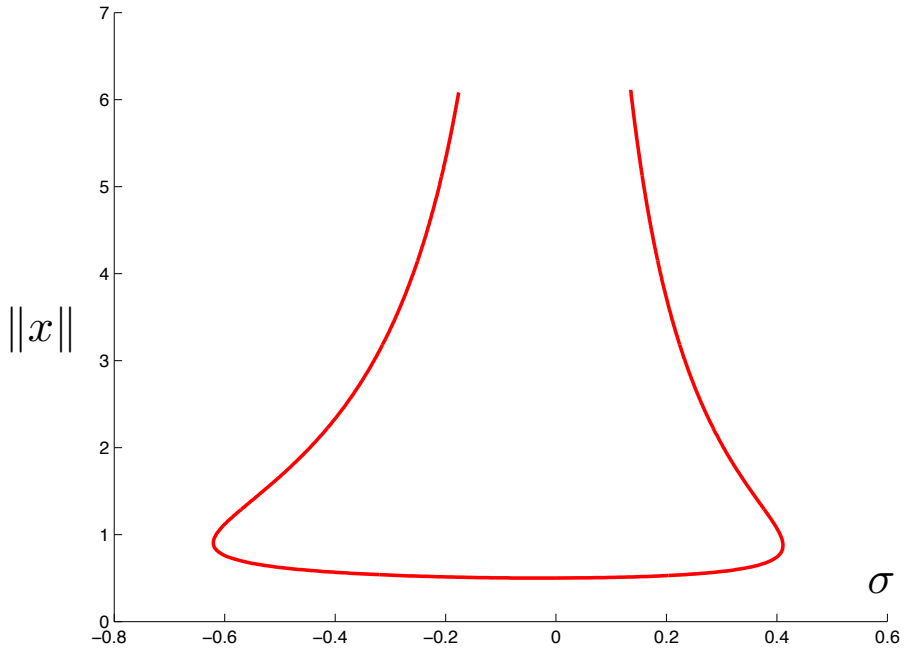


Figure 2.2 – Validated solutions of $-(2+\cos(t))u''(t) + u(t) + \sigma u^2(t) = g(t)$, with homogeneous Neumann boundary conditions, $g(t) = \frac{1}{2} + 3\cos(t) + \frac{1}{2}\cos(2t)$, for different values of σ.

2.2.4 Exposition of the results of Chapter 7

In this section we present the main results obtained in Chapter 7 of this thesis, which is about stable and unstable manifolds.

Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be $\mathcal{C}^1$ vector field, and consider the ODE

$$y' = g(y). \tag{2.27}$$

Let $p \in \mathbb{R}^n$ be an equilibrium point for g, i.e. $g(p) = 0$. Assume that $Dg(p)$ has k eigenvalues $\lambda_1 \leq \ldots \leq \lambda_k$ with negative real part. To make the presentation easier, we assume here that the

eigenvalues are simple and real. The case of complexe eigenvalues is treated in Chapter 7, for a generalization of some of the techniques presented here to the case including multiple eigenvalues we refer to [66]. Denote $V_1, \dots, V_k$ the associated eigenvectors and E^s the k dimensional subspace of $\mathbb{R}^n$ spanned by them. We recall (see for instance [186]) that their exists a k dimensional manifold $W^s_{loc}(p)$, which is tangent to the stable subspace E^s at p, is stable by the flow ϕ generated by g, and such that, for all $x \in W^s_{loc}(p)$, $\phi_t(x) \underset{t\to\infty}{\longrightarrow} p$. $W^s_{loc}(p)$ is called the *local stable manifold* of p. Similarly, there is an *local unstable manifold* associated with the eigenvalues with positive real part.

The local stable and unstable manifold play a crucial role in understanding the dynamics of the solutions of (2.27) near the equilibrium p. They can also be very useful to study connecting orbits (i.e. solutions y such that y converges to equilibrium points in $\pm\infty$). Indeed, if one can get a solution that lies in an unstable manifold (say of an equilibrium p) at a time t_0 and lies in a stable manifold (say of an equilibrium q) at a time t_1, then we are done, since by definition of the stable and unstable manifolds the solution must go to p as time goes to $-\infty$ and to q as time goes to $+\infty$. Therefore, using the stable and unstable manifolds allows to reduce the problem of finding a connecting orbit (defined on $\mathbb{R}$) between two equilibria, to the problem of finding an orbit (defined on a *finite* time interval) between two manifolds. This reduction from an infinite to a finite time interval can be fairly helpful, especially when dealing with numerical computations. However, for this reduction to be effective, one must first compute the manifolds in question.

There are several techniques to compute such manifolds, the one we use in Chapter 7 was introduced by Cabré, Fontich and de la Llave [82, 83, 84] and is called the *parameterization method*. Their method consists in finding a specific parameterization of the manifold, that semi-conjugates the dynamics of the system to the dynamics of its linearization. More precisely, denoting by Λ the diagonal matrix containing the stable eigenvalues $\lambda_1, \dots, \lambda_k$, and $\mathcal{B}_\nu$ the closed ball of radius ν in $\mathbb{R}^k$, we look for a parameterization $f : \mathcal{B}_\nu \to \mathbb{R}^n$, such that $f(0) = p$ and

$$f(e^{t\Lambda}\theta) = \phi_t(f(\theta)), \quad \forall\ \theta \in \mathcal{B}_\nu,\ \forall\ t \geq 0. \tag{2.28}$$

Notice that, if such an f exists, then $f(\mathcal{B}_\nu)$ is indeed a local stable manifold of p [19]. However, this formulation is not really convenient, as in involves the flow ϕ. Therefore, we differentiate (2.28) with respect to t, and then evaluate at $t = 0$, to obtain the so-called *invariance equation*

$$Df(\theta)\Lambda\theta = g(f(\theta)), \tag{2.29}$$

which is in fact equivalent to (2.28). Indeed, assume that f solves the invariance equation (2.29) and denote $h(t) = f(e^{t\Lambda}\theta)$. Then $h(0) = f(\theta)$ and $h'(t) = g(f(e^{t\Lambda}\theta)) = g(h(t))$, therefore $h(t) = \phi_t(f(\theta))$ and (2.28) holds.

Assume now that g is analytic. It is shown in [82] that the invariance equation (2.29) then has an analytic solution f, as soon as *non resonance condition* on the eigenvalues (detailed in Chapter 7) is satisfied. Assuming this condition holds, we can look for f as a power series

$$f(\theta) = \sum_{|\alpha|\geq 0} a_\alpha \theta^\alpha, \quad \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix} \in \mathbb{R}^k,\ a_\alpha = \begin{pmatrix} a_\alpha^{(1)} \\ \vdots \\ a_\alpha^{(n)} \end{pmatrix} \in \mathbb{R}^n, \tag{2.30}$$

with the classical multi-indices notations $|\alpha| = \alpha_1 + \ldots + \alpha_k$ and $\theta^\alpha = \theta_1^{\alpha_1} \dots \theta_{n_s}^{\alpha_k}$. Plugging this series expansion in (2.29), we get an (infinite) system of algebraic equations for the unknown coefficients $(a_\alpha)_{|\alpha|\geq 0}$, on which we can apply our validated numerics technique. This kind of combination of the parameterization method with computer-assisted proofs originates from the

19. And is of dimension k if $Df(0)$ is of rank k.

work of van den Berg, Lessard, Mireles James and Mischaikow in [65], where parameterizations of local stable manifolds are computed and validated, and then used to prove the existence of homoclinic orbits for the Gray-Scott equation.

To apply our validated numerics technique, based on Banach's fixed point theorem, to the invariance equation (2.29), we need this equation to have isolated solutions. Therefore, in addition to the condition $f(0) = p$, we have to append a condition for $Df(0)$, the natural choice being

$$Df(0) = \begin{pmatrix} V_1 & \dots & V_k \end{pmatrix}, \tag{2.31}$$

or a similar condition with rescaled eigenvectors. Therefore, asides from the parameters inherent to our validation method such as the dimension of the truncation used to define A, when trying to validate manifolds we have an additional set of free parameters given by ν (which prescribes the size of the domain of definition $\mathcal{B}_\nu$ of the parameterization f, in the parameter space $\mathbb{R}^k$), and the scaling of the eigenvectors. It is important to tune these free parameters appropriately, as they can influence the success or failure of the validation procedure, but also because they have an impact on the manifold (i.e. the geometrical object $f(\mathcal{B}_\nu)$) that is obtained. Depending on the situation, one may want to get a manifold that is as large as possible, or to get as far as possible in a given direction (this can be especially relevant when dealing with fast-slow systems).

Prior to the work of Chapter 7, these free parameters where selected by a time consuming trial and error process based on *numerical experimentations*. Chapter 7 provides a kind of automatic procedure to select those parameters, that can be easily adapted to the situation at hand. It is based on the simple but crucial observation that, at least at the theoretical level, for a given $\gamma > 0$ it is equivalent to change the radius of the parameterization domain from ν to $\gamma\nu$ or to rescale the eigenvectors by γ, that is to replace the condition (2.31) by

$$Df(0) = \begin{pmatrix} \gamma V_1 & \dots & \gamma V_k \end{pmatrix}.$$

Of course, if you want to get the largest manifold possible, the natural thing to do is to try and get the largest ν possible, to increase the domain of definition of the parameterization. In practice however, for numerical stability reasons it is better to always keep $\nu = 1$. Therefore we fix $\nu = 1$ and only change the scaling of the eigenvectors. To promote a specific direction in the manifold, we also consider *anisotropic* scalings $\gamma = (\gamma_1, \dots \gamma_k)$ and the associated condition

$$Df(0) = \begin{pmatrix} \gamma_1 V_1 & \dots & \gamma_k V_k \end{pmatrix}. \tag{2.32}$$

In Chapter 7 we first develop a method for choosing these scalings that is only based on *numerical defect*, and then another (slightly more costly) that can be used for rigorous validation. More precisely, we derive the bounds Y, Z_1 and Z_2 defined in Proposition 2.2.1 (for the function F obtained when plugging the series expansion (2.30) into the invariance equation (2.29), and complemented with the condition (2.32)), which depend explicitly on $\gamma = (\gamma_1, \dots \gamma_k)$. Therefore, once the bounds have been computed, they can be recomputed cheaply for any rescaling $\gamma = (\gamma_1, \dots \gamma_k)$.

As an example of application, we compute local stable and unstable manifolds of equilibria for the Lorenz and FitzHugh-Nagumo systems, as well as for the suspended bridge equation. This last example is studied more extensively in Chapter 9, where we use our method to compute and validate a continuation of manifolds, which are then used to prove the continuation of homoclinic orbits (more details in Section 2.2.6).

2.2.5 Exposition of the results of Chapter 8

In this section we present the main results obtained in Chapter 8 of this thesis, which is about the use of polynomial interpolation in the context of validated numerics.

The original work of Yamamoto [214], as well as more recent works (see for instance [65, 61]), have been using piecewise linear approximation to solve boundary values problems or prove the existence of connecting orbits using validated numerics. Piecewise linear approximation, and more generally finite element methods are more flexible tools than spectral methods, in the sense that they can be used to treat problems with complex geometry, whereas spectral methods are usually restricted to rectangular domains. Besides, polynomial interpolation is more fit to handle complicated (non polynomial) nonlinearities [20]. However, in situations where both can be used, spectral methods are substantially more efficient when used in the context of validated numerics.

To be more precise, assume for instance that you want to use validated numerics to compute the solution of an initial value problem [21] for an ODE

$$\begin{cases} u' = \phi(u) \quad \text{on } [0,\tau] \\ u(0) = u_0, \end{cases} \tag{2.33}$$

with a polynomial vector field ϕ. Then, with the methods we know of, the longest orbit that can be validated using a spectral method like Chebyshev series, is much longer than the longest orbit that could be validated using piecewise linear interpolation. One might immediately object here that the comparison is not fair, because piecewise linear interpolation is much rougher than approximations using Chebyshev series. However, we do not discuss directly the quality of the approximation but only the possibility to rigorously validate the obtained approximate solution. And while the two are obviously linked, we highlight through some examples in Chapter 8 that the quality of the approximation is not the main issue here, and that simply using a higher order approximation (like piecewise quadratic or piecewise cubic), does not allow to validate longer orbits.

In Chapter 8, we present a new technique, that we call *a priori bootstrap*, which allows us to take advantage of higher order polynomial approximation to validate much longer orbit. Before describing this technique, we briefly recall how piecewise linear approximation was used in the context of validated numerics. For the sake of simplicity we restrict our attention to the initial value problem (2.33). In the framework presented in Section 2.2.2, this would be a kind of $F = L+N = 0$ problem with $Lu = u'$ and $N(u) = -\phi(u)$. But as we mentioned when presenting Yamamoto's method, it is more convenient to consider the $I + L^{-1}N = 0$ formulation, which corresponds to the integral version of the ODE

$$u(t) = u_0 + \int_0^t \phi(u(s))ds, \quad \forall\ t \in [0,\tau]. \tag{2.34}$$

To fit with the notations of Section 2.2.2 we define

$$\tilde{F}(u)(t) = u(t) - u_0 - \int_0^t \phi(u(s))ds, \quad \forall\ t \in [0,\tau],$$

and aim at proving the existence of a zero of $\tilde{F}$. The validation procedure uses the space $\mathcal{X} = \mathcal{C}^0([0,\tau])$ of continuous functions on the interval $[0,\tau]$, and its decomposition $\mathcal{X} = \mathcal{X}_h \oplus \mathcal{X}_\infty$, where $\mathcal{X}_h$ is the subspace of piecewise linear functions associated with a grid of step-size h on $[0,\tau]$. That is, $\mathcal{X}_h$ is the subspace of all functions u such that $u_{|[jh,(j+1)h]}$ is linear for all $0 \leq j \leq \frac{\tau}{h} - 1$ [22]. The fixed point operator on which we then want to apply Banach's fixed point theorem is exactly the operator $\tilde{T}$ defined in (2.23). In fact, we use a kind of Newton

20. Nonetheless, we mention that techniques inspired from automatic differentiation can sometimes be used to handle non polynomial nonlinearities, in the context of validated numerics with spectral methods (see [154] and also Chapter 10).

21. The same comparison holds for solutions of boundary value problems, or periodic orbits.

22. where h is assumed to be chosen such that $\frac{\tau}{h}$ is an integer.

Kantorovich theorem very similar to Proposition 2.2.1, and therefore we need to control (among others) interpolation errors like

$$(I - \Pi_h)(F(u) - u) = (I - \Pi_h)\left(t \mapsto \int_0^t \phi(u(s))ds\right).$$

This is the limiting term in the estimates and is of order h, since for a $\mathcal{C}^0$ function u the function $t \mapsto \int_0^t \phi(u(s))ds$ is only $\mathcal{C}^1$ (precise estimates and more detailed explanations are given in Chapter 8). We already see here that simply increasing the order of the approximation (for instance considering for $\mathcal{X}_h$ the subspace of piecewise quadratic functions and Π_h the associated projection of $\mathcal{X}$ onto $\mathcal{X}_h$) does not really help. Indeed for $\mathcal{C}^1$ functions and a fixed h, doubling the degree of the polynomial interpolation roughly divides by two the interpolation error (again see Chapter 8 for detailed estimates), which is not better than simply taking h twice as small without changing the degree of the polynomial interpolation.

The goal of our a priori bootstrap procedure is to circumvent this limitation. To do so, starting from (2.33), we first take a derivative to get the equivalent second order problem (in which we replaced u' by $\phi(u)$)

$$\begin{cases} u'' = D\phi(u)\phi(u) \quad \text{on } [0,\tau] \\ u(0) = u_0, \\ u'(0) = \phi(u_0). \end{cases} \tag{2.35}$$

We then integrate it back twice to get the following equivalent integral formulation [23]

$$u(t) = u_0 + t\phi(u_0) + \int_0^t (t-s)D\phi(u(s))\phi(u(s))ds.$$

This formulation is arguably more complicated, but the crucial point is that, for a $\mathcal{C}^0$ function u, the function $t \mapsto \int_0^t (t-s)D\phi(u(s))\phi(u(s))ds$ is now $\mathcal{C}^2$, allowing for better estimates for the interpolation errors. We could reiterate this procedure, to get a $\mathcal{C}^3$ function, and so on. In some sense the a priori bootstrap technique allows us to incorporate some (a priori known) regularity of the solution into the integral formulation. It is only at this point that it becomes valuable to use higher order interpolation, since the previously limiting factor (which is related to the Z_1 bound) has now been taken care of, and the quality of the approximation (i.e. the bound Y) then starts becoming relevant.

In Chapter 8 we derive general estimates, to use validated numerics with an arbitrary number p of a priori bootstraps, and any order k of polynomial interpolation (at the Chebyshev points of the second kind). While this a priori bootstrap reformulation might seem rather naive, it results in very significant improvements for the validated numerics techniques. A detailed comparison study is made in Chapter 8, we only mention here an example that we obtained on the Lorenz system. For a given initial data u_0 and a fixed dimension of the finite dimensional subspace X_h, the maximal length τ of the orbit that we could validate went up from 0.7 without a priori bootstrap, to 5.6 when using a priori bootstrap once, and then to 8.1 when using it twice. In particular, combining this method with the technique developed in Chapter 7 for maximizing local manifolds allowed us to validate an heteroclinic orbit for the Lorenz system with standard parameter values, that could not be validated without our a priori bootstrap technique in a previous study [155].

We were also able to validate specific solutions (called ballistic spiral orbits) for ABC flows [55, 112, 126], where the vector field is given by

$$\phi_{A,B,C}(x,y,z) = \begin{pmatrix} A\sin(z) + C\cos(y) \\ B\sin(x) + A\cos(z) \\ C\sin(y) + B\cos(x) \end{pmatrix}, \quad A, B, C \in \mathbb{R}.$$

23. which could also be obtained directly from (2.34) by doing an appropriate integration by part.

This is a typical example where spectral methods struggle in the context of validated numerics, because of the non polynomial nonlinearities, but that we can now handle with polynomial interpolation combined with a priori bootstrap.

2.2.6 Exposition of the results of Chapter 9

In this section we present the main results obtained in Chapter 9 of this thesis, which deal with the existence of traveling waves in the suspended bridge equation.

Motivated by an old report of observations on the Golden Gate Bridge during a storm in 1938, Chen and McKenna started a mathematical study of traveling waves in a suspended bridge [97]. They considered the equation

$$\partial_{tt}U + \partial_{xxxx}U + (1+U)^+ - 1 = 0, \tag{2.36}$$

where $U^+ := \max(U, 0)$. Here $U(t, x)$ describes the height of the bridge deck at time t and position x (x is an unidimensional variable describing the direction of traffic), where $U = 0$ describes the bridge at rest. Chen and McKenna proved the existence of traveling waves solutions that can be computed explicitly. More precisely, they looked for solutions of the form

$$U(t, x) = u(x - ct), \tag{2.37}$$

such that the profile u goes to 0 exponentially fast when the independent variable $\xi = x - ct$ goes to $\pm\infty$. Therefore, plugging the ansatz (2.37) into the bridge equation (2.36), they ended up with having to find homoclinic orbits for the ODE

$$u'''' + \beta u'' + (1+u)^+ - 1 = 0, \tag{2.38}$$

where $\beta = c^2$. Using that the equation is piecewise linear, they found explicit homoclinic solutions for all $\beta \in [\beta_1, \beta_2]$, where $0 < \beta_1 < \beta_2 < 2$. In a further study [98], Chen and McKenna developed a method based on the mountain pass theorem that could be used to prove the existence of homoclinic orbits of (2.38) for all β in $(0, 2)$. However, some numerical observations suggested that it would be better to consider a slightly more refined model (with a smoother nonlinearity) given by

$$\partial_{tt}U + \partial_{xxxx}U + e^U - 1 = 0, \tag{2.39}$$

which, still using the ansatz (2.37) gives the following ODE

$$u'''' + \beta u'' + e^u - 1 = 0. \tag{2.40}$$

The method developed by Chen and McKenna could be applied to equations of the form

$$u'''' + \beta u'' + g(u) - 1 = 0,$$

for nonlinearity g more general than just $g(u) = (1+u)^+$, but their technique could no handle (2.40). Still, Chen and McKenna studied numerically the existence of homoclinic orbits for (2.40) and conjectured that there also existed homoclinic solutions of (2.40) for all $\beta \in (0, 2)$.

This conjecture was partially solved by van den Berg and Smets in [68], where the existence of homoclinic orbits for (2.40) is proved for *almost all* β in $(0, 2)$ using variational methods. In [193], it was then proven that homoclinic orbits exists for all β in $(0, \beta^*)$, where $\beta^* \simeq 0.5516$.

In Chapter 9, we use validated numerics to complement theses results and prove the following theorem.

Theorem 2.2.4. *For all β in $(0.5, 1.9)$, there exists a homoclinic orbit of (2.40).*

Before giving some details on this result, we mention that computer-assisted proofs were already used to study the suspended bridge equation: Plum's method was applied in [80] to prove the existence of 36 different homoclinic orbits of (2.40), for a given parameter value (in that case $\beta = 1.69$), thus answering another question that was raised by Chen and McKenna in [98].

Our proof is based on the work of Chapter 7, as we use the method developed there to efficiently compute and validate the stable manifold of the origin [24]. The finite part of the orbit landing onto the manifold is then computed and validated using Chebyshev series. We also mention that we need a generalization of Proposition 2.2.1. Indeed, as it is stated in Section 2.2.2, Proposition 2.2.1 could be applied (to validate the local stable manifold and the finite part of the orbit) to prove the existence of a homoclinic orbit for a fixed value of β. Therefore, we use in Chapter 9 a parameter dependent version of Proposition 2.2.1, which is basically obtained by using a uniform contraction theorem. This approach was already used previously in a validated numerics setting, see [64, 79, 53].

2.2.7 Exposition of the results of Chapter 10

In this section we present the main results obtained in Chapter 10 of this thesis. In this last chapter, we study the steady states of a cross-diffusion system used in population dynamics, that was introduced by Shigesada, Kawasaki and Teramoto [194]

$$\begin{cases} \partial_t u = \partial_{xx}\left((d_1 + d_{12}v)u\right) + (r_1 - a_1 u - b_1 v)u & \text{on } \mathbb{R}_+ \times (0,1), \\ \partial_t v = d_2 \partial_{xx} v + (r_2 - b_2 u - a_2 v)v & \text{on } \mathbb{R}_+ \times (0,1), \\ \partial_x u(t,0) = 0 = \partial_x u(t,1) & \text{on } \mathbb{R}_+, \\ \partial_x v(t,0) = 0 = \partial_x v(t,1) & \text{on } \mathbb{R}_+. \end{cases} \tag{2.41}$$

Here $u = u(t,x)$ and $v = v(t,x)$ represent the population densities at time t and position x of two competing species. The reaction terms are those of the standard Lotka-Volterra model, the non negative coeffcients r_i, a_i and b_i $(i = 1, 2)$ describing the unhindered growth of the species, the intra-specific competition and the inter-specific competition respectively. The term that makes this system different from a standard reaction diffusion system, is $d_{12}\partial_{xx}(vu)$ which models that the presence of the species v influences the diffusion rate of the species u.

This term was added to the standard reaction diffusion Lotka-Volterra model by Shigesada, Kawasaki and Teramoto to account for *spatial segregation*. Indeed, biological observations suggest that competing species can coexist by forming pattern to avoid each other, therefore we expect our model to exhibit stable non homogeneous steady states, which is not the case for the standard reaction diffusion Lotka-Volterra (i.e. system (2.41) with $d_{12} = 0$, this is detailed in the introduction of Chapter 10).

Since the original paper of Shigesada, Kawasaki and Teramoto in 1979, the steady states of system (2.41) were studied quite extensively, both numerically and theoretically (see for instance [137] and the references therein, or the introduction of Chapter 10). The numerical studies suggest that, depending on the biological parameters, system (2.41) can exhibit a very wide variety of nonhomogeneous steady states (see for instance Figure 2.3). However, the existing theoretical results only cover a fraction of those numerical solutions (mainly those that are close to a bifurcation from a trivial homogeneous solution, or in some asymptotic parameter regime). The same observation holds for the study of the stability of those steady states.

24. The unstable manifold is in fact not needed, because we only compute half the orbit, i.e. for $t \geq 0$, thanks to a symmetry argument.

26. Of course, for the straight line of solutions at $v(0) = 0.125$, which corresponds to trivial homogeneous solutions, the linear stability or instability could be proved by hand.

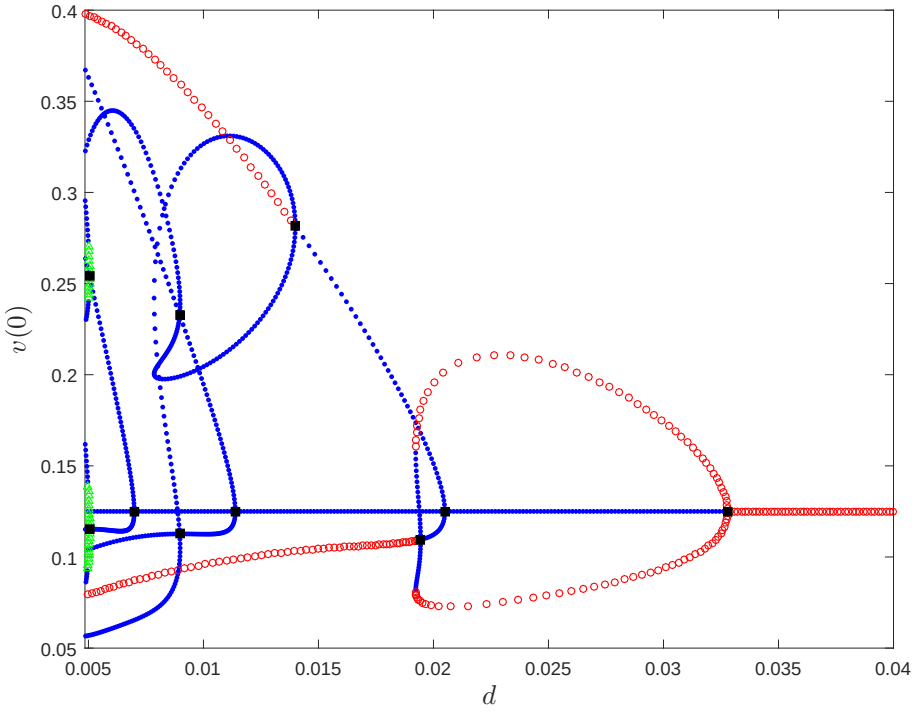


Figure 2.3 – A validated bifurcation diagram of steady states of (2.41), with $r_1 = 5$, $r_2 = 2$, $a_1 = 3$, $a_2 = 3$, $b_1 = 1$, $b_2 = 1$, $d_{12} = 3$, $d_{21} = 0$ and $d_1 = d_2 = d$ left as the bifurcation parameter. Each blue dot corresponds to a validated steady state that has been proven to be unstable. Each green triangle corresponds to a validated steady state, that is suspected to be unstable but for which the validation of the instability did not succeed. Each red circle corresponds to a validated steady state that is suspected to be stable [26].

In Chapter 10, we use validated numerics to study the steady states of (2.41). Thanks to the homogeneous Neumann boundary conditions, we are able to use a Fourier series expansion. However, because of the cross-diffusion term, we do not directly get a simple diagonal (or even tridiagonal) linear dominant structure as described in Section 2.2.3. Therefore, we first perform a change of variable $w = (d_1 + d_{12}v)u$ before applying our validated numerics technique. Once the steady states have been validated, we focus on their stability. We are able to prove that many of the validated steady states are unstable, by computing and then validating an eigenvalue with positive real part for the linearized equation around the (non homogeneous) steady state. Unfortunately we were not able to extend this technique to prove the stability of steady states for this problem (even though numerical simulations suggest that some of them are indeed stable). Our results are summarized in the following theorem.

Theorem 2.2.5. *Each blue dot, green triangle and red circle in Figure 2.3, corresponds to a validated steady state of the cross-diffusion system* (2.41)*, with* $r_1 = 5$, $r_2 = 2$, $a_1 = 3$, $a_2 = 3$, $b_1 = 1$, $b_2 = 1$, $d_{12} = 3$, $d_{21} = 0$ *and* $d_1 = d_2 = d$ *left as the bifurcation parameter. Besides, each steady states represented by a blue dot is unstable.*

In particular, for $d = 0.005$ *there exists* 13 *different steady states, among which* 11 *are unstable.*

2.2.8 Conclusion and perspectives

In Chapter 6, we extend the range of application of spectral methods for validated numerics, from the standard situation with a *dominant diagonal linear term* to include some situations with a *dominant tridiagonal linear term*. In Chapter 7, we improve the efficiency of the para-

meterization method to compute invariant manifolds in the context of validated numerics, by developing estimates allowing to easily adapt the scaling of the eigenvectors to optimize the obtained manifold. In Chapter 8 we significantly improve the efficiency of polynomial interpolation in the context of validated numerics, by introducing an *a priori bootstrap* technique. In Chapter 9 we combine the work of Chapter 7 with Chebyshev series in the context of validated numerics, to partially solve a conjecture from Chen and McKenna about the existence of traveling waves for the suspended bridge equation. Finally, in Chapter 10 we apply our validated numerics technique to the SKT cross-diffusion system, to prove the existence of many non homogeneous steady states as well as the instability of some of them.

We finish this introductory section by presenting some natural extensions suggested by the results of this thesis, and remaining open questions.

- One of the possible extension of Chapter 8, which was in fact the main motivation for this work, would be to try and combine polynomial interpolation (in time) with a spectral method like Fourier series (in space, with appropriate boundary conditions) to develop validated numerics techniques for parabolic PDEs. The only existing results (in our validated numerics framework) for parabolic PDEs concerns the existence of time periodic solutions, where (for problems with appropriate boundary conditions in space) Fourier expansions can be used in every dimension (see for instance [122]). However for some boundary value problem in time, for example to get connection orbits, Fourier expansion in time is no longer appropriate. From a theoretical perspective, it seems that polynomial interpolation in time could be used, but the previously rather poor efficiency of this technique for validated numerics has thwarted studies in that direction. With the improvement provided by the a priori bootstrap method, we believe that this is now a promising direction of research.
- To completely solve the conjecture from Chen and McKenna about the existence of traveling waves for the bridge equation, one would need to prove the existence of homoclinic orbits for (2.40) for all $\beta \in [1.9, 2)$. While our technique could certainly be pushed past $\beta = 1.9$, there is no hope of actually reaching 2 with the exact same method. Indeed, when β goes to 2 the real part of the eigenvalues goes to 0 (the system undergoes a Hamiltonian-Hopf bifurcation), and our $F = 0$ formulation to find a parameterization of the local stable manifold becomes singular. An interesting project would be to try and unfold this bifurcation, to then get the existence of homoclinic orbits in an (explicit) left-neighborhood of $\beta = 2$. One could then try to push the method of Chapter 9 to reach this neighborhood and completely solve the conjecture.
- In Chapter 10, while we were able to validate many different steady states, among which some that are suspected to be linearly stable, we were not able to rigorously prove this stability. However, stability is a very relevant property from the biological point of view, since stable solutions are the ones that we expect to observe in practice. Therefore this seems a very interesting direction to explore. We mention that validated numerics techniques were already successfully used to prove stability results in some cases (see for instance [202, 96]), but the cross-diffusion term generates additional difficulties. And while we could remove some of those difficulties by using a change of variable $w = (d_1 + d_{12}v)u$ when studying the steady states (i.e. when $\partial_t u = 0 = \partial_t v$), the same change of variable is not so helpful when considering the time dependent equation.

Chapter 3

List of publications contained in this thesis

The results presented in Chapters 4 to 10 consist of original research works that have been published or submitted for publication.

- **Chapter 4** is the result of a collaboration with L. Desvillettes and K. Fellner, and was published in *Monatshefte für Mathematik* under the title *Smoothness of moments of the solutions of discrete coagulation equations with diffusion* [7].
- **Chapter 5** is the result of an autonomous work, and has been accepted for publication in *Kinetic and Related Models* under the title *Applications of improved duality lemmas to the coagulation-fragmentation equations with diffusion* [6].
- **Chapter 6** is the result of a collaboration with L. Desvillettes and J.-P. Lessard, and was published in *Discrete and Continuous Dynamical Systems - Series A* under the title *Rigourous numerics for nonlinear operators with tridiagonal dominant linear part* [73].
- **Chapter 7** is the result of a collaboration with J.-P. Lessard and J. D. Mireless James, and was published in *Indagationes Mathematicae* under the title *Computation of maximal local (un)stable manifold patches by the parameterization method* [77].
- **Chapter 8** is the result of a collaboration with J.-P. Lessard, and has been submitted for publication under the title *Polynomial interpolation and a priori bootstrap for computer-assisted proofs in nonlinear ODEs* [75].
- **Chapter 9** is the result of a collaboration with J. B. van den Berg, J.-P. Lessard, and M. Murray, and has been submitted for publication under the title *Continuation of homoclinic orbits in the suspension bridge equation: a computer-assisted proof* [58]. Section 9.4 of this work is mainly due to M. Murray.
- **Chapter 10** is the result of a collaboration with R. Castelli, and has been submitted for publication under the title *Existence and instability of steady states for a triangular cross-diffusion system: a computer-assisted proof* [71].

Part II

Moments estimates for the discrete coagulation-fragmentation equations with diffusion

Chapter 4

The pure coagulation case

Abstract

This chapter is taken from [7]. We establish smoothness of moments of the solutions of discrete coagulation-diffusion systems. As key assumptions, we suppose that the coagulation coefficients grow at most sub-linearly and that the diffusion coefficients converge towards a strictly positive limit (those conditions also imply the existence of global weak solutions and the absence of gelation).

4.1 Introduction

In this paper we consider discrete coagulation systems with spatial diffusion. Coagulation models appear in a wide range of applications ranging from chemistry (e.g. the formation of polymers) over physics (aerosols, raindrops, smoke, sprays), astronomy (the formation of galaxies) to biology (haematology, animal grouping), see e.g. the surveys [18, 30, 16] and the references therein.

Following the pioneering works of Smoluchowski (see [42, 43]), we shall denote by $c_i := c_i(t,x) \in \mathbb{R}_+$ the concentration of polymers or clusters of mass/size $i \in \mathbb{N}^*$ at time t and position x. Here, we consider a smooth bounded domain Ω of $\mathbb{R}^N$ in which the clusters are confined via homogeneous Neumann conditions (in applications, we of course have $N \leq 3$). Moreover, for any positive time T, we denote by Ω_T the set $[0,T] \times \Omega$.

We assume that the concentrations c_i satisfy the following infinite (for all $i \in \mathbb{N}^* := \mathbb{N} \setminus \{0\}$) set of reaction-diffusion equations with homogeneous Neumann boundary conditions:

$$\begin{cases} \partial_t c_i - d_i \Delta_x c_i = Q_i(c) & \text{on } \Omega_T, \\ \nabla_x c_i \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ c_i(0,\cdot) = c_i^{in} & \text{on } \Omega, \end{cases} \tag{4.1}$$

where $d_i > 0$ are strictly positive diffusion coefficients, $\nu(x)$ denotes the outward unit normal vector at point $x \in \partial\Omega$ and c_i^{in} are given initial data, which are typically assumed nonnegative.

The coagulation terms $Q_i(c)$ depend on the whole sequence of concentrations $c = (c_i)_{i \in \mathbb{N}^*}$ and can be written as the difference between a gain term $Q_i^+(c)$ and a loss term $Q_i^-(c)$, which take the form

$$Q_i(c) := Q_i^+(c) - Q_i^-(c) = \frac{1}{2} \sum_{j=1}^{i-1} a_{i-j,j} c_{i-j} c_j - \sum_{j=1}^{\infty} a_{i,j} c_i c_j. \tag{4.2}$$

Here, the nonnegative parameters $a_{i,j}$ represent the coagulation coefficients of clusters of size i merging with clusters of size j, which are symmetric: $a_{i,j} = a_{j,i}$. In this work, we consider the case in which the coagulation coefficients additionally satisfy the following asymptotic behaviour:

$$\lim_{j\to\infty} \frac{a_{i,j}}{j} = 0, \quad \forall\ i \in \mathbb{N}^*, \qquad 0 \le a_{i,j} = a_{j,i}, \quad \forall\ i,j \in \mathbb{N}^*. \tag{4.3}$$

The conditions (4.3) are sufficient to provide the existence of global L^1-weak solutions (with nonnegative concentrations) to system (4.1)-(4.2) (for which the below estimate (4.8) on the mass holds), when suitable nonnegative initial data are considered, see [28].

Thanks to the symmetry assumption on the coagulation coefficients, we can write (at a formal level) the following weak formulation of the coagulation operator: For any test-sequence $(\varphi_i)_{i\in\mathbb{N}^*}$, we have

$$\sum_{i=1}^{\infty} \varphi_i\, Q_i(c) = \frac{1}{2} \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{i,j}\, c_i\, c_j\, (\varphi_{i+j} - \varphi_i - \varphi_j). \tag{4.4}$$

In the sequel, we shall systematically denote for any $k \in \mathbb{R}_+$ by

$$\rho_k(t,x) := \sum_{i=1}^{\infty} i^k\, c_i(t,x) \tag{4.5}$$

the moment of order k of the sequence of concentrations $(c_i)_{i\in\mathbb{N}^*}$ and, similarly, the moment of order k of the initial concentrations by

$$\rho_k^{in}(x) := \sum_{i=1}^{\infty} i^k\, c_i^{in}(x). \tag{4.6}$$

By taking $\varphi_i = i$ in (4.4), we see that (still at a formal level) the conservation of the total mass contained in all clusters/polymers holds, that is,

$$\forall t \ge 0, \qquad \int_\Omega \rho_1(t,x)\,dx = \int_\Omega \sum_{i=1}^{\infty} i\, c_i(t,x)\,dx = \int_\Omega \sum_{i=1}^{\infty} i\, c_i^{in}(x)\,dx = \int_\Omega \rho_1^{in}(x)\,dx. \tag{4.7}$$

It is a well-known phenomenon for coagulation models, called *gelation* (see for instance [18, 21]), that the formal conservation of the total mass (4.7) will not hold for solutions of coagulation models with sufficiently growing coagulation coefficients $a_{i,j}$ (already for space homogeneous models): When approximating the first order moment $\rho_1(t,x)$ as the cut-off limit $\lim_{K\to\infty} \sum_{i=1}^{\infty} \min\{i,K\}c_i$, then the weak formulation (4.4) for the test-sequence $\varphi_i = \min\{i,K\}$ shows that the map $t \mapsto \sum_{i=1}^{\infty} \min\{i,K\}c_i$ is non-increasing in time, and Fatou's lemma only implies that the total mass is non-increasing in time (for space homogeneous and space inhomogeneous coagulation with homogeneous Neumann boundary conditions models alike). The conservation law (4.7) can become a strict inequality for solutions of (4.1) with sufficiently growing coagulation coefficients, but we still get a natural uniform-in-time bound in $L^\infty(\mathbb{R}_+; L^1(\Omega))$ of the total mass ρ_1, namely

$$\forall t \ge 0, \qquad \int_\Omega \rho_1(t,x)\,dx \le \int_\Omega \rho_1^{in}(x)\,dx. \tag{4.8}$$

A standard way to prove the existence of weak solutions to system (4.1)-(4.2) satisfying (4.8) is to consider a sequence of truncated systems for which we can prove existence of smooth solutions. Then, using some compactness arguments, one extracts a solution of the limiting

system (4.1)-(4.2), again see for instance [28]. In this work, for any $n \in \mathbb{N}^*$, we define $c^n = (c_1^n, \dots, c_n^n)$ as the solution of the truncated problem: $\forall\ 1 \le i \le n$,

$$\begin{cases} \partial_t c_i^n - d_i \Delta_x c_i^n = Q_i^n(c^n) & \text{on } \Omega_T, \\ \nabla_x c_i^n \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ c_i^n(0,\cdot) = c_i^{in} & \text{on } \Omega, \end{cases} \tag{4.9}$$

where

$$Q_i^n(c^n) = \frac{1}{2} \sum_{j=1}^{i-1} a_{i-j,j} c_{i-j}^n c_j^n - \sum_{j=1}^{n-i} a_{i,j} c_i^n c_j^n. \tag{4.10}$$

This is now a finite system of reaction-diffusion equations with finite sums in the r.h.s, for which the existence and uniqueness of nonnegative, global and smooth solutions have already been proven (see for example Proposition 2.1 and Lemma 2.2 of [45], or [15]). Notice that for any sequence $(\varphi_i)_{i \in \mathbb{N}^*}$, we have

$$\sum_{i=1}^{n} \varphi_i\, Q_i^n(c^n) = \frac{1}{2} \sum_{i+j \le n; i,j \ge 1} a_{i,j}\, c_i^n\, c_j^n\, (\varphi_{i+j} - \varphi_i - \varphi_j), \tag{4.11}$$

so that we get (this time rigorously)

$$\forall t \ge 0, \qquad \int_\Omega \sum_{i=1}^{n} i c_i^n(t,x)\, dx = \int_\Omega \sum_{i=1}^{n} i c_i^{in}(x)\, dx.$$

If we manage to extract a limit from (c_n), Fatou's Lemma then yields (4.8) for the limiting concentration.

Before proceeding further, let us introduce a precise definition of weak solution, following [28].

Definition 4.1.1. *A global weak solution $c = (c_i)_{i \in \mathbb{N}^*}$ to (4.1)-(4.2) is a sequence of functions $c_i : [0,+\infty) \times \Omega \to [0,+\infty)$ such that, for all $i \in \mathbb{N}^*$ and $T > 0$*

- $c_i \in \mathcal{C}\left([0,T]; L^1(\Omega)\right)$,
- $Q_i^-(c) \in L^1(\Omega_T)$,
- $\sup\limits_{t \ge 0} \int\limits_\Omega \rho_1(t,x)dx \le \int\limits_\Omega \rho_1^{in}(x)dx$,
- *c_i is a mild solution to the i-th equation in (4.1), that is*

$$c_i(t) = e^{d_i A_1 t} c_i^{in} + \int_0^t e^{d_i A_1 (t-s)} Q_i(c(s))ds,$$

where Q_i is defined by (4.2), A_1 is the closure in $L^1(\Omega)$ of the unbounded, linear, self-adjoint operator A of $L^2(\Omega)$ defined by

$$D(A) = \left\{ w \in H^2(\Omega),\ \nabla w \cdot \nu = 0 \text{ on } \partial\Omega \right\}, \qquad Aw = \Delta w,$$

and $e^{d_i A_1 t}$ is the $\mathcal{C}^0$-semigroup generated by $d_i A_1$ in $L^1(\Omega)$.

The following result, which is a direct application of [28, Theorem 3], states that we can obtain weak solutions of (4.1)-(4.2) from the truncated systems (4.9)-(4.10). We also refer to [45, 47].

Proposition 4.1.2. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$. Assume that the coagulation coefficients satisfy* (4.3) *and that all diffusion coefficients are strictly positive, i.e. $d_i > 0\ \forall\ i \in \mathbb{N}^*$. Assume also that the initial concentrations $c_i^{in} \geq 0$ are such that $\rho_1^{in} \in L^1(\Omega)$. For every $n \in \mathbb{N}^*$, let $c^n = (c_1^n, \ldots, c_n^n)$ be the solution of the truncated system of size n* (4.9)-(4.10).

Then, there exists a sequence $c = (c_i)_{i\in\mathbb{N}^*}$ such that, up to extraction

$$c_i^n \underset{n\to\infty}{\longrightarrow} c_i \quad \text{in } L^1(\Omega_T), \quad \forall\ i \in \mathbb{N}^*,\ \forall\ T > 0,$$

and c is a weak solution to (4.1)-(4.2) *in the sense of Definition 4.1.1.*

Our first proposition states that if the diffusion rates of clusters of different sizes are sufficiently close to each others, the natural uniform L^1-bound (4.8) can be extended to L^p (with $p > 1$ depending on the closeness of the diffusion rates). To be more precise about this closeness hypothesis, let us first introduce

Definition 4.1.3. *For $m > 0$ and $q \in]1, +\infty[$, we define $\mathcal{K}_{m,q} > 0$ as the best (i.e. the smallest) constant independent of $T > 0$ in the parabolic regularity estimate*

$$\left(\int_{\Omega_T} |\partial_t v|^q + m^q \int_{\Omega_T} |\Delta_x v|^q \right)^{\frac{1}{q}} \leq \mathcal{K}_{m,q} \left(\int_{\Omega_T} |f|^q \right)^{\frac{1}{q}}, \quad \forall\ f \in L^q(\Omega_T),$$

where v is the unique solution of the heat equation with constant diffusion coefficient m, homogeneous Neumann boundary conditions and zero initial data:

$$\begin{cases} \partial_t v - m\Delta_x v = f & \text{on } \Omega_T, \\ \nabla_x v \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ v(0,\cdot) = 0 & \text{on } \Omega. \end{cases}$$

The existence of a such a constant $\mathcal{K}_{m,q} < \infty$ independent of the time $T > 0$ is explicitly stated in [27] provided that $\partial\Omega \in \mathbb{C}^{2+\alpha}, \alpha > 0$.

Next, we present the

Proposition 4.1.4. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$ (e.g. $\partial\Omega \in \mathbb{C}^{2+\alpha}, \alpha > 0$). Let $p \in]1, +\infty[$ and assume that the nonnegative initial data $c_i^{in} \geq 0$ have an initial mass ρ_1^{in} which lies in $L^p(\Omega)$. Assume that the coagulation coefficients satisfy* (4.3). *Assume that*

$$0 < \delta := \inf_{i\geq 1} d_i, \qquad \text{and} \qquad D := \sup_{i\geq 1} d_i < \infty. \tag{4.12}$$

Then, provided that for the Hölder conjugate p' of p holds the condition

$$\frac{D-\delta}{D+\delta}\, \mathcal{K}_{\frac{D+\delta}{2},p'} < 1, \tag{4.13}$$

there exists a weak solution of the coagulation system (4.1)-(4.2) *such that the mass ρ_1 lies in $L^p(\Omega_T)$ for any finite time $T > 0$.*

Remark 4.1.5. *Note that the above estimate was already proven in [10] in the particular case $p = 2$, even without assuming* (4.13). *In fact, the Hilbert space case $p = 2$ allows to prove the explicit bound $\mathcal{K}_{m,2} \leq 1$ (see Lemma 4.4.4), which leads to*

$$\frac{D-\delta}{D+\delta}\mathcal{K}_{\frac{D+\delta}{2},2} \leq \frac{D-\delta}{D+\delta} < 1, \tag{4.14}$$

and hypothesis (4.13) *is automatically satisfied for $p = 2$ for all $0 < \delta \leq D < \infty$ and $T > 0$ (hence its absence in [10]).*

Note that this global L^2-bound together with assumptions (4.3) *also ensures that no gelation can occur, so that the conservation law* (4.7) *rigorously holds for any weak solution, see [10].*

Moreover, the strict inequality in (4.14) *has been further exploited in [11] by proving a continuous upper bound of the best constant $\mathcal{K}_{m,p'}$ on $p' \leq 2$. Therefore, for all $0 < \delta \leq D < \infty$, there exists a sufficiently small $0 < \varepsilon = \varepsilon(\delta, D) \ll 1$ such that* (4.14) *can be slightly improved to*

$$\frac{D-\delta}{D+\delta}\,\mathcal{K}_{\frac{D+\delta}{2},2-O(\varepsilon)} < 1,$$

and this allows to prove a correspondingly improved a priori estimate in $L^{2+\varepsilon}(\Omega_T)$.

Proposition 4.1.4 can be improved in the case when the diffusion coefficients $(d_i)_{i\in\mathbb{N}^*}$ constitute a sequence converging towards a strictly positive limit. Note that such an assumption is not so far from the assumption that the sequence $(d_i)_{i\in\mathbb{N}^*}$ is bounded above and below (by a strictly positive constant), which is used in Proposition 4.1.4 (or also in [10]), since one expects on physical grounds that the sequence $(d_i)_{i\in\mathbb{N}^*}$ is decreasing; that is, that larger clusters diffuse less. Under this assumption and provided that the coagulation coefficients are strictly sublinear (see the precise assumption in Theorem 4.1.6 below) we can show that L^p norms of moments ρ_k are propagated for any $k \in \mathbb{N}^*$, $p \in]1, \infty[$.

Theorem 4.1.6. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$. Assume that the coagulation coefficients satisfy for a constant $C > 0$ and all $i, j \in \mathbb{N}^*$*

$$0 \leq a_{i,j} = a_{j,i} \leq C\,(i^\gamma + j^\gamma), \quad \text{for some } \gamma \in [0, 1[, \tag{4.15}$$

and that $(d_i)_{i\in\mathbb{N}}$ is a sequence of strictly positive real numbers which converges toward a strictly positive limit.

Assume that (for some $k \in \mathbb{N}^*$) the initial moment ρ_k^{in} lies in $L^p(\Omega)$ for all $p < +\infty$ and that (for all $i \in \mathbb{N}^*$) each initial concentration $c_i^{in} \geq 0$ lies in $L^\infty(\Omega)$.

Then, there exists a global weak nonnegative solution to (4.1)-(4.2) *for which the moment ρ_k lie in $L^p(\Omega_T)$ for all $p < +\infty$ and all finite time $T > 0$.*

Remark 4.1.7. *Notice that hypothesis* (4.15) *on the coagulation coefficients implies the assumption* (4.3)*, which in return yields existence of global weak solutions.*

In the existing literature, sublinear assumptions on the coagulation coefficients are often found under the form:

$$0 \leq a_{i,j} = a_{j,i} \leq \tilde{C}\,(i^\alpha j^\beta + i^\beta j^\alpha), \quad \text{for some } \alpha, \beta \in [0, 1[\text{ with } \alpha + \beta < 1. \tag{4.16}$$

Our motive for using assumption (4.15) *rather than assumption* (4.16) *is mainly that it allows for slightly shorter computations, without any loss of generality since* (4.16) *implies* (4.15) *with $\gamma = \alpha + \beta$.*

Remark 4.1.8. *We point out that Theorem 4.1.6 could be extended to the case where a finite number of diffusion coefficients d_i are equal to 0 (see Remark 4.3.7).*

In the case that $d_i = 0$ starting from some $i = I$ as considered for instance in [46], our approach should allow for a corresponding generalisation of Theorem 4.1.6 provided that a suitable "closeness" condition on the finitely many non-zero diffusion coefficients is satisfied.

The study of moments for the coagulation equation has been a longstanding strategy to get mass conservation and uniqueness results (see [36] for one of the first work in this direction for the coagulation equation with diffusion, in the continuous case).

More recently, results in the same spirit as Theorem 4.1.6 about propagation of moments have been obtained in [39] and [40], where the system (4.1)-(4.2) and its continuous counterpart

are studied on the whole space $\mathbb{R}^N$. Assuming (4.12) and a finite total increase of variation for (d_i), together with a control on the growth of the coagulation coefficients (also involving the diffusion rates) such as

$$\frac{a_{i,j}}{(i+j)(d_i+d_j)} \xrightarrow[i+j\to+\infty]{} 0,$$

and the finiteness of some of the initial moments (in different norms), bounds are obtained which look like

$$\|\rho_k(t,\cdot)\|_{L^p(\Omega)} \leq \left\|\rho_k^{in}\right\|_{L^p(\Omega)} + Ck^{-l},$$

where l depends on the degree of the initial moments assumed to be finite.

The statement of our result is therefore close to that of [40] (our requirement on the diffusion rate is however more stringent), but the proof is completly different, so that the exact conditions required on initial data are also different. Note that the limit case $a_{i,j} = i+j$ is still open (absence of gelation for this coagulation coefficient is conjectured in general, but is proven only when there is no diffusion, see for instance [44, 3]).

Although in the present work, L^p estimates for moments are only shown for $p < \infty$ (whereas $p = \infty$ can be obtained in [40]), the use of parabolic inequalities for the heat equation enables to recover this case (and also higher order derivatives).

Indeed, the estimates obtained in Theorem 4.1.6 can be improved if the initial data are assumed to be smooth enough. This leads to our main Theorem, namely:

Theorem 4.1.9. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$. Assume that the coagulation coefficients satisfy (4.15) and that $(d_i)_{i\in\mathbb{N}}$ is a sequence of strictly positive real numbers which converges toward a strictly positive limit.*

Assume that the initial data $c_i^{in} \geq 0$ are of class $\mathcal{C}^\infty(\overline{\Omega})$, compatible with the boundary conditions, and that for all $k \in \mathbb{N}^*$ the initial moments ρ_k^{in} are of class $\mathcal{C}^\infty(\overline{\Omega})$.

Then, there exists a unique smooth solution to (4.1)-(4.2) such that each (c_i) is nonnegative, of class $\mathcal{C}^\infty(\overline{\Omega}_T)$ for any finite time $T > 0$, and such that the moments ρ_k are also of class $\mathcal{C}^\infty(\overline{\Omega}_T)$, for any $k \in \mathbb{N}^*$.

Remark 4.1.10. *The $\mathcal{C}^\infty$ regularity down to time 0 requires of course the $\mathcal{C}^\infty$ hypothesis on the initial data. However, it can be seen in the various steps of the proof (see Section 4.4) that propagation of regularity in intermediate Sobolev spaces holds under suitable (less stringent) assumptions on the initial data.*

Since each c_i is solution of a heat equation subject to a r.h.s. that can be controlled once all moments are bounded in $L^p(\Omega_T)$, $p < +\infty$, we can in fact show the creation of regularity for strictly positive times. For example, under the assumption that $\rho_k^{in} \in L^p(\Omega)$ for all $p < +\infty$ and all $k \in \mathbb{N}^*$, we can prove that the concentrations c_i are of class $C^\infty(]0,T] \times \bar{\Omega})$.

Also, as will be made clear in Section 4.4, $\mathcal{C}^\infty$ regularity is not needed to ensure uniqueness. As shown in [23], uniqueness holds as soon as $\rho_2 \in L^\infty$, so that starting from initial data leading to an estimate for ρ_2 in a Sobolev space embedded in L^∞, uniqueness can already be obtained.

Finally, we point out that assumption (4.15) is not far from optimal, since it is known that gelation can occur as soon as $a_{i,j} = i^\alpha j^\beta + i^\beta j^\alpha$ with $\alpha + \beta > 1$ (see [21]) and gelation is not compatible with the conclusion of Theorem 4.1.6 or Theorem 4.1.9.

Our paper is organized as follows. In Section 4.2, we recall some lemmas existing in the literature and called duality lemmas. We also introduce modified versions of those lemmas, that are later used in Section 4.3 to prove the propagation of moments in $L^p(\Omega_T)$, $p < +\infty$ (Propositions 4.1.4 and 4.1.6). In Section 4.4, we extend these results to prove $\mathcal{C}^\infty$ regularity for the concentrations and the moments (Theorem 4.1.9). Finally, a short Appendix is devoted to technical lemmas which are useful to make the proof of some duality lemmas rigorous.

4.2 Duality estimates

We start by recalling some *a priori* estimates based on duality arguments from [11]. These estimates are key ingredients of the present work. In this section, functions said to be weak solutions ought to be understood as solutions of the equation obtained by multiplying by a test function and integrating by parts. Remember also that $\mathcal{K}_{m,q}$ is defined in Definition 4.1.3.

The first statement recalls [11, Lemma 2.2].

Lemma 4.2.1. *Let Ω be a smooth bounded subset of $\mathbb{R}^N$ and consider a function $M := M(t,x) : \Omega_T \to \mathbb{R}_+$ satisfying $a \le M \le b$ for some $a, b > 0$. For any $q \in]1, +\infty[$, if*

$$\frac{b-a}{b+a}\,\mathcal{K}_{\frac{a+b}{2},q} < 1, \tag{4.17}$$

then, there exist constants $C_0 > 0$ and $C > 0$ (depending on Ω, a, b, q, T) such that for any $f \in L^q(\Omega)$, the (unique, weak) solution v of the backward parabolic system with L^∞ coefficient $M := M(t,x)$,

$$\begin{cases} \partial_t v + M\Delta_x v = f & \text{on } \Omega_T, \\ \nabla_x v \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ v(T,\cdot) = 0 & \text{on } \Omega, \end{cases}$$

satisfies $\|v\|_{L^q(\Omega_T)} \le C \|f\|_{L^q(\Omega_T)}$ and $\|v(0,\cdot)\|_{L^q(\Omega)} \le C_0 \|f\|_{L^q(\Omega_T)}$.

Remark 4.2.2. *The bound on $\|v\|_{L^q(\Omega_T)}$ is not explicitly mentioned in Lemma 2.2 of [11], but is a direct consequence of its proof, in particular of the estimates $\|\Delta_x v\|_{L^q(\Omega_T)} \le C_1 \|f\|_{L^q(\Omega_T)}$ and $\|\partial_t v\|_{L^q(\Omega_T)} \le C_1 \|f\|_{L^q(\Omega_T)}$, which are explicitly mentioned there.*

Remark 4.2.3. *The fact that the above mentioned function v exists (and is unique) is (in particular for $q < 2$) not obvious because M is not assumed to be continuous (or at least VMO). In the Appendix (Proposition 4.4.3), we give a proof of the existence and uniqueness of v for the sake of completeness.*

Lemma 4.2.1 is used to prove the following duality lemma, which is Proposition 1.1 of [11].

Proposition 4.2.4. *Let Ω be a smooth bounded subset of $\mathbb{R}^N$ and consider a function $M := M(t,x) : \Omega_T \to \mathbb{R}_+$ satisfying $a \le M \le b$ for some $a, b > 0$. For any $p \in]1, +\infty[$, if*

$$\frac{b-a}{b+a}\,\mathcal{K}_{\frac{a+b}{2},p'} < 1,$$

then, there exists a constant $C > 0$ (depending on Ω, a, b, p, T) such that for any $u_0 \in L^p(\Omega)$, any weak solution u of the parabolic system (in divergence form)

$$\begin{cases} \partial_t u - \Delta_x (Mu) = 0 & \text{on } \Omega_T, \\ \nabla_x u \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ u(0,\cdot) = u_0 & \text{on } \Omega, \end{cases}$$

satisfies $\|u\|_{L^p(\Omega_T)} \le C \|u_0\|_{L^p(\Omega)}$.

In the sequel we will need a generalized version of Proposition 4.2.4, which is an adaptation of Theorem 3.1 in [17] (where only the case $p = 2$ is treated).

Proposition 4.2.5. *Let Ω be a smooth bounded subset of $\mathbb{R}^N$, $\mu_1, \mu_2 \ge 0$, and consider a function $M := M(t,x) : \Omega_T \to \mathbb{R}_+$ satisfying $a \le M \le b$ for some $a, b > 0$. For any $p \in]1, +\infty[$, if*

$$\frac{b-a}{b+a}\,\mathcal{K}_{\frac{a+b}{2},p'} < 1, \tag{4.18}$$

then, there exists a constant $C > 0$ (depending on $\Omega, a, b, p, \mu_1, \mu_2, T$) such that for any $u_0 \in L^p(\Omega)$, any function $u : \Omega_T \to \mathbb{R}_+$ satisfying (weakly)

$$\begin{cases} \partial_t u - \Delta_x (Mu) \leq \mu_1 u + \mu_2 & \text{on } \Omega_T, \\ \nabla_x u \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ u(0,\cdot) = u_0 & \text{on } \Omega, \end{cases} \tag{4.19}$$

belongs to $L^p(\Omega_T)$, with the estimate:

$$\|u\|_{L^p(\Omega_T)} \leq C \left(1 + \|u_0\|_{L^p(\Omega)}\right).$$

Proof. Let φ be a nonnegative smooth function on Ω_T and v be the (unique, weak) solution (cf. Prop. 4.4.3) of the dual problem

$$\begin{cases} \partial_t v + M\Delta_x v + \mu_1 v = -\varphi & \text{on } \Omega_T, \\ \nabla_x v \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ v(T,\cdot) = 0 & \text{on } \Omega. \end{cases}$$

Notice that the function $\tilde{v}$ defined by $\tilde{v}(t,x) = v(T-t,x)$ satisfies a standard, forward in time, reaction-diffusion equation

$$\begin{cases} \partial_t \tilde{v} - \tilde{M}\Delta_x \tilde{v} = \mu_1 \tilde{v} + \tilde{\varphi} & \text{on } \Omega_T, \\ \nabla_x \tilde{v} \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ \tilde{v}(0,\cdot) = 0 & \text{on } \Omega, \end{cases}$$

where $\tilde{M}(t,x) = M(T-t,x)$ and $\tilde{\varphi}(t,x) = \varphi(T-t,x)$. This ensures that $\tilde{v}$, and therefore v, are nonnegative. Multiplying (4.19) by the solution v of the dual problem and integrating on Ω_T, we end up with

$$\int_{\Omega_T} u\,\varphi \leq \int_{\Omega} u_0 v(0) + \mu_2 \int_{\Omega_T} v \leq \|u_0\|_{L^p(\Omega)} \|v(0)\|_{L^{p'}(\Omega)} + \mu_2 \left(|\Omega|T\right)^{\frac{1}{p}} \|v\|_{L^{p'}(\Omega_T)}. \tag{4.20}$$

Moreover, the rescaled function $w = e^{\mu_1 t} v$ satisfies

$$\begin{cases} \partial_t w + M\Delta_x w = -e^{\mu_1 t}\varphi & \text{on } \Omega_T, \\ \nabla_x w \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ w(T,\cdot) = 0 & \text{on } \Omega. \end{cases}$$

Thus, provided that hypothesis (4.18) is satisfied, we can apply Lemma 4.2.1 to w and get (after noticing that $|v| \leq |w|$)

$$\|v\|_{L^{p'}(\Omega_T)} \leq C \|\varphi\|_{L^{p'}(\Omega_T)}, \quad \text{and} \quad \|v(0)\|_{L^{p'}(\Omega)} \leq C_0 \|\varphi\|_{L^{p'}(\Omega_T)},$$

where the term $e^{\mu_1 T}$ is absorbed in the constants. Returning to (4.20), we finally obtain

$$\int_{\Omega_T} u\,\varphi \leq C \left(1 + \|u_0\|_{L^p(\Omega)}\right) \|\varphi\|_{L^{p'}(\Omega_T)},$$

for all nonnegative smooth functions φ, and the statement of Proposition 4.2.5 follows by duality. □

We finish this section with another variant of the duality lemma, in which L^p r.h.s. can be treated.

Proposition 4.2.6. *Let Ω be a smooth bounded subset of $\mathbb{R}^N$ and consider a function $M := M(t,x) : \Omega_T \to \mathbb{R}_+$ satisfying $a \le M \le b$ for some $a, b > 0$. Consider functions A and B defined on Ω_T and a real number $\varepsilon \in]0,1[$. Assume that for some $p \in]1, +\infty[$, the following statements hold:*

$$\frac{b-a}{b+a}\,\mathcal{K}_{\frac{a+b}{2},p'} < 1, \qquad A \in L^{\frac{p}{\varepsilon}}(\Omega_T), \quad B \in L^p(\Omega_T). \tag{4.21}$$

Then, there exists a constant C (depending on $\Omega, a, b, p, \varepsilon, T$) such that for any $u_0 \in L^p(\Omega)$, and any nonnegative $u \in L^p(\Omega_T)$ satisfying (weakly)

$$\begin{cases} \partial_t u - \Delta_x(Mu) \le A\,u^{1-\varepsilon} + B & \text{on } \Omega_T, \\ \nabla_x u \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ u(0,\cdot) = u_0 & \text{on } \Omega, \end{cases} \tag{4.22}$$

the following estimate holds:

$$\|u\|^p_{L^p(\Omega_T)} \le C\left(\|u_0\|^p_{L^p(\Omega)} + \|A\|^{\frac{p}{\varepsilon}}_{L^{\frac{p}{\varepsilon}}(\Omega_T)} + \|B\|^p_{L^p(\Omega_T)}\right).$$

.

Remark 4.2.7. *We stress the fact that Proposition 4.2.6 above requires a priori that the function u lies in $L^p(\Omega_T)$. As a consequence, we shall not be able to directly apply this result to weak solutions of (4.1)-(4.2), but only to solutions of an approximate (truncated) system (such as (4.10)), for which we have a priori regularity estimates.*

Proof. We consider v the solution (whose existence and uniqueness are once again given by Proposition 4.4.3 of the Appendix) of the dual problem

$$\begin{cases} \partial_t v + M\Delta_x v = -u^{p-1} & \text{on } \Omega_T, \\ \nabla_x v \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ v(T,\cdot) = 0 & \text{on } \Omega. \end{cases}$$

Again multiplying (4.22) by v and integrating on Ω_T, we end up with

$$\int\limits_{\Omega_T} u^p \le \int\limits_{\Omega} u_0\, v(0) + \int\limits_{\Omega_T} A u^{1-\varepsilon} v + \int\limits_{\Omega_T} B v. \tag{4.23}$$

Moreover thanks to (4.21), we can apply Lemma 4.2.1 to the above dual problem and get that

$$\|v\|_{L^{p'}(\Omega_T)} \le C\left\|u^{p-1}\right\|_{L^{p'}(\Omega_T)} \le C\,\|u\|^{p-1}_{L^p(\Omega_T)} \quad \text{and} \quad \|v(0)\|_{L^{p'}(\Omega)} \le C_0\,\|u\|^{p-1}_{L^p(\Omega_T)}.$$

Next, returning to (4.23), we can bound each term of the r.h.s. using several times Young's inequality (sometimes using a parameter $\eta > 0$): For the first term, we get

$$\int\limits_{\Omega} u_0\, v(0) \le \frac{1}{p\eta^p}\int\limits_{\Omega} u_0^p + \frac{\eta^{p'}}{p'}\int\limits_{\Omega} v(0)^{p'} \le \frac{1}{p\,\eta^p}\int\limits_{\Omega} u_0^p + \frac{C_0^{p'}\,\eta^{p'}}{p'}\int\limits_{\Omega_T} u^p,$$

while for the second one,

$$\begin{aligned} \int\limits_{\Omega_T} A u^{1-\varepsilon} v &\le \frac{1-\varepsilon}{p}\int\limits_{\Omega_T} u^p + \frac{\eta^{p'}}{p'}\int\limits_{\Omega_T} v^{p'} + \frac{\varepsilon}{p\eta^{\frac{p}{\varepsilon}}}\int\limits_{\Omega_T} A^{\frac{p}{\varepsilon}} \\ &\le \left(\frac{1-\varepsilon}{p} + \frac{C^{p'}\,\eta^{p'}}{p'}\right)\int\limits_{\Omega_T} u^p + \frac{\varepsilon}{p\,\eta^{\frac{p}{\varepsilon}}}\int\limits_{\Omega_T} A^{\frac{p}{\varepsilon}}, \end{aligned}$$

and finally for the last one,

$$\int_{\Omega_T} Bv \leq \frac{1}{p\,\eta^p} \int_{\Omega_T} B^p + \frac{\eta^{p'}}{p'} \int_{\Omega_T} v^{p'} \leq \frac{1}{p\,\eta^p} \int_{\Omega_T} B^p + \frac{C^{p'}\,\eta^{p'}}{p'} \int_{\Omega_T} u^p.$$

Putting everything together, we end up with

$$\int_{\Omega_T} u^p \leq \left(\frac{1-\varepsilon}{p} + \frac{(2\,C^{p'} + C_0^{p'})\eta^{p'}}{p'} \right) \int_{\Omega_T} u^p + \frac{1}{p\eta^p} \int_{\Omega} u_0^p + \frac{\varepsilon}{p\eta^{\frac{p}{\varepsilon}}} \int_{\Omega_T} A^{\frac{p}{\varepsilon}} + \frac{1}{p\eta^p} \int_{\Omega_T} B^p,$$

and by taking $\eta > 0$ small enough, we get the announced estimate. □

4.3 Propagation of moments in L^p norms

This Section is devoted to the proof of propagation in $L^p(\Omega_T)$ $(p < +\infty)$ of moments ρ_k. We begin with Proposition 4.1.4 and the propagation of the total mass ρ_1, when the *closeness* hypothesis (4.13) on the diffusion coefficients is satisfied.

Proof of Proposition 4.1.4. For $n \in \mathbb{N}^*$, we consider the solution $c^n = (c_1^n, \ldots, c_n^n)$ of (4.9)-(4.10), for which the existence and uniqueness of nonnegative, global and smooth solutions are classical (see for example Proposition 2.1 and Lemma 2.2 of [45], or [15]). Summing up the equations (4.9) for each i, we get

$$\partial_t \left(\sum_{i=1}^n i c_i^n \right) - \Delta_x \left(\sum_{i=1}^n i d_i c_i^n \right) = 0,$$

which rewrites, when

$$\rho_1^n = \sum_{i=1}^n i c_i^n \quad \text{and} \quad M_1^n := \frac{\sum_{i=1}^n i d_i c_i^n}{\sum_{i=1}^\infty i c_i^n},$$

as

$$\partial_t \rho_1^n - \Delta_x \left(M_1^n\, \rho_1^n \right) = 0.$$

Using

$$a = \inf_{i \geq 1} \{d_i\} \quad \text{and} \quad b = \sup_{i \geq 1} \{d_i\},$$

we get $a \leq M_1^n \leq b$ independently of n. Proposition 4.2.4 then yields

$$\|\rho_1^n\|_{L^p(\Omega_T)} \leq C\, \|\rho_1^n(0,\cdot)\|_{L^p(\Omega)} \leq C\, \left\|\rho_1^{in}\right\|_{L^p(\Omega)},$$

where C does not depend on n. By Proposition 4.1.2 (or respectively [28, Theorem 3]), we get a weak solution $c = (c_i)_{i \in \mathbb{N}*}$ of (4.1)-(4.2) defined by (up to extraction)

$$c_i = \lim_{n \to \infty} c_i^n,$$

and thanks to Fatou's Lemma, we see that $\|\rho_1\|_{L^p(\Omega_T)} \leq C\, \|\rho_1^0\|_{L^p(\Omega)}$. □

Remark 4.3.1. *In fact Proposition 4.1.4 would be valid for any weak solution to* (4.1)-(4.2) *such that* $\sum_{i=1}^\infty i Q_i(c) \in L^1(\Omega_T)$. *Indeed, one can then prove that*

$$\partial_t \left(\sum_{i=1}^\infty i c_i \right) - \Delta_x \left(\sum_{i=1}^\infty i d_i c_i \right) = 0 \tag{4.24}$$

holds weakly, and one can then directly apply Proposition 4.2.4 to (4.24).

The proof of Theorem 4.1.6 is a bit more involved but still based on the same idea. The outline of the proof is the following: First, we get $L^\infty(\Omega_T)$ bounds for each concentration c_i and for any finite time T. Thus, it is sufficient to prove propagation in L^p spaces for tail moments, in which we only consider concentrations c_i for i larger than some index I. Because we assumed that the d_i converge (when $i \to \infty$) towards a strictly positive real number, the closeness hypothesis (4.13) will always be satisfied for the coefficients $(d_i)_{i\geq I}$ when I is large enough. This allows us to use a similar argument as in Proposition 4.1.4 to prove the propagation in $L^p(\Omega_T)$ of the mass and then of all higher order moments.

Proof of Theorem 4.1.6. As for Proposition 4.1.4, the rigorous way to prove Theorem 4.1.6 is to get all the needed estimates on the solutions of the truncated problems (4.9)-(4.10) and then pass to the limit (when $n \to \infty$). However for a clearer exposition of the different arguments, we first derive (sometimes formally) estimates on the whole system (4.1)-(4.2) and then explain how to pass to the limit in the corresponding estimates on the truncated system. We begin with the following result (which was already noticed in [45]).

Lemma 4.3.2. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$. Assume that the coagulation coefficients satisfy* (4.3) *and that $d_i > 0$ for all $i \in \mathbb{N}^*$. Assume also that each $c_i^{in} \geq 0$ lies in $L^\infty(\Omega)$. We consider a global weak nonnegative solution of* (4.1)-(4.2) *(nonnegative meaning here that $c_i \geq 0$ for all $i \in \mathbb{N}^*$).*

Then, the concentration c_i lies in $L^\infty(\Omega_T)$ for each integer $i \in \mathbb{N}^*$ and any positive time $T > 0$.

Proof. Since

$$\partial_t c_1 - d_1 \Delta_x c_1 \leq Q_1^+(c) = 0,$$

the maximum principle for the heat equation yields that $c_1 \in L^\infty(\Omega_T)$. Then, we observe that for all $i \geq 2$,

$$\partial_t c_i - d_i \Delta_x c_i \leq Q_i^+(c).$$

Since the coagulation gain term $Q_i^+(c) = \dfrac{1}{2}\displaystyle\sum_{j=1}^{i-1} a_{i,j} c_{i-j} c_j$ involves only c_j for $j < i$, we can conclude the statement of the lemma by induction. □

The proof of Lemma 4.3.2 shows sufficient conditions under which each c_i is bounded on Ω_T, but explicit bounds computed in this way would grow very fast with i. Thus, there is little hope of obtaining a result on ρ_1 by directly using this method. However, the knowledge that any finite truncation of ρ_1 lies in $L^\infty(\Omega_T)$ enables us to prove another result of propagation of L^p norms for the mass ρ_1, where the assumption (4.13) is removed and replaced by the assumption of convergence of the diffusion coefficients d_i towards a strictly positive limit.

Lemma 4.3.3. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$. Assume that the coagulation coefficients satisfy* (4.15). *Assume also that all d_i are strictly positive, and that (d_i) converges toward a strictly positive limit. Finally, assume that each $c_i^{in} \geq 0$ lies in $L^\infty(\Omega)$ and that $\rho_1^{in} \in L^p(\Omega)$ for some $p \in]1, +\infty[$. We consider a global weak nonnegative solution of* (4.1)-(4.2) *(nonnegative meaning here that $c_i \geq 0$ for all $i \in \mathbb{N}^*$).*

Then, $\rho_1 \in L^p(\Omega_T)$, for any finite time $T > 0$.

Proof. We define

$$a^I := \inf_{i\geq I} d_i, \qquad \text{and} \qquad b^I := \sup_{i\geq I} d_i.$$

Since (d_i) converges toward a positive limit, there exists a positive integer I for all $p' \in]1, +\infty[$ such that

$$\frac{b^I - a^I}{b^I + a^I} \mathcal{K}_{\frac{a^I+b^I}{2}, p'} < 1. \tag{4.25}$$

We then consider

$$\rho_1^I := \sum_{i=I}^{\infty} i c_i, \qquad \text{and} \qquad M_1^I := \frac{\sum_{i=I}^{\infty} i d_i c_i}{\sum_{i=I}^{\infty} i c_i}.$$

Note that thanks to Lemma 4.3.2, it is enough to prove that $\rho_1^I \in L^p(\Omega_T)$ in order to conclude the proof of Lemma 4.3.3. We therefore compute (remember that $a_{i,j} = a_{j,i}$)

$$\begin{aligned}
\partial_t \rho_1^I - \Delta_x \left(M_1^I \rho_1^I \right) &= \frac{1}{2} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{i,j} c_i c_j \left((i+j) \mathbb{1}_{i+j \geq I} - i \mathbb{1}_{i \geq I} - j \mathbb{1}_{j \geq I} \right) \\
&= \frac{1}{2} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{i,j} c_i c_j \left(i \left(\mathbb{1}_{i+j \geq I} - \mathbb{1}_{i \geq I} \right) + j \left(\mathbb{1}_{i+j \geq I} - \mathbb{1}_{j \geq I} \right) \right) \\
&= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{i,j} c_i c_j i \left(\mathbb{1}_{i+j \geq I} - \mathbb{1}_{i \geq I} \right).
\end{aligned}$$

Next, by using assumption (4.15), and more precisely that $a_{i,j} \leq C \left(i^\gamma + j^\gamma \right) \leq C \left(i + j \right)$, we have

$$\begin{aligned}
\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{i,j} c_i c_j i \left(\mathbb{1}_{i+j \geq I} - \mathbb{1}_{i \geq I} \right) &\leq C \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} (i+j) \, c_i c_j i \left(\mathbb{1}_{i+j \geq I} - \mathbb{1}_{i \geq I} \right) \\
&\leq C \sum_{i=1}^{I-1} \sum_{j=1}^{\infty} i^2 c_i c_j + C \sum_{i=1}^{I-1} \sum_{j=1}^{\infty} i c_i j c_j \\
&\leq 2C \Big(\sum_{i=1}^{I-1} i^2 \, c_i \Big) \Big(\sum_{j=1}^{\infty} j \, c_j \Big).
\end{aligned}$$

Thus, we obtain

$$\partial_t \rho_1^I - \Delta_x \left(M_1^I \rho_1^I \right) \leq \psi_1 \rho_1^I + \psi_2,$$

where $\psi_1 = 2C \sum_{i=1}^{I-1} i^2 \, c_i$, and $\psi_2 = \psi_1 \sum_{i=1}^{I-1} j \, c_j$. Now thanks to Lemma 4.3.2, both ψ_1 and ψ_2 belong to $L^\infty(\Omega_T)$. Then, if we denote (for $i \in \{1, 2\}$), $\mu_i := \|\psi_i\|_{L^\infty(\Omega_T)}$, we get

$$\partial_t \rho_1^I - \Delta_x \left(M_1^I \rho_1^I \right) \leq \mu_1 \rho_1^I + \mu_2,$$

and we can conclude using Proposition 4.2.5. □

Remark 4.3.4. *Note that aside from symmetry, the above proof only requires the estimate $a_{i,j} \leq C \, i \, j$, which is a much weaker restriction on the coagulation coefficients than the "sublinear" assumption* (4.15)*. However, "strictly superlinear" coagulation is known to produce gelation already in the spatially homogeneous case. Nonetheless, it is known for the homogeneous case that adding sufficiently strong fragmentation in the model can prevent gelation even with "superlinear" coagulation (see for instance [12, 14]). Similar results in presence of diffusion, together with generalisations of some results of this paper to models including fragmentation are discussed in [6].*

Continuation of the proof of Theorem 4.1.6. We shall now prove the propagation of L^p, $(p < +\infty)$ regularity for moments of higher order. This is done still under the assumption that the diffusion coefficients d_i converge towards a strictly positive limit.

We first introduce

$$M_k^I := \frac{\sum_{i=I}^{\infty} i^k d_i c_i}{\sum_{i=I}^{\infty} i^k c_i}, \qquad a^I \le M_k^I \le b^I.$$

The proof of propagation for moments of higher order involves the *a priori* estimate established in Proposition 4.2.6. Therefore the results of Theorem 4.1.6 (for $k > 1$) only apply to such solutions, which are constructed as a limit of solutions of a truncated system. For a clearer exposition of the proof, we first perform the computations formally and then show afterwards how to conclude rigorously through the use of the truncated systems (4.9)-(4.10).

We proceed by induction and assume that (for some integer k), $\rho_l \in L^p(\Omega_T)$, for all $p < +\infty$ and every $l \le k-1$. Note that Lemma 4.3.3 ensures that the induction hypothesis holds for $k = 2$. For any $I \in \mathbb{N}^*$ (using (4.15)), we have

$$\begin{aligned}
\partial_t \rho_k^I - \Delta_x(M_k^I \rho_k^I) &\le \frac{1}{2}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{i,j} c_i c_j \left((i+j)^k \mathbb{1}_{i+j\ge I} - i^k \mathbb{1}_{i\ge I} - j^k \mathbb{1}_{j\ge I}\right) \\
&\le \frac{C}{2}\sum_{l=1}^{k-1}\binom{k}{l}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(i^\gamma + j^\gamma)\, i^l c_i\, j^{k-l} c_j \\
&\quad + \frac{C}{2}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(i^\gamma + j^\gamma) c_i c_j (i^k(\mathbb{1}_{i+j\ge I} - \mathbb{1}_{i\ge I}) + j^k(\mathbb{1}_{i+j\ge I} - \mathbb{1}_{j\ge I})) \\
&\le C\sum_{l=1}^{k-1}\binom{k}{l}\sum_{i=1}^{\infty} i^{\gamma+l} c_i \sum_{j=1}^{\infty} j^{k-l} c_j \\
&\quad + C\sum_{i=1}^{I-1} i^{\gamma+k} c_i \sum_{j=1}^{\infty} c_j + C\sum_{i=1}^{I-1} i^k c_i \sum_{j=1}^{\infty} j^\gamma c_j. \qquad (4.26)
\end{aligned}$$

We point out that an alternative way of getting (4.26), maybe more reminiscent of some earlier computations (see [12]), would be to use that

$$(i^\gamma + j^\gamma)\left((i+j)^k - i^k - j^k\right) \le C(k)\left(i^{k+\gamma-1} j + j^{k+\gamma-1} i\right).$$

In the special case $k = 2$, (4.26) yields (using $\gamma \le 1$)

$$\partial_t \rho_2^I - \Delta_x(M_2^I \rho_2^I) \le 2C\rho_{1+\gamma}\rho_1 + 2C\rho_1 \sum_{i=1}^{I-1} i^3 c_i.$$

Moreover, we can split the moment of order $1+\gamma$ between a finite part that we already control and a tail, and then bound the latter using Hölder's inequality:

$$\rho_{1+\gamma} \le \sum_{i=1}^{I-1} i^{1+\gamma} c_i + \rho_{1+\gamma}^I \le \sum_{i=1}^{I-1} i^2 c_i + \left(\rho_2^I\right)^{1-\varepsilon}\left(\rho_1^I\right)^{\varepsilon}, \qquad \text{where} \quad \varepsilon = 1-\gamma > 0.$$

Therefore, we end up with

$$\partial_t \rho_2^I - \Delta_x(M_2^I \rho_2^I) \le 2C\left(\rho_2^I\right)^{1-\varepsilon}(\rho_1)^{1+\varepsilon} + 4C\rho_1 \sum_{i=1}^{I-1} i^3 c_i, \qquad (4.27)$$

and the last term in the r.h.s. of (4.27) lies in $L^p(\Omega_T)$ for every $p < +\infty$ thanks to Lemma 4.3.2 and Lemma 4.3.3. Thus, taking I large enough, inequality (4.27) and Proposition 4.2.6 formally yield a (computable) bound for ρ_2^I (and therefore ρ_2) in $L^p(\Omega_T)$, for all $p < +\infty$.

Next, returning to the general case of moments of order $k > 2$, we estimate (4.26) as

$$\begin{aligned}\partial_t \rho_k^I - \Delta_x(M_k^I \rho_k^I) &\le kC \sum_{i=1}^{\infty} i^{\gamma+k-1} c_i \sum_{j=1}^{\infty} j c_j + C \sum_{l=1}^{k-2} \binom{k}{l} \sum_{i=1}^{\infty} i^{\gamma+l} c_i \sum_{j=1}^{\infty} j^{k-l} c_j \\ &\quad + C \sum_{i=1}^{I-1} i^{\gamma+k} c_i \sum_{j=1}^{\infty} c_j + C \sum_{i=1}^{I-1} i^k c_i \sum_{j=1}^{\infty} j^{\gamma} c_j \\ &\le kC \rho_{k-1+\gamma} \rho_1 + C \sum_{l=1}^{k-2} \binom{k}{l} \rho_{l+1} \rho_{k-l} + 2C \rho_1 \sum_{i=1}^{I-1} i^{k+1} c_i,\end{aligned}$$

where we used $\gamma \le 1$. Again, we can split the moment of order $k - 1 + \gamma$ between a finite part that we already control and a tail, and then bound the latter using Hölder's inequality:

$$\rho_{k-1+\gamma} \le \sum_{i=1}^{I-1} i^{k-1+\gamma} c_i + \rho_{k-1+\gamma}^I \le \sum_{i=1}^{I-1} i^k c_i + \left(\rho_k^I\right)^{1-\varepsilon} \left(\rho_1^I\right)^{\varepsilon}, \qquad \text{where} \quad \varepsilon = \frac{1-\gamma}{k-1} > 0.$$

Therefore, we end up with

$$\partial_t \rho_k^I - \Delta_x(M_k^I \rho_k^I) \le kC \left(\rho_k^I\right)^{1-\varepsilon} (\rho_1)^{1+\varepsilon} + C \sum_{l=1}^{k-2} \binom{k}{l} \rho_{l+1} \rho_{k-l} + (k+2) C \rho_1 \sum_{i=1}^{I-1} i^{k+1} c_i, \quad (4.28)$$

and the last two terms in the r.h.s. of (4.28) lie in $L^p(\Omega_T)$ for every $p < +\infty$ thanks to Lemma 4.3.2, Lemma 4.3.3 and the induction hypothesis. Thus, taking I large enough, inequality (4.28) and Proposition 4.2.6 would yield a (computable) bound for ρ_k^I (and therefore ρ_k) in $L^p(\Omega_T)$, for all $p < +\infty$, except that Proposition 4.2.6 also requires to *a priori* know that $\rho_k^I \in L^p(\Omega_T)$, which is not the case at this point.

In order to make the proof rigorous, we need to apply Proposition 4.2.6 to smooth solutions obtained by truncating the original system. Therefore, for an integer $n > I$, we consider $c^n = (c_i^n, \dots, c_n^n)$ the solution of (4.9)-(4.10) and define

$$\rho_k^n = \sum_{i=1}^{n} i^k c_i^n, \quad \rho_k^{n,I} = \sum_{i=I}^{n} i^k c_i^n \quad \text{and} \quad M_k^{n,I} = \frac{\sum_{i=I}^n i^k d_i c_i^n}{\sum_{i=I}^n i^k c_i^n}.$$

We then perform the same computations as previously, taking into account the truncation in the coagulation kernel (4.10). We get for the second order moment (and $n > I$)

$$\begin{aligned}\partial_t \rho_2^{n,I} - \Delta_x(M_2^{n,I} \rho_2^{n,I}) &\le 2C \left(\rho_2^{n,I}\right)^{1-\varepsilon} (\rho_1^n)^{1+\varepsilon} + 4C \rho_1^n \sum_{i=1}^{I-1} i^3 c_i^n \\ &= A_1^n \left(\rho_2^{n,I}\right)^{1-\varepsilon} + B_1^{n,I},\end{aligned}$$

where $\varepsilon = 1 - \gamma > 0$ and A_1^n and $B_1^{n,I}$ only depend on the approximating first order moment ρ_1^n and on a finite number of approximate concentrations c_i^n, for $i < I$. Since this time we know that $\rho_2^{n,I} \in L^p(\Omega_T)$, we can apply Proposition 4.2.6 and get the estimate

$$\begin{aligned}\int_{\Omega_T} \left(\rho_2^{n,I}\right)^p &\le C \left(\int_{\Omega} \left(\rho_2^{n,I}(0)\right)^p + \int_{\Omega_T} (A_1^n)^{\frac{p}{\varepsilon}} + \int_{\Omega_T} \left(B_1^{n,I}\right)^p \right) \\ &\le C \left(\int_{\Omega} \left(\rho_2^{in}\right)^p + \int_{\Omega_T} (A_1^n)^{\frac{p}{\varepsilon}} + \int_{\Omega_T} \left(B_1^{n,I}\right)^p \right), \end{aligned} \quad (4.29)$$

where the constant C does not depend on n. In order to complete the proof, we still have to show that A_1^n and $B_1^{n,I}$ can be bounded in L^p norms uniformly-in-n.

Prior to that, we consider also the approximation of any general moments of order $k > 2$. By estimating as in the computation leading to (4.28), we obtain for the truncated moments of order $k > 2$

$$\begin{aligned}\partial_t \rho_k^{n,I} - \Delta_x \left(M_k^{n,I} \rho_k^{n,I} \right) &\leq kC \left(\rho_k^{n,I} \right)^{1-\varepsilon} (\rho_1^n)^{1+\varepsilon} \\ &\quad + C \sum_{l=1}^{k-2} \binom{k}{l} \rho_{l+1}^n \rho_{k-l}^n + (k+2) C \rho_1^n \sum_{i=1}^{I-1} i^{k+1} c_i^n \\ &= A_{k-1}^n \left(\rho_k^{n,I} \right)^{1-\varepsilon} + B_{k-1}^{n,I},\end{aligned}$$

with $\varepsilon = \frac{1-\gamma}{k-1} > 0$, and where A_{k-1}^n and $B_{k-1}^{n,I}$ only depend on moments ρ_l^n of integer order l between 1 and $k-1$ and on a finite number of concentrations c_i^n (for $i < I$). Moreover, since $\rho_k^{n,I} \in L^p(\Omega_T)$, we can again apply Proposition 4.2.6 and estimate

$$\int_{\Omega_T} \left(\rho_k^{n,I} \right)^p \leq C \left(\int_\Omega \left(\rho_k^{in} \right)^p + \int_{\Omega_T} \left(A_{k-1}^n \right)^{\frac{p}{\varepsilon}} + \int_{\Omega_T} \left(B_{k-1}^{n,I} \right)^p \right), \tag{4.30}$$

where the constant C does not depend on n. So we again have to prove that A_{k-1}^n and $B_{k-1}^{n,I}$ can be bounded in L^p norms uniformly-in-n.

First we notice that for any given i, since $c_i^{in} \in L^\infty(\Omega)$, the concentrations c_i^n can be bounded in $L^\infty(\Omega_T)$ uniformly-in-n by the computations of Lemma 4.3.2. Indeed,

$$\partial_t c_1^n - d_1 \Delta_x c_1^n \leq Q_1^{+,n}(c^n) = 0$$

yields a uniform-in-n bound for c_1^n, and then

$$\partial_t c_i^n - d_i \Delta_x c_i^n \leq \frac{1}{2} \sum_{j=1}^{i-1} a_{i-j,j} c_{i-j}^n c_j^n$$

allows to conclude inductively in i for all $T > 0$. Now, from that fact (that is, each c_i^n is uniformly-in-n bounded in $L^\infty(\Omega_T)$), we get that $\rho_1^{n,I}$ (and also ρ_1^n) is uniformly-in-n bounded in any L^p norm, $p < +\infty$, thanks to Lemma 4.3.3 and Proposition 4.2.5 (one just needs to repeat the computations in the proof of Lemma 4.3.3 with $\rho_1^{n,I}$ instead of ρ_1^I). Therefore A_1^n and $B_1^{n,I}$ are uniformly-in-n bounded in any $L^p(\Omega_T)$, $p < +\infty$, and going back to (4.29), this yields that $\rho_2^{n,I}$ (and thus ρ_2^n) is uniformly-in-n bounded in any $L^p(\Omega_T)$, $p < +\infty$. Similarly, we can prove inductively using (4.30) that for all $k > 2$, $\rho_k^{n,I}$ (and thus ρ_k^n) is uniformly-in-n bounded in any $L^p(\Omega_T)$, $p < +\infty$. Applying Fatou's Lemma we can conclude that, for the weak solutions given by Proposition 4.1.2, ρ_k is bounded in any $L^p(\Omega_T)$, $p < +\infty$. □

Remark 4.3.5. *Fatou's Lemma is enough here to show that $\rho_k \in L^p(\Omega_T)$, but since for any $k \in \mathbb{N}^*$ we know that ρ_{k+1}^n is bounded in $L^p(\Omega_T)$ uniformly-in-n, we could show by interpolation that we do in fact have the convergence of ρ_k^n to ρ_k in $L^p(\Omega_T)$.*

Remark 4.3.6. *Note that Theorem 4.1.6 states that we have propagation of the moment ρ_k in every $L^p(\Omega_T)$, $p < +\infty$, provided that the initial moment ρ_k^{in} lies in every $L^p(\Omega)$, $p < +\infty$. If we only want to get propagation of ρ_k in $L^p(\Omega_T)$ for some fixed p, we can relax a bit the hypothesis, but to apply the above proof we still need to assume that initial moment of lower order ρ_l^{in}, $l < k$, are in some space $L^q(\Omega)$ with $q > p$ depending of the magnitude of the coagulation. For instance, if we want to get for some fixed p that $\rho_2 \in L^p(\Omega_T)$ with the method of Theorem 4.1.6, we need to assume that $\rho_2^{in} \in L^p(\Omega)$ and $\rho_1^{in} \in L^q(\Omega)$, where $q = \frac{2-\gamma}{1-\gamma} p$.*

Remark 4.3.7. *Finally, we point out that Theorem 4.1.6 would still hold if a finite number of diffusion coefficients d_i were equal to 0. Indeed Lemma 4.3.2 is still valid in this case: if $c_j \in L^\infty(\Omega_T)$ for all $j < i$ and $d_i = 0$, then*

$$\partial_t c_i \leq Q_i^+(c) = \frac{1}{2}\sum_{j=1}^{i-1} a_{i,j} c_{i-j} c_j$$

shows that $c_i \in L^\infty(\Omega_T)$; and up to taking I large enough, we would still get $\inf_{i\geq I} d_i > 0$, so we could still apply Proposition 4.2.5 and Proposition 4.2.6 to control the tail moments ρ_k^I.

4.4 Propagation of Sobolev norms for moments

In this Section, we show how the parabolic structure of equation (4.1) can be used to improve the results of Theorem 4.1.6 and get higher regularity as stated in Theorem 4.1.9. We also explain how the obtained regularity in fact implies uniqueness. For some early work on the diffusive coagulation-fragmentation equation, using extensively the parabolic structure, we refer the reader to [2].

Proof of Theorem 4.1.9. We consider a solution provided by Theorem 4.1.6, for which we already know that we have propagation of moments in L^p spaces. Remembering (4.1), we want to use the properties of the heat equation to get additional regularity, and to do so we first need to estimate the coagulation term. This is the content of the following lemma.

Lemma 4.4.1. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$ and $s \in \mathbb{N}$. Assume that $(c_i)_{i\in\mathbb{N}^*}$ is a sequence of positive functions defined on Ω_T such that*

$$\sup_{i\geq 1} \left\| i^k c_i \right\|_{W^{s,p}(\Omega_T)} < +\infty, \quad \forall k \in \mathbb{N},\ \forall p < +\infty. \tag{4.31}$$

Then, assuming (4.2) *and* (4.15)*, the following estimates hold:*

$$\sup_{i\geq 1} \left\| i^k Q_i(c) \right\|_{W^{s,p}(\Omega_T)} < +\infty, \quad \forall k \in \mathbb{N},\ \forall p < +\infty. \tag{4.32}$$

Remark 4.4.2. *We remark that analog statements to* (4.31) *and* (4.32) *could be made in terms of fractional Sobolev spaces, for instance by interpolation arguments.*

Proof. Remembering (4.2) and using the sublinearity of the coagulation coefficients (4.15) ($a_{i,j} \leq C\,(i^\gamma + j^\gamma) \leq C\,(i+j) \leq 2C\,ij$), we estimate

$$\left\| i^k Q_i(c) \right\|_{W^{s,p}(\Omega_T)} \leq C \sum_{j=1}^{i-1} i^{k+1} \left\| c_{i-j} c_j \right\|_{W^{s,p}(\Omega_T)} + 2C \left\| i^{k+1} c_i \sum_{j=1}^{\infty} j c_j \right\|_{W^{s,p}(\Omega_T)}.$$

We now use the algebra property of $\bigcap_{p<+\infty} W^{s,p}(\Omega_T)$. More precisely, by combining Cauchy-Schwarz's inequality and Leibniz's formula, the following estimate holds:

$$\|uv\|_{W^{s,p}(\Omega_T)} \leq C(s) \|u\|_{W^{s,2p}(\Omega_T)} \|v\|_{W^{s,2p}(\Omega_T)},$$

where $C(s)$ is a constant depending only on s. Therefore,

$$\begin{aligned}\left\|i^k Q_i(c)\right\|_{W^{s,p}(\Omega_T)} &\leq C(s)\Bigg(\sum_{j=1}^{i-1} i^{k+1}\left\|c_{i-j}\right\|_{W^{s,2p}(\Omega_T)}\left\|c_j\right\|_{W^{s,2p}(\Omega_T)} \\ &\qquad + \left\|i^{k+1}c_i\right\|_{W^{s,2p}(\Omega_T)}\sum_{j=1}^{\infty}\left\|jc_j\right\|_{W^{s,2p}(\Omega_T)}\Bigg) \\ &= C(s)\Bigg(\sum_{j=1}^{i-1}\frac{i^{k+1}}{(i-j)^{k+2}j^{k+2}}\left\|(i-j)^{k+2}c_{i-j}\right\|_{W^{s,2p}(\Omega_T)}\left\|j^{k+2}c_j\right\|_{W^{s,2p}(\Omega_T)} \\ &\qquad + \left\|i^{k+1}c_i\right\|_{W^{s,2p}(\Omega_T)}\sum_{j=1}^{\infty}\frac{1}{j^2}\left\|j^3 c_j\right\|_{W^{s,2p}(\Omega_T)}\Bigg).\end{aligned}$$

Using (4.31), we get

$$\left\|i^k Q_i(c)\right\|_{W^{s,p}(\Omega_T)} \leq C(s,p,k)\left(1+\sum_{j=1}^{i-1}\frac{i^{k+1}}{(i-j)^{k+2}j^{k+2}}\right),$$

where $C(s,p,k)$ depends on the quantities $\sup_{j\geq 1}\left\|j^l c_j\right\|_{W^{s,2p}(\Omega_T)}$ for $l\in\mathbb{N}$, but not on i. We then show that (for any $k\in\mathbb{N}$)

$$\sup_{i\geq 1}\sum_{j=1}^{i-1}\frac{i^{k+1}}{(i-j)^{k+2}j^{k+2}} < \infty.$$

Indeed, by symmetry, we know that (denoting by $[m]$ the integer part of m)

$$\begin{aligned}\sup_{i\geq 1}\sum_{j=1}^{i-1}\frac{i^{k+1}}{j^{k+2}(i-j)^{k+2}} &\leq 2\sup_{i\geq 1}\sum_{j=1}^{\left[\frac{i}{2}\right]}\frac{i^{k+1}}{j^{k+2}(i-j)^{k+2}} \leq 2\sup_{i\geq 1}\sum_{j=1}^{\left[\frac{i}{2}\right]}\left(\frac{i}{i-\left[\frac{i}{2}\right]}\right)^{k+2}\frac{1}{j^{k+2}} \\ &\leq 2^{k+3}\sum_{j=1}^{\infty}\frac{1}{j^{k+2}} < \infty.\end{aligned}$$

This implies that

$$\sup_{i\geq 1}\left\|i^k Q_i(c)\right\|_{W^{s,p}(\Omega_T)} < +\infty, \quad \forall k\in\mathbb{N},\ \forall p<+\infty,$$

and Lemma 4.4.1 is proven. □

Continuation of the proof of Theorem 4.1.9. Now, we can show that under the hypothesis of Theorem 4.1.9, the concentrations (c_i) considered in Theorem 4.1.6 satisfy

$$\sup_{i\geq 1}\left\|i^k c_i\right\|_{W^{s,p}(\Omega_T)} < +\infty, \quad \forall k\in\mathbb{N},\ \forall p<+\infty,\ \forall s\in\mathbb{N}. \tag{4.33}$$

We shall prove (4.33) by induction on s. The case $s=0$ is a direct consequence of Theorem 4.1.6. Then, if for some $s\in\mathbb{N}$,

$$\sup_{i\geq 1}\left\|i^k c_i\right\|_{W^{s,p}(\Omega_T)} < +\infty, \quad \forall k\in\mathbb{N},\ \forall p<+\infty,$$

we see that Lemma 4.4.1 yields the estimate

$$\sup_{i\geq 1}\left\|i^k Q_i(c)\right\|_{W^{s,p}(\Omega_T)} < +\infty, \quad \forall k\in\mathbb{N},\ \forall p<+\infty.$$

Remembering that $i^k\, c_i$ satisfies

$$\begin{cases} (\partial_t - d_i\Delta_x)\, i^k c_i = i^k Q_i(c) & \text{on } \Omega_T, \\ \nabla_x(i^k c_i)\cdot\nu = 0 & \text{on } [0,T]\times\partial\Omega, \\ i^k c_i(0,\cdot) = i^k c_i^{in} & \text{on } \Omega, \end{cases}$$

and using the regularising properties of the heat equation (they can be used uniformly w.r.t. i since the diffusion rates d_i are bounded above and below by strictly positive constants), we get the estimate

$$\sup_{i\geq 1} \left\| i^k c_i \right\|_{W^{s+1,p}(\Omega_T)} < +\infty, \quad \forall k\in\mathbb{N},\ \forall p < +\infty.$$

This concludes the proof of (4.33). Notice that we also get $W^{s,p}$ estimates for polynomial moments of any order, because

$$\|\rho_k\|_{W^{s,p}(\Omega_T)} \leq \sum_{i=1}^{\infty} \frac{1}{i^2} \left\| i^{k+2} c_i \right\|_{W^{s,p}(\Omega_T)} \leq \sup_{i\geq 1} \left\| i^{k+2} c_i \right\|_{W^{s,p}(\Omega_T)} \sum_{i=1}^{\infty} \frac{1}{i^2}.$$

The $\mathcal{C}^\infty$ regularity announced in Theorem 4.1.9 is now a straightforward consequence of Sobolev embeddings (note that while Ω is assumed to be smooth, $\Omega_T = \Omega\times]0,T[$ will be only of Lipschitz regularity, but this is enough to apply the required Sobolev embeddings (see for instance [1])). The uniqueness of such a smooth solution is given by a straightforward extension of [23, Theorem 1.4] to the case of bounded smooth domains with Neumann boundary conditions (the uniqueness Theorem of [23] is stated when the spatial domain is $\mathbb{R}^N$), where it is proven that there cannot exist more than one weak solution to (4.1)-(4.2) satisfying $\rho_2 \in L^\infty(\Omega_T)$, as soon as the coagulation coefficients satisfy $a_{i,j} \leq Cij$, which is implied by (4.15). □

Appendix

This section is devoted to the proof of the existence of solutions of dual problems such as

$$\begin{cases} \partial_t w + M\Delta_x w = f & \text{on } \Omega_T, \\ \nabla_x w\cdot\nu = 0 & \text{on }]0,T[\times\partial\Omega, \\ w(T,\cdot) = 0 & \text{on } \Omega, \end{cases}$$

when f lies in some $L^q(\Omega_T)$, provided that there exist constants $0 < a \leq b$ (sufficiently close from one another in the sense of hypothesis (4.17)) such that $a \leq M \leq b$. We emphasise that $M := M(t,x)$ is not assumed to be continuous.

Note that with the change of time variable $\tau = T - t$, the above dual problem becomes a forward heat equation with homogeneous initial data:

$$\begin{cases} \partial_t v - M\Delta_x v = -f & \text{on } \Omega_T, \\ \nabla_x v\cdot\nu = 0 & \text{on }]0,T[\times\partial\Omega, \\ v(0,\cdot) = 0 & \text{on } \Omega. \end{cases} \tag{4.34}$$

In the case of parabolic equations in divergence form, i.e. when

$$\partial_t v - \operatorname{div}_x\left(A(t,x)\nabla_x v\right) = f, \tag{4.35}$$

it is well known that the strict ellipticity property

$$\xi^{\mathrm{T}}\, A(t,x)\,\xi \geq \lambda\,|\xi|^2, \quad \lambda > 0, \quad \forall (t,x)\in\Omega_T, \quad \forall \xi\in\mathbb{R}^N,$$

guarantees the existence of weak solutions even if A is only in L^∞, see e.g. [26].

However, we are here interested in parabolic equations in non-divergence form, i.e.

$$\partial_t v - \sum_{i,j} A_{i,j}(t,x)\,\partial^2_{i,j} v = f, \tag{4.36}$$

and we recall that when $A(t,x) = (A_{i,j}(t,x))_{i,j}$ is not smooth, the two formulations (4.35) and (4.36) are not equivalent.

Unfortunately, the existence theory for parabolic equations with discontinuous coefficients is much less developed in the non-divergence case (4.36) than in the divergence case (4.35), and some additional assumptions on A (that is, stronger than strict ellipticity) are needed.

One class of available results for parabolic equations in non-divergence form assumes coefficients which are VMO in at least sufficiently many of the time/space variables, see e.g. [25, 22] for references which consider $f \in L^q(\Omega_T)$ with $1 < q < 2$.

Another approach, in [33], consist in treating equation (4.36) as a perturbation of the standard heat equation

$$\partial_t v - \Delta_x v = f,$$

and existence of a unique weak solution is then proven under the so called *Cordes condition*

$$\left\| \frac{\sum_{1\le i,j\le N} A_{i,j}^2 + 1}{\left(\sum_{1\le i\le N} a_{i,i} + 1\right)^2} \right\|_{L^\infty(\Omega_T)} < \frac{1}{N}, \tag{4.37}$$

which in some sense measures how far A is from the identity matrix and explicitly involves the dimension N.

The equation appearing in (4.34) is less general than the one treated in [33], since we only consider matrices A of the form

$$A_{i,j}(t,x) = \delta_{i,j} M(t,x),$$

but unfortunately assumption (4.37) may not be satisfied. We can however adapt the techniques of [33] to get a proof of existence (and uniqueness) under our assumptions on M. The main idea consists in considering $\partial_t - M\,\Delta_x$ as a perturbation of $\partial_t - m\,\Delta_x$ (instead of $\partial_t - \Delta_x$), where m can be seen as a mean value of M. This is the content of the following:

Proposition 4.4.3. *Let Ω be a bounded smooth subset of $\mathbb{R}^N$ and consider $M : \Omega_T \to \mathbb{R}_+$ satisfying $a \le M \le b$ for some $a, b > 0$, and $f \in L^q(\Omega_T)$. Assume that the closeness condition* (4.17) *holds.*

Then, there exists a unique $u \in L^q(]0,T[; W^{2,q}(\Omega)) \cap W^{1,q}(]0,T[; L^q(\Omega))$ such that

$$\begin{cases} \partial_t u - M\Delta_x u = f & \text{on } \Omega_T, \\ \nabla_x u \cdot \nu = 0 & \text{on }]0,T[\times\partial\Omega, \\ u(0,\cdot) = 0 & \text{on } \Omega. \end{cases} \tag{4.38}$$

Proof. We first rewrite the equation as a perturbation of a heat equation with constant diffusion coefficient:

$$\partial_t u - m\Delta_x u = -(m - M)\Delta_x u + f,$$

where $m = \frac{a+b}{2}$. Then we introduce the space

$$Z^q := \left\{ v \in L^q(]0,T[; W^{2,q}(\Omega)) \cap W^{1,q}(]0,T[; L^q(\Omega)),\ \nabla_x v \cdot \nu = 0,\ v(0,\cdot) = 0 \right\},$$

and the operator F defined on Z^q, which associates to each $v \in Z^q$ the unique solution $Fv \in Z^q$ of

$$\begin{cases} \partial_t (Fv) - m\,\Delta_x (Fv) = -(m - M)\,\Delta_x v + f & \text{on } \Omega_T, \\ \nabla_x (Fv) \cdot \nu = 0 & \text{on } [0,T]\times\partial\Omega, \\ (Fv)(0,\cdot) = 0 & \text{on } \Omega. \end{cases}$$

Proving Proposition 4.4.3 is now equivalent to showing the existence of a unique fixed point for F. We endow Z^q with the norm

$$\|v\|_{Z^q_m(\Omega_T)} = \left(\int_{\Omega_T} |\partial_t v|^q + m^q \int_{\Omega_T} |\Delta_x v|^q \right)^{\frac{1}{q}},$$

which makes Z^q a Banach space. Note that this is indeed a norm on Z^q thanks to Calderon-Zygmund inequality:

$$\int\limits_{\Omega_T} \left| D_x^2 v \right|^q \leq C \int\limits_{\Omega_T} |\Delta_x v|^q .$$

We now show that F is a contraction on $\left(Z^q, \|\cdot\|_{Z^q_m} \right)$, which will yield the existence of a unique fixed point by the contraction mapping Theorem. For any $v, w \in Z^q$, we have

$$\partial_t (Fv - Fw) - m\Delta_x (Fv - Fw) = -(m - M)\Delta_x (v - w),$$

so that remembering Definition 4.1.3,

$$\|Fv - Fw\|_{Z^q_m(\Omega_T)} \leq \frac{b-a}{2} \mathcal{K}_{m,q} \|\Delta_x (v - w)\|_{L^q(\Omega_T)} \leq \frac{b-a}{2} \frac{\mathcal{K}_{m,q}}{m} \|v - w\|_{Z^q_m(\Omega_T)} .$$

Thus, thanks to $m = \frac{a+b}{2}$ and assumption (4.17), F is a contraction. □

Note that in the Hilbert space case $q = 2$, it is easily possible (see e.g. [37]) to obtain an explicit bound on $\mathcal{K}_{m,2}$, namely $\mathcal{K}_{m,2} \leq 1$, which shows that assumption (4.17) is always satisfied for $q = 2$ (and hence for q sufficiently close to 2, see Remark 4.1.5). This is the content of the following

Lemma 4.4.4. *For all $m > 0$, we have $\mathcal{K}_{m,2} \leq 1$, see e.g. [37].*

Proof. By multiplying

$$\partial_t v - m\Delta_x v = f$$

once by $\partial_t v$, once by $-m\Delta_x v$, and adding the results, we get

$$(\partial_t v)^2 + m^2 (\Delta_x v)^2 - 2m\, \partial_t v \Delta_x v = f^2.$$

We now show that $\int_{\Omega_T} \partial_t v\, \Delta_x v \leq 0$. Integrating by parts and using the Neumann boundary conditions and the homogeneous initial data, we see indeed that

$$\int_{\Omega_T} \partial_t v\, \Delta_x v = -\int_0^T \int_\Omega \partial_t \nabla_x v \cdot \nabla_x v = -\frac{1}{2} \int_\Omega |\nabla_x v|^2 (T) \leq 0.$$

Therefore, we obtain

$$\int_{\Omega_T} (\partial_t v)^2 + m^2 \int_{\Omega_T} (\Delta_x v)^2 \leq \int_{\Omega_T} f^2,$$

so that $\mathcal{K}_{m,2} \leq 1$. □

Chapter 5

Including (possibly strong) fragmentation

Abstract

This chapter is taken from [6]. We investigate the use of so called "duality lemmas" to study the system of discrete coagulation-fragmentation equations with diffusion. When the fragmentation is strong enough with respect to the coagulation, we show that we have creation and propagation of superlinear moments. In particular this implies that strong enough fragmentation can prevent gelation even for superlinear coagulation, a statement which was only known up to now in the homogeneous setting. We also use this control of superlinear moments to extend a recent result from [7], about the regularity of the solutions in the pure coagulation case, to strong fragmentation models.

5.1 Introduction

In this work we consider the diffusive coagulation-fragmentation system describing the dynamics of clusters coalescing to build larger clusters and breaking apart into smaller pieces. Coagulation models were first introduced by Smoluchowski (see [42, 43]) and then complexified to take into account other effects like fragmentation. These models are used in numerous and diverse applications at very different scales, in physics (smoke, sprays), chemistry (polymers), or biology (hematology, animal collective behavior).

In this work we consider only discrete (in size) models, i.e. we assume that clusters can be of size $i \in \mathbb{N}^*$, and we denote by $c_i = c_i(t,x)$ the concentration of clusters of size i at time t and position x. We also assume that the clusters are confined in a smooth bounded domain Ω of $\mathbb{R}^N$. For any positive time T, we denote by Ω_T the set $[0,T] \times \Omega$. The concentrations c_i satisfy the following set of equations, for all $i \in \mathbb{N}^*$,

$$\begin{cases} \partial_t c_i - d_i \Delta_x c_i = Q_i(c) + F_i(c), & \text{on } \Omega_T, \\ \nabla_x c_i \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ c_i(0,\cdot) = c_i^{in} & \text{on } \Omega, \end{cases} \tag{5.1}$$

where $\nu(x)$ is a unit normal vector at $x \in \partial\Omega$ and the initial concentrations $c_i^{in} \geq 0$ are given. The coagulation terms $Q_i(c)$ and the fragmentation terms $F_i(c)$ respectively write:

$$Q_i(c) = Q_i^+(c) - Q_i^-(c) = \frac{1}{2}\sum_{j=1}^{i-1} a_{i-j,j} c_{i-j} c_j - \sum_{j=1}^{\infty} a_{i,j} c_i c_j, \tag{5.2}$$

$$F_i(c) = F_i^+(c) - F_i^-(c) = \sum_{j=1}^{\infty} B_{i+j}\beta_{i+j,i}c_{i+j} - B_i c_i. \tag{5.3}$$

The nonnegative parameters B_i, $\beta_{i,j}$ and $a_{i,j}$ represent the total rate B_i of fragmentation of clusters of size i, the average number $\beta_{i,j}$ of clusters of size j produced due to fragmentation of a cluster of size i, and the coagulation rate $a_{i,j}$ of clusters of size i with clusters of size j. The fragmentation of one cluster into smaller pieces should conserve mass, clusters of size 1 should not fragment further and the coagulation rates should be symmetric, so for all $i, j \in \mathbb{N}^*$, we impose

$$i = \sum_{j=1}^{i-1} j\beta_{i,j}, \quad B_1 = 0, \quad a_{i,j} = a_{j,i} \quad \text{and} \quad a_{i,j}, B_i, \beta_{i,j} \geq 0. \tag{5.4}$$

For more details on both the modeling and the applications, we refer the reader to the surveys [18, 30, 16] and the references therein.

Assumption (5.4) allows us to write (formally for any sequence (φ_i)) the weak formulation:

$$\sum_{i=1}^{\infty} \varphi_i Q_i(c) = \frac{1}{2}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{i,j}c_i c_j(\varphi_{i+j} - \varphi_i - \varphi_j), \tag{5.5}$$

$$\sum_{i=1}^{\infty} \varphi_i F_i(c) = -\sum_{i=2}^{\infty} B_i c_i \left(\varphi_i - \sum_{j=1}^{i-1} \beta_{i,j}\varphi_j\right). \tag{5.6}$$

For any $k \geq 0$, we define the moment of order k (associated to solutions (c_i) of (5.1)-(5.4)) as

$$\rho_k(t,x) = \sum_{i=1}^{\infty} i^k c_i(t,x),$$

and similarly the initial moment of order k as

$$\rho_k^{in}(x) = \sum_{i=1}^{\infty} i^k c_i^{in}(x).$$

Considering $\varphi_i = i$ in the weak formulation (5.5)-(5.6), it is easy to see that (at the formal level), the total mass $\int_\Omega \rho_1$ is conserved. Indeed we get that

$$\partial_t\left(\sum_{i=1}^{\infty} i c_i\right) - \Delta_x\left(\sum_{i=1}^{\infty} i d_i c_i\right) = 0, \tag{5.7}$$

and an integration by part yields (thanks to the homogeneous Neumann boundary conditions)

$$\frac{d}{dt}\int_\Omega \rho_1(t,x)dx = 0 \quad \text{and thus} \quad \int_\Omega \rho_1(t,x)dx = \int_\Omega \rho_1^{in}(x)dx, \quad \forall t \geq 0. \tag{5.8}$$

Let us point out that (5.8) can fail to be true if the coagulation is strong enough, the mass $\int_\Omega \rho_1$ then becoming strictly smaller than the initial mass $\int_\Omega \rho_1^{in}$ after some finite time t^*. This phenomenon is called gelation (in the homogeneous case see for instance [24], or [21] for continuous (in size) models).

In all cases, weak solutions satisfy

$$\int_\Omega \rho_1(t,x)dx \leq \int_\Omega \rho_1^{in}(x)dx, \quad \forall t \geq 0.$$

which gives a first *a priori* estimate stating that the mass ρ_1 lies in $L^1(\Omega_T)$.

Before proceeding further, let us introduce a precise definition of weak solutions, following [28], that we will use throughout this paper.

Definition 5.1.1. *A global weak solution $c = (c_i)_{i\in\mathbb{N}^*}$ to (5.1)-(5.4) is a sequence of nonnegative functions $c_i : [0,+\infty)\times\Omega \to [0,+\infty)$ such that, for all $i \in \mathbb{N}^*$ and $T > 0$*

- $c_i \in \mathcal{C}\left([0,T]; L^1(\Omega)\right)$,
- $Q_i^-(c), F_i^+(c) \in L^1(\Omega_T)$,
- $\sup\limits_{t\geq 0} \int\limits_\Omega \rho_1(t,x)dx \leq \int\limits_\Omega \rho_1^{in}(x)dx$,
- *c_i is a mild solution to the i-th equation in (5.1), that is*

$$c_i(t) = e^{d_iA_1t}c_i^{in} + \int_0^t e^{d_iA_1(t-s)}\left(Q_i(c(s)) + F_i(c(s))\right)ds,$$

where Q_i and F_i are defined by (5.2) and (5.3), A_1 is the closure in $L^1(\Omega)$ of the unbounded, linear, self-adjoint operator A of $L^2(\Omega)$ defined by

$$D(A) = \left\{w \in H^2(\Omega),\ \nabla w\cdot\nu = 0 \text{ on } \partial\Omega\right\}, \qquad Aw = \Delta w,$$

and $t \mapsto e^{d_iA_1t}$ is the $\mathcal{C}^0$-semigroup generated by d_iA_1 in $L^1(\Omega)$.

In [28], existence of weak solutions to (5.1)-(5.4) was proven under the following assumption on the asymptotic behavior of the coagulation and fragmentation coefficients:

$$\lim_{j\to\infty}\frac{a_{i,j}}{j} = 0 = \lim_{j\to\infty}\frac{B_{i+j}\beta_{i+j,i}}{i+j}, \quad \forall\ i\in\mathbb{N}^*. \tag{5.9}$$

This existence result relies on a sequence of truncated versions of the system (5.1)-(5.4), for which existence of global smooth solutions is known. Some compactness argument is then used to extract a solution of the full system (5.1)-(5.4). We detail this procedure in Section 5.4 (see also [28, 45, 47]).

Within the existence framework of [28], equation (5.7) was rewritten in [10], as

$$\partial_t\rho_1 - \Delta_x\left(M_1\rho_1\right) = 0, \quad \text{where} \quad M_1 = \frac{\sum_{i=1}^\infty id_ic_i}{\sum_{i=1}^\infty ic_i}, \tag{5.10}$$

and duality techniques were then used to get another *a priori* estimate on the mass ρ_1, namely that it lies in $L^2(\Omega_T)$, provided that

$$\inf_{i\in\mathbb{N}^*} d_i > 0 \quad \text{and} \quad \sup_{i\in\mathbb{N}^*} d_i < \infty,$$

and that the coagulation coefficients are strictly sublinear, i.e.

$$a_{i,j} \leq C(i^\alpha j^\beta + i^\beta j^\alpha), \quad \text{with } \alpha+\beta<1, \quad \forall i,j\in\mathbb{N}^*. \tag{5.11}$$

Thanks to new duality estimates from [11], it was recently proven in [7], in the pure coagulation case and still assuming (5.11), that moments ρ_k of any order in fact lie in every $L^p(\Omega_T)$, $p<+\infty$, if the initial moments ρ_0^k lie in every $L^p(\Omega)$ and under some additional assumption (see (5.12)) on the diffusion rates d_i. For other applications of these "duality estimates", see the survey [37] and the references therein.

We mention that similar results of propagation of moments were also obtained in [39, 40], with different techniques and a slightly different set of hypothesis. The main result of [7] is then to deduce propagation of smoothness of the solutions of (5.1)-(5.4) (Theorem 5.1.5 with no fragmentation) from these estimates in $L^p(\Omega_T)$ on the mass ρ_1.

In this paper, we investigate other consequences of the L^p estimates of [7], this time in presence of fragmentation. Our main theorem states that, when the fragmentation is strong

enough compared to the coagulation, we have creation and propagation of superlinear moments. This allows us to deduce that strong enough fragmentation can prevent gelation. To put this result in perspective, let us give a brief review (which is far from being exhaustive) of some of the main known results concerning the mass conserving solutions and the occurrence of gelation. For the homogeneous case, as long as the coagulation coefficients are sublinear, i.e.

$$a_{i,j} \leq C(i+j), \quad \forall i,j \in \mathbb{N}^*,$$

the solutions of (5.1)-(5.4) are mass conserving, that is (5.8) rigorously holds (see for instance [3]). In the inhomogeneous case, it was shown in [10] that mass conservation still holds at least for strictly sublinear coagulation coefficients (5.11), the limit case $a_{i,j} = i + j$ still being open. However, it is known in the homogeneous case that gelation occurs as soon as the coagulation coefficients are strictly superlinear if there is no fragmentation (see [24] for instance). Still in the homogeneous case, it was proven that having strong enough fragmentation could prevent this gelation phenomenon and ensure the conservation of mass even for superlinear coagulation coefficients (see [12, 14], and also [20] for the continuous case). Here we prove that the same result holds for the inhomogeneous system with diffusion, under an additional assumption on the diffusion rates d_i:

$$d_i > 0, \ \forall \ i \in \mathbb{N}^* \quad \text{and} \quad d_i \underset{i\to\infty}{\longrightarrow} d_\infty > 0. \tag{5.12}$$

Here is the precise statement of our result.

Theorem 5.1.2. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$, $N \leq 2$. Assume the following behavior for the coagulation and fragmentation coefficients:*

$$a_{i,j} \leq C_Q \left(i^\alpha j^\beta + i^\beta j^\alpha \right) \quad \text{and} \quad B_i \geq C_F i^\gamma, \tag{5.13}$$

where $C_Q, C_F > 0$, $0 \leq \alpha, \beta \leq 1$, $\gamma \geq 1$ and $\alpha + \beta < \gamma$. Assume also (5.12)*, together with*

$$K_i^Q := \sup_{j\in\mathbb{N}^*} \frac{a_{i,j}}{j} < +\infty, \quad K_i^F := \sup_{j\in\mathbb{N}^*} \frac{B_{i+j}\beta_{i+j,i}}{i+j} < +\infty, \quad \forall \ i \in \mathbb{N}^*. \tag{5.14}$$

Finally, assume that c_i^{in} lies in $L^\infty(\Omega)$ for all $i \geq 1$, that for some $k > 1$ also satisfying $k > 2 - (\gamma - \alpha)$ and $k > 2 - (\gamma - \beta)$, ρ_k^{in} lies in $L^1(\Omega)$, and that ρ_1^{in} lies in $L^p(\Omega)$ for all $p < \infty$.

Then there exists a weak solution to the coagulation-fragmentation system (5.1)-(5.4)*, whose moments satisfy*

$$\int_0^T t^{m-1} \int_\Omega \rho_{k+m(\gamma-1)}(t,x)dtdx < \infty, \quad \forall \ m \in \mathbb{N}^*. \tag{5.15}$$

In particular, the superlinear moment $\rho_{k+\gamma-1}$ lies in $L^1(\Omega_T)$, and thus gelation cannot occur.

Before proceeding further, let us make a few comments about the hypothesis used in Theorem 5.1.2.

- We point out that assumption (5.14) is more general than assumption (5.9), which does not hold when $\alpha = 1$ or $\beta = 1$ in (5.13). Therefore the existence of weak solution could not be obtained by simply applying the results of of [28], and we had to develop new existence results (see Theorem 5.1.3 and Proposition 5.2.1). Also the part about the coagulation in (5.14) is in fact implied by (5.13).
- It will be made apparent in the proof of Theorem 5.1.2 that the creation of moments displayed in (5.15) is obtained through an iterative procedure (on m), and that assuming $\rho_1^{in} \in L^p(\Omega)$ for all $p < \infty$ is needed only if one want to get (5.15) for all $m \in \mathbb{N}^*$. However, just assuming that $\rho_1^{in} \in L^p(\Omega)$ for a given p can be enough to get at least a bound on one superlinear moment, and thus ensure that gelation cannot happen. For instance, if ρ_1^{in} lies in $L^p(\Omega)$ for $p = 1 + \frac{\gamma+k-2}{\gamma-(\alpha+\beta)}$, $p > 2$, then the proof of Theorem 5.1.2 shows that $\rho_{k+\gamma-1} \in L^1(\Omega_T)$.

- It is also worth mentioning that in the case where ρ_1^{in} lies in $L^p(\Omega)$ for a given p, the restrictions $N \leq 2$ and $c_i^{in} \in L^\infty(\Omega)$ for all $i \geq 1$ can be dropped, provided than (5.12) is replaced by a different *closeness* assumption on the diffusion rates d_i (depending on p) which we will introduce later (5.25). This will be explained in Section 5.4 along the proof of Theorem 5.1.2 (see Remark 5.4.1).
- Finally, we point out that our assumption $\alpha + \beta < \gamma$, which prescribes how strong the fragmentation should be compared to the coagulation, is more restrictive than the one made in [14] for the homogeneous case, where only $\alpha+\beta-1 < \gamma$ is assumed. However, this difference is not related to the fact that our model includes spatial inhomogeneity, but comes from the fact that [14] only studies the case of binary fragmentation (where one cluster only breaks into exactly two smaller ones) and makes an additionnal assumption on the *distribution of the fragmentation* (represented in our cases by the coefficients $\beta_{i,j}$). If we also make an additionnal assumption on the $\beta_{i,j}$, similiar to the one of [14], we can show that Theorem 5.1.2 still holds with the weaker assumption $\alpha + \beta - 1 < \gamma$ (more details after the proof, in Remark 5.4.2).

As we pointed out earlier, to treat the strong fragmentation case we have to develop new existence results (see Proposition 5.2.1). In the process, we obtain a new proof of existence of weak solutions of (5.1)-(5.4) that is also valid in many weak fragmentation cases. This proof was already presented in [10] but only in dimension 1. Here the improved duality estimates allow us to treat any dimension. Compared to the existence result of [28], our theorem requires an additional assumption (5.16) on the diffusion coefficients and a more stringent assumption on the initial data ($\rho_1^{in} \in L^{2+\varepsilon}(\Omega)$ instead of $\rho_1^{in} \in L^1(\Omega)$), but we then get that ρ_1 lies in $L^{2+\varepsilon}(\Omega_T)$. Our proof has the advantage of being rather simple, and its other virtue is that it can be adapted for a variation of the model presented here, where fragmentation is induced by binary collisions between clusters (see [10] and the references therein), leading to new quadratic terms that do not seem to be easy to treat with the inductive approach of [28]. The precise statement of our existence result is the following:

Theorem 5.1.3. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$. Assume that the coagulation and fragmentation coefficients satisfy the asymptotic behavior* (5.9) *and that the diffusion rates are such that*

$$\inf_{i\in\mathbb{N}^*} d_i > 0 \quad \textit{and} \quad \sup_{i\in\mathbb{N}^*} d_i < +\infty. \tag{5.16}$$

Assume also that the initial data $c_i^{in} \geq 0$ are such that there exists $p > 2$ such that the initial mass ρ_1^{in} lies in $L^p(\Omega)$.

Then there exists a weak solution to the coagulation-fragmentation system (5.1)-(5.4) *such that ρ_1 lies in $L^p(\Omega_T)$ for all positive T.*

Remark 5.1.4. *We point out that the only assumption on the coagulation coefficients in Theorem 5.1.3 is* (5.9)*, therefore this result includes situations where gelation can occur.*

Finally, it was shown in [7], with sublinear coagulation and no fragmentation, that having a control on all moments ρ_k, $k \geq 1$, in all $L^p(\Omega_T)$, $p < \infty$, was enough to get a result of propagation of smoothness for solutions of (5.1)-(5.4). Since we have such a control in the strong fragmentation case thanks to Theorem 5.1.2, we are now able to complete this smoothness result so that it encompasses nearly all situations where it is known that gelation cannot happen.

Theorem 5.1.5. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$, $N \leq 2$, and $T > 0$. Assume the behavior* (5.12) *for the diffusion coefficients and assume the following behavior for the coagulation and fragmentation coefficients:*

$$a_{i,j} \leq C_Q \left(i^\alpha j^\beta + i^\beta j^\alpha \right) \quad \textit{and} \quad B_i \leq C_{max} i^{\gamma_{max}},$$

where $0 \leq \alpha, \beta \leq 1$ *and* $\gamma_{max} < \infty$. *Assume also that the initial data* $c_i^{in} \geq 0$ *are of class* $\mathcal{C}^\infty(\overline{\Omega})$, *compatible with the boundary conditions and that for all* $k \in \mathbb{N}^*$ *the initial moments* ρ_k^{in} *are of class* $\mathcal{C}^\infty(\overline{\Omega})$. *Finally, assume that we are in one of the two following cases.*

Sublinear coagulation case: (5.9) *holds and* $\alpha + \beta < 1$.

Strong fragmentation case: (5.14) *holds and there exists* $C_F > 0$ *and* $\gamma \geq 1$, $\gamma > \alpha + \beta$, *such that* $B_i \geq C_F i^\gamma$ *for all* $i \in \mathbb{N}^*$.

Then there exists a smooth solution to the coagulation-fragmentation system (5.1)-(5.4) *such that each* c_i *is of class* $\mathcal{C}^\infty(\overline{\Omega}_T)$, *and such that the moments* ρ_k *are also of class* $\mathcal{C}^\infty(\overline{\Omega}_T)$, *for any* $k \in \mathbb{N}^*$. *Besides, if* $\sup_{i \in \mathbb{N}^*} B_i < \infty$, *the solution is unique.*

Let us make a few comments about Theorem 5.1.5

- The $\mathcal{C}^\infty$ regularity down to time 0 requires of course the $\mathcal{C}^\infty$ hypothesis on the initial data. However, it can be seen in the various steps of the proof (see Section 5.5) that propagation of regularity in intermediate Sobolev spaces holds under suitable (less stringent) assumptions on the initial data.
- Since each c_i is solution of a heat equation subject to a r.h.s. that can be controlled once all moments are bounded in $L^p(\Omega_T)$, $p < +\infty$, we can in fact show the creation of regularity for strictly positive times. For example, under the assumption that $\rho_k^{in} \in L^p(\Omega)$ for all $p < +\infty$ and all $k \in \mathbb{N}^*$, we can prove that the concentrations c_i are of class $\mathcal{C}^\infty(]0,T] \times \bar{\Omega})$.
- In the strong fragmentation case, we can suppose even less, since we saw in Theorem 5.1.2 that moments bounds in $L^p(\Omega_T)$ are created. Indeed, if we simply assume $\rho_{k_0}^{in} \in L^1(\Omega)$ for a $k_0 > 1$, (5.15) yields that $\rho_k \in L^1([\varepsilon, T] \times \Omega)$ for all $\varepsilon > 0$ and all $k \geq 0$, and this is then enough to show smoothness on $]0,T] \times \Omega$.

Our paper is organised as follow. In Section 5.2 we present the truncated system we use to approximate the caogulation-fragmentation system (5.1)-(5.4). We also give sufficient conditions, in terms of *a priori* estimates on the mass ρ_1, under which the solutions of the truncated system converge (up to extraction) to weak solutions of the full system (5.1)-(5.4). In Section 5.3 we then recall some key results from [11] and [7]: improved duality lemmas that can be used to get the *a priori* estimates introduced in the previous section, allowing us to prove Theorem 5.1.3. In Section 5.4 we make further use of these *a priori* estimates, and show that they enable creation of higher order moments in the strong fragmentation case, to get the proof of Theorem 5.1.2. We conclude in Section 5.5 by showing how control on higher order moments translates into smoothness and prove Theorem 5.1.5.

5.2 Approximation scheme and existence results

In this section, we explain how to obtain weak solutions of the coagulation-fragmentation system (5.1)-(5.4) from a truncated system. For all $n \in \mathbb{N}^*$, we consider $c^n = (c_1^n, \dots, c_n^n)$ the solution of

$$\begin{cases} \partial_t c_i^n - d_i \Delta_x c_i^n = Q_i^n(c^n) + F_i^n(c^n), & \text{on } \Omega_T, \\ \nabla_x c_i^n \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ c_i^n(0,\cdot) = c_i^{in} & \text{on } \Omega, \end{cases} \tag{5.17}$$

with

$$Q_i^n(c^n) = Q_i^{+,n}(c^n) - Q_i^{-,n}(c^n) = \frac{1}{2} \sum_{j=1}^{i-1} a_{i-j,j} c_{i-j}^n c_j^n - \sum_{j=1}^{n-i} a_{i,j} c_i^n c_j^n, \tag{5.18}$$

and

$$F_i^n(c^n) = F_i^{+,n}(c^n) - F_i^{-,n}(c^n) = \sum_{j=1}^{n-i} B_{i+j}\beta_{i+j,i}c_{i+j}^n - B_i c_i^n. \tag{5.19}$$

For this reaction diffusion system (of finite dimension and with finite sums in the r.h.s.), existence and uniqueness of nonnegative, global and smooth solution are known (see for instance [45]). Notice that, if one assumes (5.4), the truncated version of the weak formulation (5.5)-(5.6) becomes (for any sequence (φ_i))

$$\sum_{i=1}^{n} \varphi_i Q_i^n(c^n) = \frac{1}{2} \sum_{\substack{1\le i,j\le n \\ i+j\le n}} a_{i,j}c_i^n c_j^n(\varphi_{i+j} - \varphi_i - \varphi_j), \tag{5.20}$$

and

$$\sum_{i=1}^{n} \varphi_i F_i^n(c^n) = -\sum_{i=2}^{n} B_i c_i^n \left(\varphi_i - \sum_{j=1}^{i-1} \beta_{i,j}\varphi_j\right). \tag{5.21}$$

For all $k \ge 0$, we then define the moment of order k, associated to the solution c^n, as

$$\rho_k^n = \sum_{i=1}^{n} i^k c_i^n.$$

We now given sufficient conditions on those moments, under which one can extract from $(c^n)_n$ a subsequence converging to a solution of (5.1)-(5.4).

Proposition 5.2.1. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$, and initial concentrations $c_i^{in} \ge 0$ such that ρ_1^{in} lies in $L^1(\Omega)$. Assume that $d_i > 0$ for all $i \in \mathbb{N}^*$, that there exists $p > 2$ such that, for all $n \in \mathbb{N}^*$, the mass $(\rho_1^n)_n$ associated to the solution c^n of (5.17)-(5.19) is bounded in $L^p(\Omega_T)$. Assume also that one of the two following statements holds:*

(i) the coagulation and fragmentation coefficients satisfy the asymptotic behavior given in (5.9),

(ii) the coagulation and fragmentation coefficients satisfy the (weaker) asymptotic behavior given in (5.14), and there exists $k > 1$ such that $(\rho_k^n)_n$ is bounded in $L^1(\Omega_T)$.

Then, up on extraction, $(c^n)_n$ converges to a weak solution c of (5.1)-(5.3), for which ρ_1 lies in $L^p(\Omega_T)$.

Remark 5.2.2. • *As mentioned in the introduction, for the case (i) a more general result was already proven in [28]. However, our proof is much simpler because we assume that we have an a priori estimate on the mass in L^p for some $p > 2$, from which we can deduce the convergence (up to extraction) of $(c_i^n)_n$ in $L^2(\Omega_T)$. We show situations where we can get this a priori estimate in the next sections. A proof similar to the one given here was also presented in [10], but it was only valid in dimension $N = 1$ because they only had an a priori estimate in L^2 for the mass.*

• *While assumption (5.9) covers a wide variety of coagulation and fragmentation coefficients, it does not hold in some of the cases of strong fragmentation (which allows for stronger coagulation) considered in Theorem 5.1.2, for instance if $\alpha = 1$ or $\beta = 1$. In that case the existence will be provided by the case (ii) of Proposition 5.2.1.*

Proof. Since $p \ge 2$, we start by noticing that, for all $i \in \mathbb{N}^*$ $(Q_i^n(c^n))_n$ and $(F_i^n(c^n))_n$ are bounded, in $L^1(\Omega_T)$ and $L^2(\Omega_T)$ respectively. Indeed, remembering the notations introduced in (5.14),

$$\begin{aligned} |Q_i^n(c^n)| &\le \frac{1}{2}\sum_{j=1}^{i-1} a_{i-j,j}\left(\left(c_{i-j}^n\right)^2 + \left(c_j^n\right)^2\right) \\ &\quad + c_i^n \sum_{j=1}^{n-i} \frac{a_{i,j}}{j} j c_j^n \le (\rho_1^n)^2\left(\sum_{j=1}^{i-1} a_{i-j,j} + K_i^Q\right) \end{aligned}$$

and

$$|F_i^n(c^n)| \leq \sum_{j=1}^{n-i} \frac{B_{i+j}\beta_{i+j,i}}{i+j}(i+j)c_{i+j}^n + B_i c_i^n \leq \rho_1^n \left(K_i^F + B_i\right).$$

Therefore $\left(\partial_t c_i^n - d_i \Delta_x c_i^n\right)_n$ is bounded in $L^1(\Omega_T)$, which yields that $\left(c_i^n\right)_n$ lies in a (strongly) compact subset of $L^1(\Omega_T)$. Thus, up to a diagonal extraction, we can assume that for each i in $\mathbb{N}^*$,

$$c_i^n \longrightarrow c_i, \quad \text{in } L^1(\Omega_T) \text{ and almost surely}, \quad \forall i \in \mathbb{N}^*.$$

Then, since $\left(c_i^n\right)_n$ is bounded in $L^p(\Omega_T)$, (by $\left(\rho_1^n\right)_n$), Fatou's Lemma yields that c_i does in fact lie in $L^p(\Omega_T)$ and since $p > 2$, we get by interpolation that

$$c_i^n \longrightarrow c_i, \quad \text{in } L^2(\Omega_T), \quad \forall i \in \mathbb{N}^*.$$

Still by Fatou's Lemma, we also have that ρ_1 lies in $L^p(\Omega_T)$.

To prove that the limit $c = (c_i)$ is a weak solution of (5.1)-(5.4), it suffices to show that $\left(Q_i^n(c^n)\right)_n$ and $\left(F_i^n(c^n)\right)_n$ converge in $L^1(\Omega_T)$ to $Q_i(c)$ and $F_i(c)$ respectively. The convergence of $\left(Q_i^{+,n}(c^n)\right)_n$ and $\left(F_i^{-,n}(c^n)\right)_n$ to Q_i^+ and F_i^- respectively is obvious, because these terms are only composed of linear or quadratic finite sums (and we have the convergence of c_i^n in L^2 so the quadratic terms do converge in L^1). For the remaining terms, we need a different argument for the cases (i) and (ii). If (i) holds, we estimate

$$\begin{aligned}
\left|Q_i^{-,n}(c^n) - Q_i^-(c)\right| &\leq \sum_{j=1}^{j_0} a_{i,j}\left|c_i^n c_j^n - c_i c_j\right| \\
&\quad + \sup_{j>j_0} \frac{a_{i,j}}{j}\left(c_i \sum_{j=j_0+1}^{\infty} j c_j + c_i^n \sum_{j=j_0+1}^{n} j c_j^n\right) \\
&\leq \sum_{j=1}^{j_0} a_{i,j}\left|c_i^n c_j^n - c_i c_j\right| + \sup_{j>j_0} \frac{a_{i,j}}{j}\left(\rho_1^2 + (\rho_1^n)^2\right),
\end{aligned}$$

and

$$\begin{aligned}
\left|F_i^{+,n}(c^n) - F_i^+(c)\right| &\leq \sum_{j=1}^{j_0} B_{i+j}\beta_{i+j,i}\left|c_{i+j}^n - c_{i+j}\right| \\
&\quad + \sup_{j>j_0} \frac{B_{i+j}\beta_{i+j,i}}{i+j}\left(\sum_{j=j_0+1}^{\infty}(i+j)c_{i+j} + \sum_{j=j_0+1}^{n}(i+j)c_{i+j}^n\right) \\
&\leq \sum_{j=1}^{j_0} B_{i+j}\beta_{i+j,i}\left|c_{i+j}^n - c_{i+j}\right| + \sup_{j>j_0} \frac{B_{i+j}\beta_{i+j,i}}{i+j}\left(\rho_1 + \rho_1^n\right).
\end{aligned}$$

Since $\left(\rho_1^n\right)_n$ is bounded in $L^p(\Omega_T)$, $p \geq 2$, by (5.9) $\sup\limits_{j>j_0} \frac{a_{i,j}}{j}\left(\rho_1^2 + (\rho_1^n)^2\right)$ can be made arbitrarily small by taking j_0 large enough. Once j_0 is fixed, we can then pass to the limit in the finite sum, using that $\left(c_i^n\right)_n$ converges to c_i in L^2, to get that $\left(Q_i^{-,n}(c^n)\right)_n$ converges in $L^1(\Omega_T)$ to $Q_i^-(c)$. By a similar argument, we get the convergence of $\left(F_i^{+,n}(c^n)\right)_n$ to $F_i^+(c)$ in $L^1(\Omega_T)$. This concludes the proof in the case (i).

If (ii) holds instead of (i), we can only get

$$\left|Q_i^{-,n}(c^n) - Q_i^-(c)\right| \leq \sum_{j=1}^{j_0} a_{i,j}\left|c_i^n c_j^n - c_i c_j\right| + K_i^Q\left(c_i \sum_{j=j_0+1}^{\infty} j c_j + c_i^n \sum_{j=j_0+1}^{n} j c_j^n\right),$$

and

$$\left|F_i^{+,n}(c^n) - F_i^+(c)\right| \leq \sum_{j=1}^{j_0} B_{i+j}\beta_{i+j,i}\left|c_{i+j}^n - c_{i+j}\right| + K_i^F\left(\sum_{j=j_0+1}^{\infty}(i+j)c_{i+j} + \sum_{j=j_0+1}^{n}(i+j)c_{i+j}^n\right).$$

However, the knowledge that a superlinear moment (i.e. ρ_k^n for $k>1$) is bounded in $L^1(\Omega_T)$, will allow us to control (uniformly in n) reminders terms of the form $\sum_{j=j_0+1}^n jc_j^n$. Indeed, we already know that $\sum_{j=j_0+1}^{\infty} jc_j$ can be made arbitrarily small in $L^2(\Omega_T)$ by taking j_0 large enough, since ρ_1 lies in $L^p(\Omega_T)$ with $p\geq 2$. To show that $\sum_{j=j_0+1}^n jc_j^n$ can also be made arbitrarily small in $L^2(\Omega_T)$ (uniformly in n), it is sufficient to prove that $(\rho_1^n)_n$ converges to ρ_1 in $L^2(\Omega_T)$. To do so, we estimate

$$\begin{aligned}\int_{\Omega_T}|\rho_1-\rho_1^n| &\leq \sum_{i=1}^{i_0-1} i\int_{\Omega_T}|c_i-c_i^n| + \int_{\Omega_T}\sum_{i=i_0}^{\infty} ic_i + \int_{\Omega_T}\sum_{i=i_0}^{n} ic_i^n\\ &\leq \sum_{i=1}^{i_0-1} i\int_{\Omega_T}|c_i-c_i^n| + \frac{1}{i_0^{k-1}}\left(\int_{\Omega_T}\sum_{i=i_0}^{\infty} i^k c_i + \int_{\Omega_T}\sum_{i=i_0}^{n} i^k c_i^n\right)\\ &\leq \sum_{i=1}^{i_0-1} i\int_{\Omega_T}|c_i-c_i^n| + \frac{1}{i_0^{k-1}}\left(\|\rho_k\|_{L^1(\Omega_T)} + \|\rho_k^n\|_{L^1(\Omega_T)}\right), \end{aligned} \tag{5.22}$$

where we know that ρ_k lies in $L^1(\Omega_T)$, again by Fatou's Lemma. Therefore, $(\rho_1^n)_n$ converges to ρ_1 in $L^1(\Omega_T)$, but since $(\rho_1^n)_n$ is bounded in $L^p(\Omega_T)$, $p>2$, we get by interpolation that $(\rho_1^n)_n$ also converges to ρ_1 in $L^2(\Omega_T)$. Thus, we finally get that

$$\left(c_i\sum_{j=j_0+1}^{\infty} jc_j + c_i^n\sum_{j=j_0+1}^{n} jc_j^n\right) \quad \text{and} \quad \left(\sum_{j=j_0+1}^{\infty}(i+j)c_{i+j} + \sum_{j=j_0+1}^{n}(i+j)c_{i+j}^n\right)$$

can be made arbitrarily small in $L^1(\Omega_T)$ (uniformly in n) by taking j_0 large enough, which shows that $\left(Q_i^{-,n}(c^n)\right)_n$ and $\left(F_i^{+,n}(c^n)\right)_n$ converge in $L^1(\Omega_T)$, to $Q_i^-(c)$ and $F_i^+(c)$ respectively. This concludes the proof in the case (ii). □

Remark 5.2.3. *We point out that for case (ii), the fragmentation part of hypothesis* (5.14) *is not optimal. Indeed when proving the convergence of F_i^n to F_i, we only used information on the first moment (more precisely that ρ_1 lies in $L^1(\Omega_T)$ and that ρ_1^n converges to ρ_1 in $L^1(\Omega_T)$). Knowing that $(\rho_k^n)_n$ is bounded in $L^1(\Omega_T)$, for some $k>1$, we could show the convergence of F_i^n to F_i in $L^1(\Omega_T)$ with a weaker hypothesis than* (5.14). *Indeed estimating*

$$\left|F_i^{+,n}(c^n) - F_i^+(c)\right| \leq \sum_{j=1}^{j_0} B_{i+j}\beta_{i+j,i}\left|c_{i+j}^n - c_{i+j}\right| + \sup_{j>j_0}\frac{B_{i+j}\beta_{i+j,i}}{(i+j)^k}\left(\sum_{j=j_0+1}^{\infty}(i+j)^k c_{i+j} + \sum_{j=j_0+1}^{n}(i+j)^k c_{i+j}^n\right),$$

we see that is suffices to assume that

$$\lim_{j\to\infty}\frac{B_{i+j}\beta_{i+j,i}}{(i+j)^k} = 0, \quad \forall\, i\in\mathbb{N}^*.$$

Here it seems mandatory to assume that the limit is 0 (and not simply that the supremum over j is finite) because we do not know that $(\rho_k^n)_n$ converges to ρ_k but only that is is bounded in $L^1(\Omega_T)$. Also, we cannot so easily extend similarly the hypothesis on the coagulation coefficients because of the quadratic nature of Q_i.

5.3 Duality estimates and propagation of the mass in L^p norms

In this section, we present some duality lemmas and their applications to the proof of propagation of mass in L^p norm for the coagulation-fragmentation system (5.1)-(5.4). First we need to introduce the

Definition 5.3.1. *For $m > 0$ and $q \in]1, +\infty[$, we define $\mathcal{K}_{m,q} > 0$ as the best (i.e. the smallest) constant independent of $T > 0$ in the parabolic maximal regularity estimate*

$$\left(\int_{\Omega_T} |\partial_t v|^q + m^q \int_{\Omega_T} |\Delta_x v|^q \right)^{\frac{1}{q}} \leq \mathcal{K}_{m,q} \left(\int_{\Omega_T} |f|^q \right)^{\frac{1}{q}},$$

where v is the unique solution of the heat equation with constant diffusion coefficient m, homogeneous Neumann boundary conditions, and zero initial data:

$$\begin{cases} \partial_t v - m\Delta_x v = f & \text{on } \Omega_T, \\ \nabla_x v \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ v(0,\cdot) = 0 & \text{on } \Omega. \end{cases}$$

We recall that for all $m > 0$ and $q \in]1, +\infty[$, $\mathcal{K}_{m,q}$ is finite, and for the particular case $q = 2$ we have an explicit bound $\mathcal{K}_{m,2} \leq 1$ (see [7] and the references therein). We now recall some duality results, the first one being Proposition 1.1 of [11] (see also Proposition 2.4 of [7] for this exact formulation) and the second one being Proposition 2.5 of [7]. In the sequel, we shall systematically denote by p' the conjugate exponent of $p \in]1, +\infty[$, e.g. satisfying $\frac{1}{p} + \frac{1}{p'} = 1$.

Lemma 5.3.2. *Let Ω be a smooth bounded subset of $\mathbb{R}^N$ and consider a function $M : \Omega_T \to \mathbb{R}_+$ satisfying $a \leq M \leq b$ for some $a, b > 0$. For any $p \in]1, +\infty[$, if*

$$\frac{b-a}{b+a} \mathcal{K}_{\frac{a+b}{2}, p'} < 1,$$

then there exists a constant C (depending only on Ω, T, a, b, p) such that for any $u_0 \in L^p(\Omega)$, any weak solution u of the parabolic system

$$\begin{cases} \partial_t u - \Delta_x (Mu) = 0 & \text{on } \Omega_T, \\ \nabla_x u \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ u(0,\cdot) = u_0 & \text{on } \Omega, \end{cases}$$

satisfies $\|u\|_{L^p(\Omega_T)} \leq C \|u_0\|_{L^p(\Omega)}$.

Lemma 5.3.3. *Let Ω be a smooth bounded subset of $\mathbb{R}^N$, $\mu_1, \mu_2 \geq 0$ and consider a function $M : \Omega_T \to \mathbb{R}_+$ satisfying $a \leq M \leq b$ for some $a, b > 0$. For any $p \in]1, +\infty[$, if*

$$\frac{b-a}{b+a} \mathcal{K}_{\frac{a+b}{2}, p'} < 1,$$

then there exists a constant C (depending only on Ω, T, a, b, p) such that for any $u_0 \in L^p(\Omega)$, any function $u : \Omega_T \to \mathbb{R}_+$ satisfying (weakly)

$$\begin{cases} \partial_t u - \Delta_x (Mu) \leq \mu_1 u + \mu_2 & \text{on } \Omega_T, \\ \nabla_x u \cdot \nu = 0 & \text{on } [0,T] \times \partial\Omega, \\ u(0,\cdot) = u_0 & \text{on } \Omega, \end{cases} \tag{5.23}$$

belongs to $L^p(\Omega_T)$, and satisfies the estimate:

$$\|u\|_{L^p(\Omega_T)} < C\left(1 + \|u_0\|_{L^p(\Omega)}\right).$$

Let us now briefly explain how the above duality lemmas are used in the context of the coagulation-fragmentation system (5.1)-(5.4).

Proposition 5.3.4. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$ and $T > 0$. Assume* (5.4)*. Let $p \in]1, +\infty[$ and assume also that the initial data $c_i^{in} \geq 0$ are such that the initial mass ρ_1^{in} lies in $L^p(\Omega)$. Assume*

$$a := \inf_{i\in\mathbb{N}^*} d_i > 0 \quad \text{and} \quad b := \sup_{i\in\mathbb{N}^*} d_i < \infty, \tag{5.24}$$

and

$$\frac{b-a}{b+a}\mathcal{K}_{\frac{a+b}{2},p'} < 1. \tag{5.25}$$

Then there exists $C > 0$ such that, for all $n \in \mathbb{N}^*$, the mass ρ_1^n associated to the solution of the truncated system (5.17)-(5.19) *satisfies*

$$\|\rho_1^n\|_{L^p(\Omega_T)} \leq C\left\|\rho_1^{in}\right\|_{L^p(\Omega)}.$$

Proof. Looking at (5.20)-(5.21), we see that

$$\sum_{i=1}^{n} iQ_i^n = 0 = \sum_{i=1}^{n} iF_i^n,$$

and thus the mass ρ_1^n satisfies

$$\partial_t \rho_1^n - \Delta_x\left(M_1^n \rho_1^n\right) = 0,$$

where $M_1^n = \frac{\sum_{i=1}^n id_i c_i^n}{\sum_{i=1}^n ic_i^n}$. Thanks to assumptions (5.24)-(5.25), we can apply Lemma 5.3.2 to ρ_1^n, and the results follows since $\|\left(\rho_1^n\right)^{in}\|_{L^p(\Omega)} \leq \|\rho_1^{in}\|_{L^p(\Omega)}$ for all $n \in \mathbb{N}^*$. □

As already pointed out in [11], the closeness hypothesis (5.25) can be removed when p is close to 2, providing a variant of Proposition 5.3.4.

Proposition 5.3.5. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$ and $T > 0$. Assume* (5.4)*. Assume that the diffusion coefficients satisfy* (5.24) *and that we have nonnegative initial data $c_i^{in} \geq 0$.*

There exists $p_0 > 2$ such that, for all $p \in [2, p_0[$, if the initial mass ρ_1^{in} lies in $L^p(\Omega)$, then there exists $C > 0$ such that, for all $n \in \mathbb{N}^*$, the mass ρ_1^n associated to the solution of the truncated system (5.17)-(5.19) *satisfies*

$$\|\rho_1^n\|_{L^p(\Omega_T)} \leq C\left\|\rho_1^{in}\right\|_{L^p(\Omega)}.$$

Proof. As already pointed out after Definition 5.3.1, for $q = 2$ we have an explicit bound $\mathcal{K}_{m,2} \leq 1$. This yields that for any $a, b > 0$,

$$\frac{b-a}{b+a}\mathcal{K}_{\frac{a+b}{2},2} \leq \frac{b-a}{b+a} < 1.$$

Therefore, by continuity of $p \mapsto \mathcal{K}_{m,p'}$ (see [11, Lemma 3.2] for an explicit computation) there exists $p_0 > 2$ (depending on $a = \inf_{i\in\mathbb{N}^*} d_i$ and $b = \sup_{i\in\mathbb{N}^*} d_i$), such that for all $p \in [2, p_0[$

$$\frac{b-a}{b+a}\mathcal{K}_{\frac{a+b}{2},p'} < 1.$$

We then conclude by applying again Lemma 5.3.2 (or directly Proposition 5.3.4). □

Finally, if one assume (5.12), e.g. that the diffusion rates d_i are converging toward a positive limit, hypothesis (5.25) can be removed for any p, basically because the coefficients d_i will then be arbitrarily close from one another for i large enough (if fragmentation is non zero, we have to restrict ourselves to low dimension $N \leq 2$). This leads to yet another variant of Proposition 5.3.4, which was already stated and proven in [7] in the particular case of no fragmentation, that is when one assumes that $F_i = 0$ for all $i \in \mathbb{N}^*$.

Proposition 5.3.6. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$, $N \leq 2$, and $T > 0$. Assume* (5.4)*. Assume also that the coagulation coefficients satisfy*

$$a_{i,j} \leq Cij, \qquad \forall\ i,j \in \mathbb{N}^*, \tag{5.26}$$

that the fragmentation coefficients satisfy (5.14) *and that the diffusion coefficients satisfy* (5.12)*. Let $p \in]2, +\infty[$ and assume that the initial data $c_i^{in} \geq 0$ lie in $L^\infty(\Omega)$ for all $i \in \mathbb{N}^*$, and that the initial mass ρ_1^{in} lies in $L^p(\Omega)$.*

Then there exists $C > 0$ such that, for all $n \in \mathbb{N}^*$, the mass ρ_1^n associated to the solution of the truncated system (5.17)-(5.19) *satisfies*

$$\|\rho_1^n\|_{L^p(\Omega_T)} \leq C\left(1 + \left\|\rho_1^{in}\right\|_{L^p(\Omega)}\right).$$

We point out that the assumption (5.12) stating that (d_i) converges, which allows us to get rid of the closeness hypothesis (5.25), is not much more stringent from the physical point of view than just assuming that the diffusions coefficients are bounded below, because it is expected that the clusters diffuse less when they become larger, and thus the sequence (d_i) is expected to be decreasing. Also notice that while assumption (5.25) depends on p (and gets more and more stringent when p goes to infinity), assuming (5.12) allows to get the propagation of the mass in every L^p, $p < +\infty$. To prove Proposition 5.3.6, we first need a control on a finite number of concentrations c_i. This is the content of the following lemma.

Lemma 5.3.7. *Let Ω be a smooth bounded domain of $\mathbb{R}^N$, $N \leq 2$, and $T > 0$. Assume* (5.4)*. Assume that the fragmentation coefficients satisfy the asymptotic behavior* (5.14) *and that the diffusion coefficients satisfy* (5.24)*. Assume also that the initial concentrations $c_i^{in} \geq 0$ each lie in $L^\infty(\Omega)$ and that there exists $p > 2$ such that the initial mass ρ_1^{in} lies in $L^p(\Omega)$.*

Then, for all $i \in \mathbb{N}^*$, there exists $\tilde{c}_i \geq 0$ such that the solution c^n of (5.17)-(5.19) *satisfies*

$$\|c_i^n\|_{L^\infty(\Omega_T)} \leq \tilde{c}_i, \quad \forall\ n \in \mathbb{N}^*.$$

Proof. By Proposition 5.3.5, $(\rho_1^n)_n$ is bounded in $L^r(\Omega_T)$, for $r = \min(p, p_0) > 2$. From this we deduce that $\left(F_i^{+,n}(c^n)\right)_n$ is bounded in $L^r(\Omega_T)$. Indeed, remembering (5.14),

$$F_i^{+,n}(c^n) = \sum_{j=1}^{n-i} \frac{B_{i+j}\beta_{i+j,i}}{i+j}(i+j)c_{i+j}^n \leq K_i^F \rho_1^n.$$

Noticing that $F_i^{-,n}(c^n), Q_i^{-,n}(c^n) \leq 0$ for all $i \in \mathbb{N}^*$, we get

$$\partial_t c_i^n - d_i \Delta_x c_i^n \leq Q_i^{+,n}(c^n) + F_i^{+,n}(c^n).$$

For $i = 1$ we have that $Q_i^{+,n}(c^n) = 0$, and therefore

$$\partial_t c_1^n - d_1 \Delta_x c_1^n \leq F_1^{+,n}(c^n) \in L^r(\Omega_T).$$

Since $r > 2$, $N \leq 2$ and $c_1 \geq 0$, the regularizing properties of the heat equation yield that $(c_1^n)_n$ is bounded in $L^\infty(\Omega_T)$. Then for any integer $i \geq 2$, since

$$\begin{aligned}\left\|Q_i^{+,n}(c^n)\right\|_{L^r(\Omega_T)} &\leq (|\Omega|T)^{\frac{1}{r}} \left\|Q_i^{+,n}(c^n)\right\|_{L^\infty(\Omega_T)} \\ &\leq \frac{(|\Omega|T)^{\frac{1}{r}}}{2} \sum_{j=1}^{i-1} a_{i,j} \left\|c_{i-j}^n\right\|_{L^\infty(\Omega_T)} \left\|c_j^n\right\|_{L^\infty(\Omega_T)}\end{aligned}$$

involves only c_j^n for $j < i$, we can conclude by induction. □

We can now get L^p estimates on the mass ρ_1, assuming the convergence (5.12) of the diffusion coefficients.

Proof of Proposition 5.3.6. For any $I \in \mathbb{N}^*$, we define

$$a^I := \inf_{i \geq I} d_i \quad \text{and} \quad b^I := \sup_{i \geq I} d_i.$$

Since (d_i) converges toward a positive limit, there exists a positive integer I such that

$$\frac{b^I - a^I}{b^I + a^I} \mathcal{K}_{\frac{a^I + b^I}{2}, p'} < 1. \tag{5.27}$$

We fix such an I and then consider, for $n \geq I$,

$$\rho_1^{I,n} := \sum_{i=I}^n i c_i^n \quad \text{and} \quad M_1^{I,n} := \frac{\sum_{i=I}^n i d_i c_i^n}{\sum_{i=I}^n i c_i^n}.$$

Since (5.12) implies (5.24), Lemma 5.3.7 holds and it is enough to prove that $(\rho_1^{I,n})_n$ is bounded in $L^p(\Omega_T)$ to get that the full first moment $(\rho_1^n)_n$ is bounded in $L^p(\Omega_T)$.

Using (5.21) and the hypothesis $i = \sum_{j=1}^{i-1} j\beta_{i,j}$ of (5.4), we get that the contribution of fragmentation to the evolution of $\rho_1^{I,n}$ is non-positive:

$$\sum_{i=I}^n i F_i^n(c^n) = -\sum_{i=I}^n B_i c_i^n \left(i - \sum_{j=I}^{i-1} j\beta_{i,j} \right) \leq 0,$$

Therefore we get (using (5.20) and again the symmetry assumption in (5.4))

$$\begin{aligned}\partial_t \rho_1^{I,n} - \Delta_x \left(M_1^{I,n} \rho_1^{I,n} \right) &\leq \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n a_{i,j} c_i^n c_j^n \left((i+j) \mathbb{1}_{i+j \geq I} - i \mathbb{1}_{i \geq I} - j \mathbb{1}_{j \geq I} \right) \\ &\leq \sum_{i=1}^n \sum_{j=1}^n a_{i,j} c_i^n c_j^n i \left(\mathbb{1}_{i+j \geq I} - \mathbb{1}_{i \geq I} \right).\end{aligned}$$

Using (5.26), we end up with

$$\begin{aligned}\partial_t \rho_1^{I,n} - \Delta_x \left(M_1^{I,n} \rho_1^{I,n} \right) &\leq C \sum_{i=1}^{I-1} i^2 c_i^n \sum_{j=1}^n j c_j^n \\ &= \psi_1^n \rho_1^{I,n} + \psi_2^n,\end{aligned}$$

where

$$\psi_1^n = C \sum_{i=1}^{I-1} i^2 c_i^n \quad \text{and} \quad \psi_2^n = \psi_1^n \sum_{i=1}^{I-1} i c_i^n.$$

But $(\psi_1^n)_n$ and $(\psi_2^n)_n$ are bounded in $L^\infty(\Omega_T)$ by Lemma 5.3.7. Considering, for $k \in \{1, 2\}$, μ_k a bound of $\|\psi_k^n\|_{L^\infty(\Omega_T)}$, we get

$$\partial_t \rho_1^{I,n} - \Delta_x \left(M_1^{I,n} \rho_1^{I,n} \right) \leq \mu_1 \rho_1^{I,n} + \mu_2,$$

and we can conclude the proof of Proposition 5.3.6 thanks to Lemma 5.3.3. □

Remark 5.3.8. *Propositions 5.3.4, 5.3.5 and 5.3.6 provide the kind of estimates on the mass ρ_1^n that are needed to apply the existence result of Proposition 5.2.1 for the full system* (5.1)-(5.3)*. In particular, Theorem 5.1.3 is nothing more than the combination of Proposition 5.3.5 and Proposition 5.2.1 case (i).*

5.4 Superlinear moments and absence of gelation in the case of strong fragmentation

In this section we show how the result of Propositions 5.3.4 and 5.3.6, namely knowing that the mass ρ_1 lies in $L^p(\Omega_T)$, $p < +\infty$, allow us to prove Theorem 5.1.2, that is the creation and propagation of superlinear moments in presence of strong fragmentation, which prevents gelation. The outline of the proof is the following. First, we use the assumption (5.13) to get a bound for any superlinear moment $\rho_{l+\gamma-1}^n$ in terms of ρ_l^n and other lower order moments. Then we use interpolation estimates to bound all these lower order moments in terms of ρ_1^n and $\rho_{l+\gamma-1}^n$. But $(\rho_1^n)_n$ can be bounded by Proposition 5.3.6, and from this we deduce a bound on $\rho_{l+\gamma-1}^n$ that does not depend on n. We then bootstrap the estimate to get a bound for $\rho_{l+m(\gamma-1)}^n$ for all $m \in \mathbb{N}^*$. Finally, these *a priori* estimates enable us to apply Proposition 5.2.1, and to get a bound on the moments $\rho_{l+m(\gamma-1)}$ of the full system.

Proof of Theorem 5.1.2. For $n \in \mathbb{N}^*$, we consider c^n the solution of (5.17)-(5.19). For $l > 1$ we introduce

$$M_l^n = \frac{\sum_{i=1}^n i^l d_i c_i^n}{\sum_{i=1}^n i^l c_i^n}.$$

Using the weak formulation (5.20)-(5.21), assumption (5.13), the symmetry of (5.4) and the fact that (still thanks to (5.4)), $i^l - \sum_{j=1}^{i-1} j^l \beta_{i,j} \geq 0$, we estimate

$$\begin{aligned}
&\partial_t \rho_l^n - \Delta_x \left(M_l^n \rho_l^n\right) \\
&\quad = \frac{1}{2} \sum_{\substack{1 \leq i,j \leq n \\ i+j \leq n}} a_{i,j} c_i^n c_j^n \left((i+j)^l - i^l - j^l\right) - \sum_{i=2}^n B_i c_i^n \left(i^l - \sum_{j=1}^{i-1} j^l \beta_{i,j}\right) \\
&\quad \leq C_Q \sum_{i=1}^n \sum_{j=1}^n i^\alpha j^\beta \left((i+j)^l - i^l - j^l\right) c_i^n c_j^n - C_F \sum_{i=2}^n i^\gamma c_i^n \left(i^l - \sum_{j=1}^{i-1} j^l \beta_{i,j}\right). \qquad (5.28)
\end{aligned}$$

To bound from above the coagulation term, we use the existence of a constant $C_{Q,l} > 0$ such that,

$$(i+j)^l - i^l - j^l \leq C_{Q,l}(i^{l-1} j + i j^{l-1}), \quad \forall\, i, j \in \mathbb{N}^*,$$

see for instance [12]. To bound from below the fragmentation term, we estimate (using (5.4))

$$\begin{aligned}
i^l - \sum_{j=1}^{i-1} j^l \beta_{i,j} &= i^l \left(1 - \frac{1}{i} \sum_{j=1}^{i-1} \left(\frac{j}{i}\right)^{l-1} j \beta_{i,j}\right) \\
&\geq i^l \left(1 - \left(\frac{i-1}{i}\right)^{l-1} \frac{1}{i} \sum_{j=1}^{i-1} j \beta_{i,j}\right) \\
&= i^l \left(1 - \left(1 - \frac{1}{i}\right)^{l-1}\right) \\
&\geq \min(l-1, 1) i^{l-1}, \qquad (5.29)
\end{aligned}$$

the last inequality coming from the concavity (if $1 < l \leq 2$) or the convexity (if $l \geq 2$) of $x \mapsto (1-x)^{l-1}$. Introducing $C_{F,l} = \min(l-1, 1) > 0$ and going back to (5.28), we end up with

$$\partial_t \rho_l^n - \Delta_x \left(M_l^n \rho_l^n\right) \leq C_Q C_{Q,l} \left(\rho_{\alpha+l-1}^n \rho_{\beta+1}^n + \rho_{\alpha+1}^n \rho_{\beta+l-1}^n\right) - C_F C_{F,l} \left(\rho_{\gamma+l-1}^n - c_1^n\right).$$

Integrating on Ω and using the homogeneous Neumann boundary conditions, we get that

$$\begin{aligned} \frac{d}{dt} \int_\Omega \rho_l^n + C_F C_{F,l} \int_\Omega \rho_{\gamma+l-1}^n \\ \leq C_Q C_{Q,l} \int_\Omega \left(\rho_{\alpha+l-1}^n \rho_{\beta+1}^n + \rho_{\alpha+1}^n \rho_{\beta+l-1}^n\right) + C_F C_{F,l} \int_\Omega c_1^n. \end{aligned} \tag{5.30}$$

Assuming that $l > 2 - (\gamma - \alpha)$ and $l > 2 - (\gamma - \beta)$, i.e. $\alpha + 1 \leq \gamma + l - 1$ and $\beta + 1 \leq \gamma + l - 1$ (these are the assumptions on k in Theorem 5.1.2), Hölder's inequality yields the following interpolation estimates

$$\rho_{\alpha+1}^n \leq (\rho_1^n)^{\frac{\gamma+l-\alpha-2}{\gamma+l-2}} \left(\rho_{\gamma+l-1}^n\right)^{\frac{\alpha}{\gamma+l-2}}, \qquad \rho_{\beta+1}^n \leq (\rho_1^n)^{\frac{\gamma+l-\beta-2}{\gamma+l-2}} \left(\rho_{\gamma+l-1}^n\right)^{\frac{\beta}{\gamma+l-2}},$$

and

$$\rho_{\alpha+l-1}^n \leq (\rho_1^n)^{\frac{\gamma-\alpha}{\gamma+l-2}} \left(\rho_{\gamma+l-1}^n\right)^{\frac{\alpha+l-2}{\gamma+l-2}}, \qquad \rho_{\beta+l-1}^n \leq (\rho_1^n)^{\frac{\gamma-\beta}{\gamma+l-2}} \left(\rho_{\gamma+l-1}^n\right)^{\frac{\beta+l-2}{\gamma+l-2}},$$

Notice that in the case where $\alpha + l - 1 < 1$ or $\beta + l - 1 < 1$, the last two interpolation estimates are no longer valid, but we can then simply use $\rho_{\alpha+l-1}^n \leq \rho_1^n$ and $\rho_{\beta+l-1}^n \leq \rho_1^n$. The rest of the proof is then identical, up to different exponents for ρ_1^n and $\rho_{\gamma+l-1}^n$. The only property that we are going to use, which holds in every case, is that the obtained exponent for $\rho_{\gamma+l-1}^n$ is strictly less than 1.

In the case $\alpha + l - 1 \geq 1$ and $\beta + l - 1 \geq 1$, (5.30) then becomes

$$\begin{aligned} \frac{d}{dt} \int_\Omega \rho_l^n + C_F C_{F,l} \int_\Omega \rho_{\gamma+l-1}^n \\ \leq 2 C_Q C_{Q,l} \int_\Omega (\rho_1^n)^{\frac{2\gamma+l-(\alpha+\beta+2)}{\gamma+l-2}} \left(\rho_{\gamma+l-1}^n\right)^{\frac{\alpha+\beta+l-2}{\gamma+l-2}} + C_F C_{F,l} \int_\Omega c_1^n. \end{aligned}$$

Then, for any $0 \leq t \leq T$, if we integrate between t and T and use Hölder's inequality, we end up with

$$\begin{aligned} &\int_\Omega \rho_l^n(T) + C_F C_{F,l} \int_t^T \int_\Omega \rho_{\gamma+l-1}^n \\ &\leq \int_\Omega \rho_l^n(t) + 2 C_Q C_{Q,l} \int_t^T \int_\Omega (\rho_1^n)^{\frac{2\gamma+l-(\alpha+\beta+2)}{\gamma+l-2}} \left(\rho_{\gamma+l-1}^n\right)^{\frac{\alpha+\beta+l-2}{\gamma+l-2}} + C_F C_{F,l} \int_t^T \int_\Omega c_1^n \\ &\leq \int_\Omega \rho_l^n(t) + 2 C_Q C_{Q,l} \left\|\rho_1^n\right\|_{L^p(\Omega_T)}^{\frac{2\gamma+l-(\alpha+\beta+2)}{\gamma+l-2}} \left(\int_t^T \int_\Omega \rho_{\gamma+l-1}^n\right)^{\frac{\alpha+\beta+l-2}{\gamma+l-2}} \\ &\quad + C_F C_{F,l} \left\|\rho_1^n\right\|_{L^1(\Omega_T)}, \end{aligned}$$

where $p = \frac{2\gamma+l-(\alpha+\beta+2)}{\gamma-(\alpha+\beta)} = 1 + \frac{\gamma+l-2}{\gamma-(\alpha+\beta)}$. But thanks to Proposition 5.3.6, we have that for any $p < \infty$, $(\rho_1^n)_n$ is bounded in $L^p(\Omega_T)$. Renaming the constants, we have shown that, for all $l > 1$ also satisfying $l > 2 - (\gamma - \alpha)$ and $l > 2 - (\gamma - \beta)$, for all $n \in \mathbb{N}^*$ and all $0 \leq t \leq T$,

$$\int_\Omega \rho_l^n(T) + \tilde{C}_{1,l} \int_t^T \int_\Omega \rho_{l+\gamma-1}^n \leq \int_\Omega \rho_l^n(t) + \tilde{C}_{2,l} \left(\int_t^T \int_\Omega \rho_{l+\gamma-1}^n\right)^{1-\theta_l} + \tilde{C}_{3,l},$$

where $0 < \theta_l < 1$, $0 < \tilde{C}_{j,l} < \infty$, $1 \leq j \leq 3$, and crucially none of these constants depend on n. Using Young's inequality

$$a^{1-\theta_l} \leq (1-\theta_l)\epsilon^{\frac{1}{1-\theta_l}} a + \frac{\theta}{\epsilon^{\frac{1}{\theta}}},$$

with $a = \int_t^T \int_\Omega \rho^n_{l+\gamma-1}$ and any ϵ such that $0 < (1-\theta_l)\epsilon^{\frac{1}{1-\theta_l}} < \frac{\tilde{C}_{1,l}}{\tilde{C}_{2,l}}$, we end up with an estimate of the form

$$\int_t^T \int_\Omega \rho^n_{l+\gamma-1} \leq \hat{C}_{1,l} \int_\Omega \rho^n_l(t) + \hat{C}_{2,l}, \tag{5.31}$$

where, $0 < \hat{C}_{j,l} < \infty$, $1 \leq j \leq 2$, and again none of these constants depend on n.

We are now ready to prove by induction that, for all $m \in \mathbb{N}^*$ and all $T > 0$, there exists $\check{C}_{m,k} < \infty$ such that

$$\int_0^T t^{m-1} \int_\Omega \rho^n_{k+m(\gamma-1)}(t,x)dtdx \leq \check{C}_{m,k}, \quad \forall\ n \in \mathbb{N}^*, \tag{5.32}$$

k being defined in the assumptions of Theorem 5.1.2. We point out that, even if we chose a notation that only highlights the dependency of $C_{m,k}$ on m and k, it also depends on T and $|\Omega|$.

We start by proving (5.32) for $m = 1$. Using (5.31) with $l = k$ and $t = 0$ we get

$$\int_0^T \int_\Omega \rho^n_{l+\gamma-1} \leq \hat{C}_{1,k} \int_\Omega \rho^n_k(0) + \hat{C}_{2,k} \leq \hat{C}_{1,k} \int_\Omega \rho^{in}_k + \hat{C}_{2,k}$$

which immediately yields (5.32) for $m = 1$. We now take $m \in \mathbb{N}^*$, consider (5.31) with $l = k + m(\gamma-1)$, multiply it by t^{m-1} and integrate for t between 0 and T, which gives

$$\int_0^T t^m \int_\Omega \rho^n_{k+(m+1)(\gamma-1)} \leq m\hat{C}_{1,k+m(\gamma-1)} \int_0^T t^{m-1} \int_\Omega \rho^n_{k+m(\gamma-1)}(t) + \hat{C}_{2,k+m(\gamma-1)} T^m.$$

Therefore, assuming (5.32) holds for m, we get that (5.32) also holds for $m+1$, and then for all $m \in \mathbb{N}^*$ by induction.

Then, notice that (5.32) with $m = 1$ says exactly that the superlinear moment $\left(\rho^n_{k+\gamma-1}\right)_n$ is bounded in $L^1(\Omega_T)$. Therefore we can apply Proposition 5.2.1 case (ii) to extract from $(c^n)_n$ a weak solution c of (5.1)-(5.4). Besides, as we saw in the proof of Proposition 5.2.1 case (ii), the L^1 bound of the superlinear moment $\left(\rho^n_{k+\gamma-1}\right)_n$ allows us to prove that $(\rho^n_1)_n$ converges to ρ_1 in $L^1(\Omega_T)$. But for all $t \geq 0$ and all $n \in \mathbb{N}^*$ we have (rigorously)

$$\int_\Omega \rho^n_1(t) = \int_\Omega \left(\rho^{in}_1\right)^n,$$

and since

$$\int_\Omega \left(\rho^{in}_1\right)^n \underset{n\to\infty}{\longrightarrow} \int_\Omega \rho^{in}_1,$$

we get that, for almost all $t \geq 0$,

$$\int_\Omega \rho_1(t) = \int_\Omega \rho^{in}_1,$$

i.e. there is no gelation.

Finally, by Fatou's Lemma the bound (5.32) is carried over to the moments of the limiting solution c, i.e. for all $m \in \mathbb{N}^*$

$$\int_0^T \frac{t^{m-1}}{(m-1)!} \int_\Omega \rho_{k+m(\gamma-1)}(t,x)dtdx < \infty. \tag{5.33}$$

□

Remark 5.4.1. *The assumption* (5.12) *on the convergence of the diffusion coefficients, the assumption* $c_i^{in} \in L^\infty(\Omega)$ *for all* $i \geq 1$*, and the assumption* $N \leq 2$ *were only used here to apply Proposition 5.3.6 and get a bound on* $(\rho_1^n)_n$ *in* $L^p(\Omega_T)$*, for all* $p < \infty$*. Therefore, if we fix a* $p > 2$ *such that* $\rho_1^{in} \in L^p(\Omega)$*, consider an arbitrary dimension* N*, assume* (5.25) *instead of* (5.12) *and remove the assumption* $c_i^{in} \in L^\infty(\Omega)$ *for all* $i \geq 1$*, we can now use Proposition 5.3.4 to get a bound on* $(\rho_1^n)_n$ *in* $L^p(\Omega_T)$ *for this fixed* p*. The estimate* (5.31) *then holds for all* $l > 1$ *such that* $1 + \frac{\gamma+l-2}{\gamma-(\alpha+\beta)} \leq p$*, and so do* (5.32) *and* (5.33)*, for all* m *such that* $1 + \frac{\gamma+k+(m-1)(\gamma-1)-2}{\gamma-(\alpha+\beta)} \leq p$*.*

Remark 5.4.2. *As already pointed out, in Theorem 5.1.2 we made no assumption on the coefficients* $\beta_{i,j}$ *(aside from the microscopic mass conservation* (5.4)*). A fairly generic assumption one can add is the following:*

$$\forall\ l > 1,\ \exists C_l > 0,\ \forall\ i \in \mathbb{N}^*, \quad i^l - \sum_{j=1}^{i-1} j^l \beta_{i,j} \geq C_l i^l. \tag{5.34}$$

A similar assumption is made (in the particular case of binary fragmentation) in [14] for the homogeneous case.

While (5.34) *may not hold for some coagulation-fragmentation models (it is for instance not satisfied for the Becker-Döring [4] model), it does still hold for a broad range of models. Indeed, typical examples of fragmentation coefficients are*

$$B_i = i^\gamma,\ \gamma \in \mathbb{R} \quad \textit{and} \quad \beta_{i,j} = \frac{i}{\sum_{k=1}^{i-1} k^{1+\nu}} j^\nu, \nu > -2,$$

see [28] and the references therein. For such coefficients, one can check that (5.34) *is always satisfied.*

Assuming (5.34)*, we can then use it instead of* (5.29) *in the proof of Theorem 5.1.2, and gain one power of* i*. The moment of higher order that appears from the fragmentation term is then* $\rho_{l+\gamma}$ *(in place of* $\rho_{l+\gamma-1}$*), and the rest of the proof still holds if we only assume* $\alpha+\beta < \gamma+1$ *and* $\gamma \geq 0$ *(instead of* $\alpha + \beta < \gamma$ *and* $\gamma \geq 1$*).*

Finally we point out that, if we wanted to generalize Theorem 5.1.2 to continuous versions of the coagulation-fragmentation equations with diffusion, an assumption like (5.34) *would probably be needed, since* (5.29) *does not readily extend to the continuous setting.*

5.5 Propagation of smoothness

This section is devoted to the proof of Theorem 5.1.5. The main argument is to show that a control of all moments ρ_k, $k \geq 0$, in all spaces $L^p(\Omega_T)$, $p < \infty$, allows for propagation of smoothness (Proposition 5.5.2). We begin with a lemma to control the coagulation and fragmentation terms Q_i and F_i.

Lemma 5.5.1. *Let* Ω *be a smooth bounded domain of* $\mathbb{R}^N$ *and* $s \in \mathbb{N}$*. Assume that* $(c_i)_{i\in\mathbb{N}^*}$ *is a sequence of positive functions defined on* Ω_T *such that*

$$\sup_{i\in\mathbb{N}^*} \left\| i^k c_i \right\|_{W^{s,p}(\Omega_T)} < +\infty, \quad \forall k \geq 0,\ \forall p < +\infty. \tag{5.35}$$

Assume also (5.2)-(5.4). *For the coagulation and fragmentation coefficients, assume that there exists* $C \geq 0$ *and* $\gamma_{max} \in \mathbb{R}$ *such that, for all* $i, j \in \mathbb{N}^*$

$$a_{i,j} \leq Cij \quad \text{and} \quad B_i \leq Ci^{\gamma_{\max}}. \tag{5.36}$$

Then, the following estimates hold

$$\sup_{i\in\mathbb{N}^*} \left\| i^k Q_i(c) \right\|_{W^{s,p}(\Omega_T)} < +\infty, \quad \forall k \geq 0,\ \forall p < +\infty. \tag{5.37}$$

and

$$\sup_{i\in\mathbb{N}^*} \left\| i^k F_i(c) \right\|_{W^{s,p}(\Omega_T)} < +\infty, \quad \forall k \geq 0,\ \forall p < +\infty. \tag{5.38}$$

Proof. The bound (5.37) for the coagulation term was already proven in [7]. To get the bound (5.38) for the fragmentation term, we estimate (remembering (5.3))

$$\left\| i^k F_i(c) \right\|_{W^{s,p}(\Omega_T)} \leq \sum_{j=1}^{\infty} B_{i+j}\beta_{i+j,i} i^k \left\| c_{i+j} \right\|_{W^{s,p}(\Omega_T)} + B_i i^k \left\| c_i \right\|_{W^{s,p}(\Omega_T)}.$$

But by (5.4), $\beta_{i+j,i} \leq \frac{i+j}{i}$, and using (5.36) we end up with

$$\begin{aligned}
\left\| i^k F_i(c) \right\|_{W^{s,p}(\Omega_T)} &\leq C \sum_{j=1}^{\infty} (i+j)^{\gamma_{max}+1} i^{k-1} \left\| c_{i+j} \right\|_{W^{s,p}(\Omega_T)} \\
&\quad + C i^{\gamma_{max}+k} \left\| c_i \right\|_{W^{s,p}(\Omega_T)} \\
&\leq C \sum_{j=0}^{\infty} \left\| (i+j)^{\gamma_{max}+k} c_{i+j} \right\|_{W^{s,p}(\Omega_T)} \\
&\leq C \sup_{i\in\mathbb{N}^*} \left\| i^{\gamma_{max}+k+2} c_i \right\|_{W^{s,p}(\Omega_T)} \sum_{j=0}^{\infty} \frac{1}{(i+j)^2} \\
&\leq C \sup_{i\in\mathbb{N}^*} \left\| i^{\gamma_{max}+k+2} c_i \right\|_{W^{s,p}(\Omega_T)} \sum_{j=0}^{\infty} \frac{1}{(1+j)^2}
\end{aligned}$$

and this bound is finite by (5.35). □

As already stated in [7], Lemma 5.5.1 can then be used to get $W^{s,p}$ estimates for solutions of (5.1)-(5.4).

Proposition 5.5.2. *Let* Ω *be a smooth bounded domain of* $\mathbb{R}^N$, *and* $T > 0$. *Assume* (5.36) *for the coagulation and fragmentation coefficients and* (5.16) *for the diffusion coefficients. Let* c *be a solution of* (5.1)-(5.4) *such that* ρ_k *lies in* $L^p(\Omega_T)$, *for all* $k \geq 0$ *and all* $p < \infty$. *Then* ρ_k *lies in the Sobolev space* $W^{s,p}(\Omega_T)$, *for all* $k \geq 0$, *all* $s \in \mathbb{N}$ *and all* $p < \infty$.

Proof. We are going to show, by induction on s, that for all $s \in \mathbb{N}$,

$$\sup_{i\geq 1} \left\| i^k c_i \right\|_{W^{s,p}(\Omega_T)} < +\infty, \quad \forall k \in \mathbb{N},\ \forall p < +\infty. \tag{5.39}$$

For all $i \in \mathbb{N}^*$, we have $\left\| i^k c_i \right\|_{L^p(\Omega_T)} \leq \left\| \rho_k \right\|_{L^p(\Omega_T)}$, therefore (5.39) holds for $s = 0$. Assuming (5.39) for a given $s \in \mathbb{N}$, Lemma 5.5.1 shows that

$$\sup_{i\geq 1} \left\| i^k \left(Q_i(c) + F_i(c) \right) \right\|_{W^{s,p}(\Omega_T)} < +\infty, \quad \forall k \in \mathbb{N},\ \forall p < +\infty.$$

Since

$$(\partial_t - d_i\Delta_x)\, i^k c_i = i^k \left(Q_i(c) + F_i(c)\right),$$

the regularizing properties of the heat equation yield that

$$\sup_{i\geq 1} \left\| i^k c_i \right\|_{W^{s+1,p}(\Omega_T)} < +\infty, \quad \forall k \in \mathbb{N},\ \forall p < +\infty,$$

where we also used (5.16), i.e. that the d_i are bounded above and below, which ensure that the regularity estimates are uniform w.r.t. i. Therefore (5.39) holds for all $s \in \mathbb{N}$. Notice that we also get $W^{s,p}$ estimates for moments ρ_k of any order, because

$$\|\rho_k\|_{W^{s,p}(\Omega_T)} \leq \sum_{i=1}^{\infty} \frac{1}{i^2} \left\| i^{k+2} c_i \right\|_{W^{s,p}(\Omega_T)} \leq \sup_{i\geq 1} \left\| i^{k+2} c_i \right\|_{W^{s,p}(\Omega_T)} \sum_{i=1}^{\infty} \frac{1}{i^2}.$$

□

Finally, we can give the

Proof of Theorem 5.1.5. We want to apply Proposition 5.5.2. Notice that (5.36) is implied by the assumptions of Theorem 5.1.5.

In the sublinear coagulation case, the fact that ρ_k lies in $L^p(\Omega_T)$ for all $k \geq 0$ and all $p < \infty$ was already proven in [7, Theorem 1.9], in the case of pure coagulation ($F_i = 0$ for all $i \in \mathbb{N}^*$). The proof is a more technical version of the one of Proposition 5.3.6 that can be readily extended to cases including fragmentation. One only needs to assume $N \leq 2$ (to get Lemma 5.3.7, which is valid in any dimension only when there is no fragmentation). Then, since the contribution of the fragmentation to the evolution of truncated moment $\rho_k^{I,n}$ of any order is non positive (thanks to (5.4)):

$$\sum_{i=I}^{\infty} i^k F_i = -\sum_{i=I}^{\infty} B_i c_i \left(i^k - \sum_{j=I}^{i-1} j^k \beta_{i,j} \right) \leq 0,$$

the proof from [7] still holds without further modifications, even when the fragmentation is nonzero, and we get that ρ_k lies in $L^p(\Omega_T)$ for all $k \geq 0$ and all $p < \infty$, in the sublinear coagulation case.

In the strong fragmentation case, since the assumptions implies that ρ_k^{in} lies in $L^1(\Omega)$ for all $k \geq 0$, Theorem 5.1.2 yields that ρ_k lies in $L^1(\Omega_T)$ for all $k \geq 0$. But by Proposition 5.3.6, we also have that ρ_1 lies in $L^p(\Omega_T)$ for all $p < \infty$. By interpolation we then get that ρ_k lies in $L^p(\Omega_T)$ for all $k \geq 0$ and all $p < \infty$.

Therefore, we can use Proposition 5.5.2 in both the sublinear coagulation and the strong fragmentation cases. The $\mathcal{C}^\infty$ regularity announced in Theorem 5.1.5 is then a straightforward consequence of Sobolev embeddings.

We finish with the proof of the uniqueness statement, which is an extension from a result in [23] where only the case without fragmentation is treated. We consider $c = (c_i)_{i\in\mathbb{N}^*}$ and $\tilde{c} = (\tilde{c}_i)_{i\in\mathbb{N}^*}$ two smooth solutions to the coagulation-fragmentation system (5.1)-(5.4) and compute

$$\begin{aligned}
\frac{d}{dt} \int_\Omega \sum_{i=1}^{\infty} i|c_i - \tilde{c}_i| &= \int_\Omega \sum_{i=1}^{\infty} \varphi_i \left(Q_i(c) - Q_i(\tilde{c}) + F_i(c) - F_i(\tilde{c})\right) \\
&= \frac{1}{2} \int_\Omega \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{i,j}(c_i c_j - \tilde{c}_i \tilde{c}_j)(\varphi_{i+j} - \varphi_i - \varphi_j) \\
&\quad - \int_\Omega \sum_{i=2}^{\infty} B_i (c_i - \tilde{c}_i) \left(\varphi_i - \sum_{j=1}^{i-1} \beta_{i,j} \varphi_j \right),
\end{aligned}$$

where

$$\varphi_i = i\,\mathrm{sgn}(c_i - \tilde{c}_i).$$

We first bound the first term. Rewriting $(c_i c_j - \tilde{c}_i\tilde{c}_j)$ as $(c_i - \tilde{c}_i)c_j + (c_j - \tilde{c}_j)\tilde{c}_i$, estimating

$$\begin{aligned}(c_i - \tilde{c}_i)c_j(\varphi_{i+j} - \varphi_i - \varphi_j) &\le (i+j)|c_i - \tilde{c}_i|c_j - i|c_i - \tilde{c}_i|c_j + j|c_i - \tilde{c}_i|c_j \\ &= 2j|c_i - \tilde{c}_i|c_j,\end{aligned}$$

and using the symmetry of the coagulation coefficients, we end up with

$$\frac{1}{2}\int_\Omega \sum_{i=1}^\infty\sum_{j=1}^\infty a_{i,j}(c_i c_j - \tilde{c}_i\tilde{c}_j)(\varphi_{i+j} - \varphi_i - \varphi_j) \le \int_\Omega \sum_{i=1}^\infty\sum_{j=1}^\infty a_{i,j}|c_i - \tilde{c}_i|j(c_j + \tilde{c}_j).$$

Since $a_{i,j} \le Cij$, we obtain

$$\begin{aligned}&\frac{1}{2}\int_\Omega \sum_{i=1}^\infty\sum_{j=1}^\infty a_{i,j}(c_i c_j - \tilde{c}_i\tilde{c}_j)(\varphi_{i+j} - \varphi_i - \varphi_j) \\ &\le C\int_\Omega \sum_{i=1}^\infty i|c_i - \tilde{c}_i|\sum_{j=1}^\infty j^2(c_j + \tilde{c}_j) \\ &\le C\left(\|\rho_2(c)\|_{L^\infty(\Omega_T)} + \|\rho_2(\tilde{c})\|_{L^\infty(\Omega_T)}\right)\int_\Omega \sum_{i=1}^\infty i|c_i - \tilde{c}_i|.\end{aligned}$$

For the second term, we have (using (5.4))

$$\begin{aligned}-\int_\Omega \sum_{i=2}^\infty B_i(c_i - \tilde{c}_i)\left(\varphi_i - \sum_{j=1}^{i-1}\beta_{i,j}\varphi_j\right) &\le \int_\Omega \sum_{i=2}^\infty B_i|c_i - \tilde{c}_i|\left(i + \sum_{j=1}^{i-1}\beta_{i,j}j\right) \\ &= 2\int_\Omega \sum_{i=2}^\infty B_i i|c_i - \tilde{c}_i| \\ &\le 2\sup_{i\in\mathbb{N}^*} B_i \int_\Omega \sum_{i=2}^\infty i|c_i - \tilde{c}_i|.\end{aligned}$$

Putting everything back together, we get

$$\begin{aligned}&\frac{d}{dt}\int_\Omega \sum_{i=1}^\infty i|c_i - \tilde{c}_i| \\ &\qquad \le \left(C\left(\|\rho_2(c)\|_{L^\infty(\Omega_T)} + \|\rho_2(\tilde{c})\|_{L^\infty(\Omega_T)}\right) + 2\sup_{i\in\mathbb{N}^*} B_i\right)\int_\Omega \sum_{i=1}^\infty i|c_i - \tilde{c}_i|,\end{aligned}$$

and by Gronwall's Lemma we conclude that, if the solutions c and $\tilde{c}$ have the same initial data, then they remain equal for all positive time. □

Part III

Validated numerics: theory and applications

Chapter 6

Extension of the framework of validated numerics techniques to include operators with a tridiagonal dominant linear part

Abstract

This chapter is taken from [73]. We present a method designed for computing solutions of infinite dimensional nonlinear operators $f(x) = 0$ with a tridiagonal dominant linear part. We recast the operator equation into an equivalent Newton-like equation $x = T(x) = x - Af(x)$, where A is an approximate inverse of the derivative $Df(\bar{x})$ at an approximate solution $\bar{x}$. We present rigorous computer-assisted calculations showing that T is a contraction near $\bar{x}$, thus yielding the existence of a solution. Since $Df(\bar{x})$ does not have an asymptotically diagonal dominant structure, the computation of A is not straightforward. This chapter provides ideas for computing A, and proposes a new rigorous method for proving existence of solutions of nonlinear operators with tridiagonal dominant linear part.

6.1 Introduction

Tridiagonal operators naturally arise in the theory of orthogonal polynomials, ordinary differential equations (ODEs), continued fractions, numerical analysis of partial differential equations (PDEs), integrable systems, quantum mechanics and solid state physics. Some differential operators can be represented by infinite tridiagonal matrices acting in sequence spaces, as it is the case for instance for differentiation in frequency space of the Hermite functions. Other examples come from the study of ODEs like the Mathieu equation, the spheroidal wave equation, the Whittaker-Hill equation and the Lamé equation.

While many well-developed methods and efficient algorithms already exist in the literature for solving linear tridiagonal matrix equations and computing their inverses, our own method has a different flavour. We aim at developing a computational method in order to prove, in a mathematically rigorous and constructive sense, existence of solutions to infinite dimensional nonlinear equations of the form

$$f(x) = \mathcal{L}(x) + N(x) = 0, \tag{6.1}$$

where $\mathcal{L}$ is a tridiagonal linear operator and N is a nonlinear operator. The domain of the operator f is the space of algebraically decaying sequences

$$\Omega^s := \left\{ x = (x_k)_{k\geq 0} : \|x\|_s := \sup_{k\geq 0}\{|x_k|\omega_k^s\} < \infty \right\}, \tag{6.2}$$

where

$$\omega_k^s := \begin{cases} 1, & k = 0, \\ k^s, & k \geq 1. \end{cases}$$

The assumptions on the linear and nonlinear parts of (6.1) are that $\mathcal{L} : \Omega^s \to \Omega^{s-s_L}$ and $N : \Omega^s \to \Omega^{s-s_N}$, for some $s_L > s_N$. Intuitively, this means that the linear part *dominates* the nonlinear part. Since $\Omega^{s_1} \subset \Omega^{s_2}$ for $s_1 > s_2$, one can see that f maps Ω^s into Ω^{s-s_L}.

General nonlinear operator equations of the form $f(x) = 0$ defined on the Banach space Ω^s arise in the study of bounded solutions of finite and infinite dimensional dynamical systems. For instance, $x = (x_k)_{k\geq 0}$ may be the infinite sequence of Fourier coefficients of a periodic solution of an ODE, a periodic solution of a delay differential equation (DDE) or an equilibrium solution of a PDE with Dirichlet, periodic or Neumann boundary conditions. The unknown x may also be the infinite sequence of Chebyshev coefficients of a solution of a boundary value problem (BVP), the Hermite coefficients of a solution of an ODE defined on an unbounded domain, or the Taylor coefficients of the solution of a Cauchy problem. In the case when the differential equation is smooth, the decay rate of the coefficients of x will be algebraic or even exponential [70]. In the present paper, we chose to solve (6.1) in the weighed ℓ^∞ Banach space Ω^s which corresponds to C^k solutions. In order to exploit the analyticity of the solutions, we could follow the idea of [136] and solve (6.1) in weighed ℓ^1 Banach spaces. This choice of space is not considered in the present paper.

Recently, several attempts to solve $f(x) = 0$ in Ω^s have been successful. They belong to a field now called *rigorous numerics*. This field aims at constructing algorithms that provide approximate solutions to a given problem, together with precise bounds implying the existence of an exact solution in the mathematically rigorous sense. Equilibria of PDEs [118, 133, 220], periodic solutions of DDEs [143], fixed points of infinite dimensional maps [107] and periodic solutions of ODEs [57, 93] have been computed using such methods.

One popular idea in rigorous numerics is to recast the problem $f(x) = 0$ as a problem of fixed point of a Newton-like equation of the form $T(x) = x - Af(x)$, where A is an approximate inverse of $Df(\bar{x})$, and $\bar{x}$ is a numerical approximation obtained by computing a finite dimensional projection of f. In [93, 107, 118, 133, 143, 220], the nonlinear equations under study have asymptotically diagonal or block-diagonal dominant linear part, which helps a lot in the computation of approximate inverses. In contrast, the present work considers problems with tridiagonal dominant linear part. To the best of our knowledge, this is the first attempt to compute rigorously solutions of such problems. While our proposed approach is designed for a specific class of operators (see assumptions (6.4) and (6.5)), we believe that it can be seen as a first step toward rigorously solving more complicated nonlinear operators with tridiagonal dominant linear part.

The paper is organized as follows. In Section 6.2, we present a method enabling to compute (with the help of the computer) pseudo-inverses of tridiagonal operators of a certain class. In Section 6.3, we recast the problem $f(x) = 0$ as a fixed point problem $T(x) = x - Af(x)$, where A is a pseudo-inverse, and we present the rigorous computational method to prove existence of fixed points of T. In Section 6.4, we present an application and finally, in Section 6.5, we conclude by presenting some interesting future directions.

6.2 Computing pseudo-inverses of tridiagonal operators

This Section is devoted to the construction of a pseudo-inverse of a linear operator with tridiagonal tail (see (6.6)). We begin this Section by specifying the assumptions that we make on the growth of the tridiagonal terms. Then we use an LU-decomposition to formally obtain a formula for the pseudo-inverse. Finally, we check that the (formally defined) pseudo-inverse has good mapping properties (see Proposition 6.2.3).

Given three sequences $(\lambda_k)_{k\geq 0}$, $(\mu_k)_{k\geq 0}$, $(\beta_k)_{k\geq 0}$ and $x \in \Omega^s$, we define the tridiagonal linear operator (acting on x) $\mathcal{L}(x) = (\mathcal{L}_k(x))_{k\geq 0}$ of (6.1) by

$$\mathcal{L}_k(x) = \lambda_k x_{k-1} + \mu_k x_k + \beta_k x_{k+1}, \quad k \geq 1, \tag{6.3}$$

and $\mathcal{L}_0(x) = \mu_0 x_0 + \beta_0 x_1$. Assume that there exist real numbers $s_L > 0$, $0 < C_1 \leq C_2$ and an integer k_0 such that

$$\forall\ k \geq 0, \quad \left|\frac{\lambda_k}{\omega_k^{s_L}}\right|, \left|\frac{\mu_k}{\omega_k^{s_L}}\right|, \left|\frac{\beta_k}{\omega_k^{s_L}}\right| \leq C_2 \quad \text{and} \quad \forall\ k \geq k_0, \quad C_1 \leq \left|\frac{\mu_k}{\omega_k^{s_L}}\right|. \tag{6.4}$$

Assume further the existence of $\delta \in \left(0, \frac{1}{2}\right)$ and $k_0 \geq 0$ such that

$$\forall\ k \geq k_0, \quad \left|\frac{\lambda_k}{\mu_k}\right|, \left|\frac{\beta_k}{\mu_k}\right| \leq \delta. \tag{6.5}$$

Then, under assumptions (6.4) and (6.5), $\mathcal{L}$ defined by (6.3) is a tridiagonal operator which maps Ω^s into Ω^{s-s_L}. Indeed, if $x \in \Omega^s$, then

$$\begin{aligned} \|\mathcal{L}(x)\|_{s-s_L} &= \sup_{k\geq 0}\{|\mathcal{L}_k(x)|\omega_k^{s-s_L}\} \\ &\leq C_2\left(\sup_{k\geq 1}\{|x_{k-1}|\omega_k^s\} + \sup_{k\geq 0}\{|x_k|\omega_k^s\} + \sup_{k\geq 0}\{|x_{k+1}|\omega_k^s\}\right) < \infty. \end{aligned}$$

From now on, assume for the sake of simplicity that $s_N = 0$, that is the nonlinear part N of (6.1) maps Ω^s into Ω^s. Since Ω^s is an algebra under discrete convolutions when $s > 1$ (e.g. see [79, 118]), then any N which is a combination of such convolutions maps Ω^s into Ω^s. Assume that using a finite dimensional projection $f^{(m)} : \mathbb{R}^m \to \mathbb{R}^m$ of (6.1), we computed a numerical approximation $\bar{x}$ such that $f^{(m)}(\bar{x}) \approx 0$. We identify $\bar{x} \in \mathbb{R}^m$ and $\bar{x} = (\bar{x}, 0, 0, 0, 0, \dots) \in \Omega^s$. We then try to construct a ball

$$B_{\bar{x}}(r) = \bar{x} + B_0(r) = \bar{x} + \{x \in \Omega^s : \|x\|_s \leq r\} = \{x \in \Omega^s : \|x - \bar{x}\|_s \leq r\}$$

centered at $\bar{x}$ and containing a unique solution of (6.1), by showing that a specific Newton-like operator $T(x) = x - Af(x)$ is a contraction on $B_{\bar{x}}(r)$. This requires the construction of an approximate inverse A of $Df(\bar{x}) = \mathcal{L}(\bar{x}) + DN(\bar{x})$. In order to do so, the structures of $\mathcal{L}(\bar{x})$ and $DN(\bar{x})$ need to be understood. From (6.3) and (6.4), $\mathcal{L}(\bar{x})$ is a tridiagonal operator with entries growing to infinity at the rate k^{s_L}. Moreover, since $DN(\bar{x})$ maps Ω^s into Ω^s, it is a bounded linear operator. As mentioned above, the expectation is that the coefficients of $\bar{x}$ decay fast to zero. This implies that a reasonable approximation $A^\dagger$ of $Df(\bar{x})$ is given by

$$A^\dagger := \begin{pmatrix} D & & & 0 & \\ & & \beta_{m-1} & & \\ & \lambda_m & \mu_m & \beta_m & \\ 0 & & \lambda_{m+1} & \mu_{m+1} & \beta_{m+1} \end{pmatrix}, \tag{6.6}$$

with $D := Df^{(m)}(\bar{x})$ for m large enough. We wish to find the inverse of $A^\dagger$ in terms of D, $(\beta_k)_{k\geq m-1}$, $(\mu_k)_{k\geq m}$ and $(\lambda_k)_{k\geq m}$. We assume therefore that

$$A^\dagger x = y, \tag{6.7}$$

where x and y are the infinite vectors

$$x = \begin{pmatrix} x_0 \\ x_1 \\ . \\ . \end{pmatrix}, \qquad y = \begin{pmatrix} y_0 \\ y_1 \\ . \\ . \end{pmatrix}.$$

The infinite part of (6.7) writes

$$\begin{pmatrix} \mu_m & \beta_m & 0 & 0 & ... \\ \lambda_{m+1} & \mu_{m+1} & \beta_{m+1} & 0 & ... \\ 0 & \lambda_{m+2} & \mu_{m+2} & \beta_{m+2} & ... \\ . & . & . & . & ... \end{pmatrix} \begin{pmatrix} x_m \\ x_{m+1} \\ . \\ . \end{pmatrix} = \begin{pmatrix} y_m - \lambda_m\, x_{m-1} \\ y_{m+1} \\ . \\ . \end{pmatrix}. \tag{6.8}$$

We introduce the notations of the book of P.G. Ciarlet (see Theorem 4.3-2 on page 142 in [102]):

$$a_2 = \lambda_{m+1},\ a_3 = \lambda_{m+2}, ..., \quad b_1 = \mu_m,\ b_2 = \mu_{m+1}, ..., \quad c_1 = \beta_m,\ c_2 = \beta_{m+1}, ...,$$

and $(\delta_n)_{n\in\mathbb{N}}$ defined by the induction formula

$$\delta_0 = 1, \quad \delta_1 = b_1, \quad \text{and} \quad \delta_n = b_n\,\delta_{n-1} - a_n\,c_{n-1}\,\delta_{n-2}, \quad \text{for} \quad n \geq 2.$$

Note that only the δ_n are really useful.

Let us define the tridiagonal operator $\mathcal{T}$ by

$$\mathcal{T} := \begin{pmatrix} b_1 & c_1 & 0 & 0 & ... \\ a_2 & b_2 & c_2 & 0 & ... \\ 0 & a_3 & b_3 & c_3 & ... \\ . & . & . & . & ... \end{pmatrix}. \tag{6.9}$$

For any infinite vector $x = (x_0, \ldots, x_k, \ldots)^T$, we introduce the notation

$$x_F := (x_0, \ldots, x_{m-1})^T \quad \text{and} \quad x_I := (x_m, \ldots, x_{m+k}, \ldots)^T.$$

Using the notation $\mathbf{e_1} = (1, 0, 0, 0, 0, \cdots)^T$, the system (6.8) becomes

$$\mathcal{T} x_I = y_I - \lambda_m\, x_{m-1}\mathbf{e_1}.$$

From Theorem 4.3-2 in [102], we compute an LU-decomposition of the tridiagonal operator defined in (6.9) as $\mathcal{T} = L_I U_I$, where

$$L_I := \begin{pmatrix} 1 & 0 & 0 & ... \\ a_2\,\frac{\delta_0}{\delta_1} & 1 & 0 & ... \\ 0 & a_3\,\frac{\delta_1}{\delta_2} & 1 & ... \\ . & . & . & ... \end{pmatrix} \quad \text{and} \quad U_I := \begin{pmatrix} \frac{\delta_1}{\delta_0} & c_1 & 0 & ... \\ 0 & \frac{\delta_2}{\delta_1} & c_2 & ... \\ 0 & 0 & \frac{\delta_3}{\delta_2} & ... \\ . & . & . & ... \end{pmatrix}. \tag{6.10}$$

Hence, the system (6.8) becomes $L_I z_I = y_I - \lambda_m\, x_{m-1}\mathbf{e_1}$ combined with $U_I x_I = z_I$, that is

$$\begin{pmatrix} 1 & 0 & 0 & ... \\ a_2\,\frac{\delta_0}{\delta_1} & 1 & 0 & ... \\ 0 & a_3\,\frac{\delta_1}{\delta_2} & 1 & ... \\ . & . & . & ... \end{pmatrix} \begin{pmatrix} z_m \\ z_{m+1} \\ . \\ . \end{pmatrix} = \begin{pmatrix} y_m - \lambda_m\, x_{m-1} \\ y_{m+1} \\ . \\ . \end{pmatrix}, \tag{6.11}$$

combined with

$$\begin{pmatrix} \frac{\delta_1}{\delta_0} & c_1 & 0 & ... \\ 0 & \frac{\delta_2}{\delta_1} & c_2 & ... \\ 0 & 0 & \frac{\delta_3}{\delta_2} & ... \\ . & . & . & ... \end{pmatrix} \begin{pmatrix} x_m \\ x_{m+1} \\ . \end{pmatrix} = \begin{pmatrix} z_m \\ z_{m+1} \\ . \end{pmatrix}. \tag{6.12}$$

Both infinite systems (6.11) and (6.12) can be explicitly solved.

System (6.11) leads to

$$z_m = y_m - \lambda_m x_{m-1},$$

and for any $k \geq 1$

$$z_{m+k} = y_{m+k} + \sum_{l=1}^{k} (-1)^l\, a_{k-l+2}\,..\,a_{k+1}\,\frac{\delta_{k-l}}{\delta_k}\,y_{m+k-l} + (-1)^{k+1}\,a_2\,..\,a_{k+1}\,\frac{\delta_0}{\delta_k}\,\lambda_m x_{m-1},$$

which we rewrite with infinite matrix/vectors notations as

$$z_I = {L_I}^{-1}[y_I - \lambda_m\, x_{m-1}\mathbf{e_1}] = {L_I}^{-1} y_I - \lambda_m x_{m-1}\, v_I, \tag{6.13}$$

where

$$z_I = \begin{pmatrix} z_m \\ z_{m+1} \\ z_{m+2} \\ \\ \vdots \end{pmatrix}, \quad y_I = \begin{pmatrix} y_m \\ y_{m+1} \\ y_{m+2} \\ \\ \vdots \end{pmatrix}, \quad v_I := {L_I}^{-1}\mathbf{e_1} = \begin{pmatrix} 1 \\ -a_2\,\frac{\delta_0}{\delta_1} \\ a_3\,a_2\,\frac{\delta_0}{\delta_2} \\ -a_4\,a_3\,a_2\,\frac{\delta_0}{\delta_3} \\ \vdots \end{pmatrix}.$$

The second system (6.12) leads to the infinite sum (for any $k \geq 0$)

$$x_{m+k} = \frac{\delta_k}{\delta_{k+1}} z_{m+k} + \sum_{l=1}^{\infty} (-1)^l\,\frac{\delta_k}{\delta_{k+l+1}}\,c_{k+1}\,..\,c_{k+l}\,z_{m+k+l},$$

which we also rewrite with infinite matrix/vector notations as

$$x_I = {U_I}^{-1} z_I. \tag{6.14}$$

Coupling (6.13) and (6.14), we end up with

$$x_I = {U_I}^{-1} z_I = {U_I}^{-1}[{L_I}^{-1} y_I - \lambda_m x_{m-1}\, v_I] = {U_I}^{-1}{L_I}^{-1} y_I - \lambda_m x_{m-1} w_I, \tag{6.15}$$

where $w_I := U_I^{-1} v_I$. Denoting $\left({U_I}^{-1}{L_I}^{-1}\right)_{r_0}$ the first row of the infinite matrix ${U_I}^{-1}{L_I}^{-1}$ and $(w_I)_0$ the first element of w_I, we can rewrite the first line of (6.15) as

$$x_m = \left({U_I}^{-1}{L_I}^{-1}\right)_{r_0} y_I - \lambda_m x_{m-1}\,(w_I)_0\,. \tag{6.16}$$

We now investigate the finite part of the linear system (6.7), which is given by

$$D \begin{pmatrix} x_0 \\ x_1 \\ . \\ . \\ x_{m-2} \\ x_{m-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ 0 \\ \beta_{m-1}\, x_m \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ . \\ . \\ y_{m-2} \\ y_{m-1} \end{pmatrix},$$

or, according to (6.16),

$$D \begin{pmatrix} x_0 \\ x_1 \\ . \\ . \\ x_{m-2} \\ x_{m-1} \end{pmatrix} + \beta_{m-1} \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ 0 \\ \left({U_I}^{-1}{L_I}^{-1}\right)_{r_0} y_I - \lambda_m x_{m-1}\,(w_I)_0 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ . \\ . \\ y_{m-2} \\ y_{m-1} \end{pmatrix}.$$

Letting

$$K := D - \beta_{m-1}\lambda_m \begin{pmatrix} 0 & 0 & ... & 0 & 0 \\ 0 & 0 & ... & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & ... & 0 & 0 \\ 0 & 0 & ... & 0 & (w_I)_0 \end{pmatrix},$$

we consider its inverse K^{-1}. We denote the last column of K^{-1} by $(K^{-1})_{c_{m-1}}$, its last row by $(K^{-1})_{r_{m-1}}$, and its last ("south-east") element by $(K^{-1})_{m-1,m-1}$. Then we obtain

$$\begin{aligned} x_F &= K^{-1}y_F - \beta_{m-1}\left\{\left({U_I}^{-1}{L_I}^{-1}\right)_{r_0} y_I\right\}(K^{-1})_{c_{m-1}} \\ &= K^{-1}y_F - \beta_{m-1}\left(\left\{(K^{-1})_{c_{m-1}}\right\} \otimes \left\{\left(U_I^{-1}L_I^{-1}\right)_{r_0}\right\}\right) y_I, \end{aligned} \tag{6.17}$$

using the tensor product notation. The last line of this identity reads

$$x_{m-1} = (K^{-1})_{r_{m-1}}\, y_F - \beta_{m-1}\left\{\left({U_I}^{-1}{L_I}^{-1}\right)_{r_0} y_I\right\}(K^{-1})_{m-1,m-1}. \tag{6.18}$$

Coming back to (6.15) and using (6.18), we see that

$$\begin{aligned} x_I &= {U_I}^{-1}{L_I}^{-1}y_I - \lambda_m x_{m-1} w_I \\ &= {U_I}^{-1}{L_I}^{-1}y_I \\ &\quad -\lambda_m\left[(K^{-1})_{r_{m-1}}\, y_F - \beta_{m-1}\left\{\left({U_I}^{-1}{L_I}^{-1}\right)_{r_0} y_I\right\}(K^{-1})_{m-1,m-1}\right] w_I \\ &= {U_I}^{-1}{L_I}^{-1}y_I - \lambda_m\, w_I\left\{(K^{-1})_{r_{m-1}}\, y_F\right\} \\ &\quad +\beta_{m-1}\lambda_m\,(K^{-1})_{m-1,m-1}\, w_I\left\{\left({U_I}^{-1}{L_I}^{-1}\right)_{r_0} y_I\right\} \\ &= -\lambda_m\left(\left\{w_I\right\} \otimes \left\{(K^{-1})_{r_{m-1}}\right\}\right) y_F \\ &\quad +\left({U_I}^{-1}{L_I}^{-1} + \beta_{m-1}\lambda_m\,(K^{-1})_{m-1,m-1}\left\{w_I\right\} \otimes \left\{\left({U_I}^{-1}{L_I}^{-1}\right)_{r_0}\right\}\right) y_I. \end{aligned} \tag{6.19}$$

Putting together (6.17) and (6.19), we end up with $(A^\dagger)^{-1} =$

$$\begin{pmatrix} K^{-1} & -\beta_{m-1}\left(\left\{(K^{-1})_{c_{m-1}}\right\}\otimes\left\{\left(U_I^{-1}L_I{}^{-1}\right)_{r_0}\right\}\right) \\ -\lambda_m\left\{w_I\right\}\otimes\left\{(K^{-1})_{r_{m-1}}\right\} & U_I{}^{-1}L_I{}^{-1}+\tilde{\Lambda} \end{pmatrix},$$

where

$$\tilde{\Lambda} := \beta_{m-1}\lambda_m\,(K^{-1})_{m-1,m-1}\left\{w_I\right\}\otimes\left\{\left(U_I{}^{-1}L_I{}^{-1}\right)_{r_0}\right\}.$$

In order to get an approximate (pseudo) inverse of $A^\dagger$, we would like to get a numerical approximation of K^{-1}. However the definition of K involves $(w_I)_0$, which cannot be explicitly computed. By definition, $w_I = U_I^{-1}L_I^{-1}\mathbf{e_1}$, so using again the computations made in this Section, we get

$$\begin{aligned}(w_I)_0 &= \left(U_I^{-1}v_I\right)_0 \\ &= \frac{\delta_0}{\delta_1}v_m + \sum_{l=1}^{\infty}(-1)^l\frac{\delta_0}{\delta_{l+1}}c_1\dots c_l\,v_{m+l} \\ &= \frac{\delta_0}{\delta_1} + \sum_{l=1}^{\infty}\frac{\delta_0^2}{\delta_l\delta_{l+1}}c_1\dots c_l\,a_2\dots a_{l+1}.\end{aligned}$$

Given a computational parameter L, we define

$$\tilde{w} := \frac{\delta_0}{\delta_1} + \sum_{l=1}^{L-1}\frac{\delta_0^2}{\delta_l\delta_{l+1}}c_1\dots c_l\,a_2\dots a_{l+1}, \tag{6.20}$$

and

$$\tilde{K} := D - \beta_{m-1}\lambda_m\begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & \tilde{w} \end{pmatrix}.$$

We now can consider A_m a numerically computed inverse of $\tilde{K}$ and then define the approximate (pseudo) inverse of $A^\dagger$ as $A :=$

$$\begin{pmatrix} A_m & -\beta_{m-1}\left(\left\{(A_m)_{c_{m-1}}\right\}\otimes\left\{\left(U_I^{-1}L_I{}^{-1}\right)_{r_0}\right\}\right) \\ -\lambda_m\left\{w_I\right\}\otimes\left\{(A_m)_{r_{m-1}}\right\} & U_I{}^{-1}L_I{}^{-1}+\Lambda \end{pmatrix}, \tag{6.21}$$

where

$$\Lambda := \beta_{m-1}\lambda_m\,(A_m)_{m-1,m-1}\left\{w_I\right\}\otimes\left\{\left(U_I{}^{-1}L_I{}^{-1}\right)_{r_0}\right\}.$$

Lemma 6.2.1. *Assume that $m \geq k_0$ and $\delta < \frac{1}{2}$. Then U_I^{-1} maps Ω^s into Ω^{s+s_L}.*

Proof. Let $z_I \in \Omega^s$ and $x_I = U_I^{-1}z_I$. Using (6.14) and the formula above, we get

$$\begin{aligned}|x_{m+k}| &\leq \frac{|\delta_k|}{|\delta_{k+1}|}\,|z_{m+k}| + \sum_{l=1}^{\infty}\frac{|\delta_k|}{|\delta_{k+l+1}|}\;|c_{k+1}|\;..\;|c_{k+l}|\;|z_{m+k+l}| \\ &\leq \frac{|\delta_k|}{|\delta_{k+1}|}\,|z_{m+k}| + \sum_{l=1}^{\infty}\delta^l\frac{|\delta_k|}{|\delta_{k+l+1}|}\;|b_{k+1}|\;..\;|b_{k+l}|\;|z_{m+k+l}|\,.\end{aligned} \tag{6.22}$$

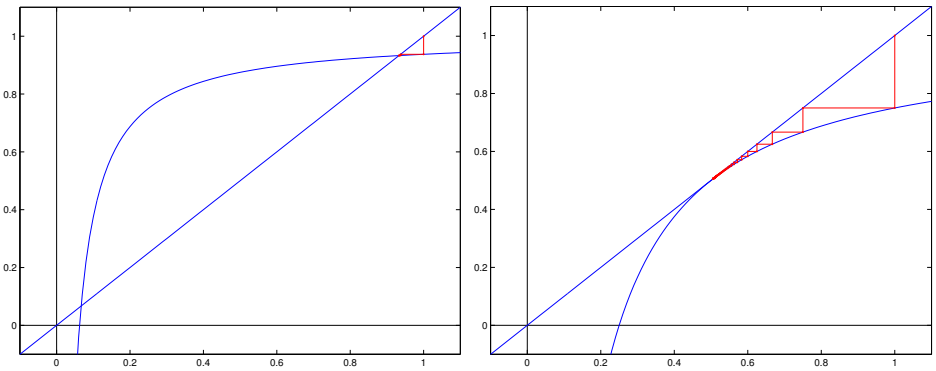
Now remember that for all $k \geq 2$, $\delta_k = b_k\,\delta_{k-1} - a_k\,c_{k-1}\delta_{k-2}$, so

$$\begin{aligned}\frac{|\delta_k|}{|\delta_{k-1}|\,|b_k|} &\geq 1 - \frac{|a_k|\,|c_{k-1}|\,|\delta_{k-2}|}{|b_k|\,|\delta_{k-1}|} \\ &\geq 1 - \frac{\delta^2\,|b_{k-1}|\,|\delta_{k-2}|}{|\delta_{k-1}|}.\end{aligned}$$

We introduce $u_k := \dfrac{|\delta_k|}{|\delta_{k-1}|\,|b_k|}$ which then satisfies

$$\begin{cases} u_1 = 1, \\ u_k \geq 1 - \dfrac{\delta^2}{u_{k-1}}, \quad \forall\ k \geq 2. \end{cases}$$

The study of the inductive sequence defined as above, but with $\geq$ replaced by $=$, yields that for any k, $\gamma \leq u_k \leq 1$, where $\gamma := \frac{1}{2} + \sqrt{\frac{1}{4} - \delta^2}$ is the largest root of $x = 1 - \frac{\delta^2}{x}$ (see Figure 6.1).



(a) $\delta = 0.2$ (b) $\delta = 0.5$

Figure 6.1 – The iterations of $u_{n+1} = 1 - \delta^2/u_n$ with $u_1 = 1$.

We can then rewrite (6.22) in order to get

$$\begin{aligned}|x_{m+k}| &\leq \frac{|\delta_k|}{|\delta_{k+1}|}\,|z_{m+k}| + \sum_{l=1}^{\infty} \delta^l \frac{|\delta_k|\,..\,|\delta_{k+l}|}{|\delta_{k+1}|\,..\,|\delta_{k+l+1}|}\,|b_{k+1}|\,..\,|b_{k+l}|\,|z_{m+k+l}| \\ &\leq \frac{|\delta_k|}{|\delta_{k+1}|}\,|z_{m+k}| + \sum_{l=1}^{\infty} \delta^l \frac{1}{u_{k+1}}\,..\,\frac{1}{u_{k+l}}\frac{|\delta_{k+l}|}{|\delta_{k+l+1}|}\,|z_{m+k+l}| \\ &\leq \sum_{l=0}^{\infty} \left(\frac{\delta}{\gamma}\right)^l \frac{|\delta_{k+l}|}{|\delta_{k+l+1}|}\,|z_{m+k+l}| \\ &\leq \sum_{l=0}^{\infty} \left(\frac{\delta}{\gamma}\right)^l \frac{1}{\gamma\,|b_{k+l+1}|}\,|z_{m+k+l}| \\ &\leq \frac{\|z_I\|_s}{C_1\gamma} \sum_{l=0}^{\infty} \left(\frac{\delta}{\gamma}\right)^l \frac{1}{(k+l+1)^{s_L}\,(m+k+l)^s}.\end{aligned} \tag{6.23}$$

Finally, since $\delta < \dfrac{1}{2} < \gamma$,

$$|x_{m+k}|\,(m+k)^{s+s_L} \leq \frac{\|z_I\|_s}{C_1\gamma}\,\frac{1}{1-\frac{\delta}{\gamma}}\,\frac{(m+k)^{s+s_L}}{(k+1)^{s_L}\,(m+k)^s}$$

and $x_I \in \Omega^{s+s_L}$. □

Lemma 6.2.2. *Assume that $m \geq k_0$ and $\delta < \frac{1}{2}$, Then L_I^{-1} maps Ω^s into Ω^s.*

Proof. Let $y_I \in \Omega^s$ and $z_I = L_I^{-1} y_I$. Using (6.13) and the formula above (without the last term since we do not consider here $L_I^{-1}(y_I - \lambda_m \, x_{m-1}\mathbf{e_1})$), we get

$$\begin{aligned}
|z_{m+k}| &\leq |y_{m+k}| + \sum_{l=1}^{k} \frac{|\delta_{k-l}|}{|\delta_k|} \, |a_{k-l+2}| \, .. \, |a_{k+1}| \, |y_{m+k-l}| \\
&\leq |y_{m+k}| + \sum_{l=1}^{k} \delta^l \frac{|\delta_{k-l}|}{|\delta_k|} \, |b_{k-l+2}| \, .. \, |b_{k+1}| \, |y_{m+k-l}| \\
&\leq |y_{m+k}| + \sum_{l=1}^{k} \delta^l \frac{|\delta_{k-l}| \, .. \, |\delta_{k-1}|}{|\delta_{k-l+1}| \, .. \, |\delta_k|} \, \frac{|b_{k-l+1}| \, |b_{k-l+2}| \, .. \, |b_{k+1}|}{|b_{k-l+1}|} \, |y_{m+k-l}| \\
&\leq |y_{m+k}| + \sum_{l=1}^{k} \delta^l \frac{1}{u_{k-l+1}} .. \frac{1}{u_k} \, \frac{|b_{k+1}|}{|b_{k-l+1}|} \, |y_{m+k-l}| \, ,
\end{aligned}$$

where we use the sequence u_k introduced in the previous proof. We get

$$|z_{m+k}| \leq \sum_{l=0}^{k} \left(\frac{\delta}{\gamma}\right)^l \frac{|b_{k+1}|}{|b_{k-l+1}|} \, |y_{m+k-l}| \, , \tag{6.24}$$

and

$$\begin{aligned}
|z_{m+k}| \, (m+k)^s &\leq \frac{C_2 \, \|y\|_s}{C_1} \sum_{l=0}^{k} \left(\frac{\delta}{\gamma}\right)^l \left(\frac{k+1}{k+1-l}\right)^{s_L} \left(\frac{m+k}{m+k-l}\right)^s \\
&\leq \frac{C_2 \, \|y\|_s}{C_1} \sum_{l=0}^{k} \left(\frac{\delta}{\gamma}\right)^l \left(\frac{m+k}{k+1-l}\right)^{s+s_L} .
\end{aligned}$$

For any $k \geq m$, we then have

$$\begin{aligned}
|z_{m+k}| \, (m+k)^s &\leq \frac{2^{s+s_L} C_2 \, \|y\|_s}{C_1} \left(\sum_{l=0}^{\left[\frac{k}{2}\right]} \left(\frac{\delta}{\gamma}\right)^l \left(\frac{k}{k+1-l}\right)^{s+s_L} \right. \\
&\qquad \left. + \sum_{l=\left[\frac{k}{2}\right]+1}^{k} \left(\frac{\delta}{\gamma}\right)^l \left(\frac{k}{k+1-l}\right)^{s+s_L} \right) \\
&\leq \frac{2^{s+s_L} C_2 \, \|y\|_s}{C_1} \left(2^{s+s_L} \sum_{l=0}^{\left[\frac{k}{2}\right]} \left(\frac{\delta}{\gamma}\right)^l + \left(\frac{\delta}{\gamma}\right)^{\frac{k}{2}} \sum_{l=\left[\frac{k}{2}\right]+1}^{k} k^{s+s_L} \right) \\
&\leq \frac{2^{s+s_L} C_2 \, \|y\|_s}{C_1} \left(\frac{2^{s+s_L}}{1-\frac{\delta}{\gamma}} + \left(\frac{\delta}{\gamma}\right)^{\frac{k}{2}} \frac{k^{s+s_L+1}}{2} \right) ,
\end{aligned}$$

which is bounded uniformly in k since the last term goes to 0 when k goes to ∞, and the proof is complete. □

Proposition 6.2.3. *Assume that $m \geq k_0$ and $\delta < \frac{1}{2}$. Then A maps Ω^s into Ω^{s+s_L}.*

Proof. Consider $y = (y_F, y_I)^T \in \Omega^s$. Let $x = (x_F, x_I)^T = Ay$. Then, by definition of the operator A in (6.21),

$$\begin{aligned}
x_F &= A_m y_F - \beta_{m-1} \left(\left\{ (A_m)_{c_{m-1}} \right\} \otimes \left\{ \left({U_I}^{-1} {L_I}^{-1} \right)_{r_0} \right\} \right) y_I \\
&= A_m y_F - \beta_{m-1} \left\{ \left({U_I}^{-1} {L_I}^{-1} \right)_{r_0} y_I \right\} (A_m)_{c_{m-1}} .
\end{aligned}$$

By the previous lemmas, ${U_I}^{-1}{L_I}^{-1}y_I \in \Omega^{s+sL}$, and in particular $\left({U_I}^{-1}{L_I}^{-1}\right)_{r_0} y_I = \left({U_I}^{-1}{L_I}^{-1}y_I\right)_0$ is well defined and so is x_F.

Using (6.21) again,

$$x_I = -\lambda_m \left(\left\{ w_I \right\} \otimes \left\{ (A_m)_{r_{m-1}} \right\} \right) y_F + {U_I}^{-1}{L_I}^{-1}y_I + \Lambda y_I.$$

Remember that $w_I = U_I^{-1}{L_I}^{-1}\mathbf{e_1}$, so that $w_I \in \Omega^s$ for any s. According to the previous lemmas and the definition of Λ (see (6.21)), we see that $x_I \in \Omega^{s+s_L}$. □

6.3 Computations of fixed points of the operator T

Our main motivation for computing approximate inverses is to prove existence, in a mathematically rigorous sense, of a fixed point of the Newton-like operator T in a set centered at a numerical approximation $\bar{x}$. The Newton-like operator has the form

$$T(x) = x - Af(x), \tag{6.25}$$

where A is the approximate inverse (6.21) of $Df(\bar{x})$ computed using the theory of Section 6.2. Since f maps Ω^s into Ω^{s-s_L} and A maps Ω^s into Ω^{s+s_L} (thanks to Proposition 6.2.3), we see that T maps the Banach space Ω^s into itself. Our goal is to obtain explicit bounds allowing us to show that a given T is a contraction on the ball $B_{\bar{x}}(r)$, which yields the existence of a fixed point of T (and thus of a zero of f). The fixed point theorem that we use (see Theorem 6.3.1) requires bounds on T and its derivative. We get formulas for these bounds in Sections 6.3.2 and 6.3.3, and then explain in Section 6.3.4 how to use the so-called radii polynomials in order to find a radius $r > 0$ such that $T(B_{\bar{x}}(r)) \subset B_{\bar{x}}(r)$, and such that T is a contraction on $B_{\bar{x}}(r)$.

Before proceeding further, we endow Ω^s with the operation of discrete convolution. More precisely, given $x = (x_k)_{k\geq 0}, y = (y_k)_{k\geq 0} \in \Omega^s$, we extend x, y symmetrically by $\tilde{x} = (x_k)_{k\in\mathbb{Z}}, \tilde{y} = (y_k)_{k\in\mathbb{Z}}$ where $\tilde{x}_{-k} = x_k$, $\tilde{y}_{-k} = y_k$, for $k \geq 1$. The discrete convolution of x and y is then denoted by $x * y$, and defined by the (infinite) sum

$$(x * y)_k = \sum_{\substack{k_1+k_2=k \\ k_1,k_2\in\mathbb{Z}}} \tilde{x}_{k_1}\tilde{y}_{k_2}.$$

It is known that for $s > 1$, $(\Omega^s, *)$ is an algebra (e.g. see [79]), that is, if $x, y \in \Omega^s$, then $x * y \in \Omega^s$. This will be useful when we shall look for a bound such as (6.27) below. We start with a classical theorem, whose proof is standard (e.g. see the proof of Lemma 3.3 in [118]) and is a direct consequence of the contraction mapping theorem.

Theorem 6.3.1. *For a given $s > 1$, consider $T\colon \Omega^s \to \Omega^s$ with $T = (T_k)_{k\geq 0}$, $T_k \in \mathbb{R}$. Assume that there exists a point $\bar{x} \in \Omega^s$ and vectors $Y = \{Y_k\}_{k\geq 0}$ and $Z = \{Z_k(r)\}_{k\geq 0}$, with $Y_k, Z_k(r) \in \mathbb{R}$, satisfying (for all $k \geq 0$)*

$$|(T(\bar{x}) - \bar{x})_k| \leq Y_k, \tag{6.26}$$

and

$$\sup_{b_1,b_2\in B_0(r)} \left| \left[DT(\bar{x} + b_1)b_2 \right]_k \right| \leq Z_k(r). \tag{6.27}$$

If there exists $r > 0$ such that $\|Y + Z(r)\|_s < r$, then the operator T is a contraction in $B_{\bar{x}}(r)$ and there exists a unique $\hat{x} \in B_{\bar{x}}(r)$ such that $T(\hat{x}) = \hat{x}$.

We shall see how to get the bounds Y (Section 6.3.2) and the bounds $Z(r)$ (Section 6.3.3), and we shall provide an efficient way of finding a radius $r > 0$ such that $\|Y + Z(r)\|_s < r$ (Section 6.3.4). The first step however consists in looking for bounds on A. More precisely, we need some estimates in order to control the action of $U_I^{-1}L_I^{-1}$. This is the goal of the following Subsection.

6.3.1 Some preliminary computations

We introduce the notations

$$\theta := \frac{\delta}{\gamma} \quad \text{and} \quad \eta := \frac{1}{\gamma(1-\theta^2)}. \tag{6.28}$$

Lemma 6.3.2. *Let $y_I = (y_m, y_{m+1}, \ldots)^T$ be an infinite vector and $x_I = U_I^{-1} L_I^{-1} y_I$. Assume that $m \geq k_0$ and $\delta < \frac{1}{2}$. Then, for all $k \geq 0$,*

$$|x_{m+k}| \leq \eta \left(\sum_{j=0}^{k} \theta^{k-j} \frac{|y_{m+j}|}{|\mu_{m+j}|} + \sum_{j=k+1}^{\infty} \theta^{j-k} \frac{|y_{m+j}|}{|\mu_{m+j}|} \right).$$

Proof. We again introduce $z_I = L_I^{-1} y_I$. Combining (6.23) from Lemma 6.2.1 and (6.24) from Lemma 6.2.2, we get

$$\begin{aligned}
|x_{m+k}| &\leq \frac{1}{\gamma} \sum_{l=0}^{\infty} \sum_{j=0}^{k+l} \theta^{k+2l-j} \frac{|y_{m+j}|}{|b_{j+1}|} \\
&= \frac{1}{\gamma} \left(\sum_{j=0}^{k} \frac{|y_{m+j}|}{|b_{j+1}|} \sum_{l=0}^{\infty} \theta^{k+2l-j} + \sum_{j=k+1}^{\infty} \frac{|y_{m+j}|}{|b_{j+1}|} \sum_{l=j-k}^{\infty} \theta^{k+2l-j} \right) \\
&= \frac{1}{\gamma} \left(\sum_{j=0}^{k} \frac{|y_{m+j}|}{|b_{j+1}|} \frac{\theta^{k-j}}{1-\theta^2} + \sum_{j=k+1}^{\infty} \frac{|y_{m+j}|}{|b_{j+1}|} \frac{\theta^{j-k}}{1-\theta^2} \right) \\
&= \eta \left(\sum_{j=0}^{k} \theta^{k-j} \frac{|y_{m+j}|}{|\mu_{m+j}|} + \sum_{j=k+1}^{\infty} \theta^{j-k} \frac{|y_{m+j}|}{|\mu_{m+j}|} \right).
\end{aligned}$$

□

In particular, we immediately obtain the two following corollaries (always under the assumptions of Lemma 6.3.2) which will be useful in the sequel.

Corollary 6.3.3. *Recall* (6.28). *Then, for $w_I = (w_m, w_{m+1}, \ldots)^T := U_I^{-1} L_I^{-1} \mathbf{e_1}$, we have*

$$|w_{m+k}| \leq \eta \theta^k \frac{1}{|\mu_m|}, \qquad \textit{for all } k \geq 0. \tag{6.29}$$

Corollary 6.3.4. *If y is such that $y_{m+k} = 0$ for any $k \geq n$, then*

$$\forall\ k \leq n-2, \quad |x_{m+k}| \leq \eta \left(\sum_{l=0}^{k} \theta^{k-l} \frac{|y_{m+l}|}{|\mu_{m+l}|} + \sum_{l=k+1}^{n-1} \theta^{l-k} \frac{|y_{m+l}|}{|\mu_{m+l}|} \right) \tag{6.30}$$

and

$$\forall\ k \geq n-1, \quad |x_{m+k}| \leq \eta \theta^k \sum_{l=0}^{n-1} \frac{|y_{m+l}|}{\theta^l\, |\mu_{m+l}|}. \tag{6.31}$$

More generally, we will also need in the next two Subsections a uniform bound on $|x_{m+k}|\,(m+k)^{s+s_L}$ for k large enough. We assume here that $m \geq 2$ (which will always be the case in practice), and define for any integer M

$$\chi := \theta^{\frac{M}{2}} \frac{M}{2} \left(\frac{m+M}{m} \right)^{s+s_L} + \theta^{\sqrt{M}} \frac{M}{2} 2^{s+s_L} + \frac{1}{1-\theta} \left(\frac{m+M}{m+M-\sqrt{M}-1} \right)^{s+s_L}.$$

Proposition 6.3.5. *Suppose that M satisfies*

$$M \geq \max\left(\frac{-m\ln\sqrt{\theta}-s-s_L-1-\sqrt{(m\ln\sqrt{\theta}+s+s_L+1)^2-4m\ln\sqrt{\theta}}}{2\ln\sqrt{\theta}}, \frac{4}{(\ln\theta)^2}, m\right). \tag{6.32}$$

Then for all $k < M$,

$$|x_{m+k}|\,(m+k)^{s+s_L} \leq \frac{\eta\|y_I\|_s}{C_1}\left(\sum_{l=0}^{k}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L}+\frac{\theta}{1-\theta}\right), \tag{6.33}$$

and for all $k \geq M$

$$|x_{m+k}|\,(m+k)^{s+s_L} \leq \frac{\eta\|y_I\|_s}{C_1}\left(\chi+\frac{\theta}{1-\theta}\right). \tag{6.34}$$

Proof. Thanks to Lemma 6.3.2,

$$\begin{aligned}|x_{m+k}|\,(m+k)^{s+s_L} &\leq \frac{\eta\|y_I\|_s}{C_1}\left(\sum_{l=0}^{k}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L}+\sum_{l=k+1}^{\infty}\theta^{l-k}\left(\frac{m+k}{m+l}\right)^{s+s_L}\right)\\ &\leq \frac{\eta\|y_I\|_s}{C_1}\left(\sum_{l=0}^{k}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L}+\frac{\theta}{1-\theta}\right).\end{aligned}$$

Then for $k \geq M$, we split the remaining sum

$$\begin{aligned}\sum_{l=0}^{k}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L} &= \sum_{l=0}^{\left[\frac{k}{2}\right]-1}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L}+\sum_{l=\left[\frac{k}{2}\right]}^{k-\left[\sqrt{k}\right]-1}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L}\\ &\quad+\sum_{l=k-\left[\sqrt{k}\right]}^{k}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L}\\ &\leq \theta^{\frac{k}{2}}\frac{k}{2}\left(\frac{m+k}{m}\right)^{s+s_L}+\theta^{\sqrt{k}}\frac{k}{2}2^{s+s_L}\\ &\quad+\frac{1}{1-\theta}\left(\frac{m+k}{m+k-\sqrt{k}-1}\right)^{s+s_L}\\ &\leq \theta^{\frac{M}{2}}\frac{M}{2}\left(\frac{m+M}{m}\right)^{s+s_L}+\theta^{\sqrt{M}}\frac{M}{2}2^{s+s_L}\\ &\quad+\frac{1}{1-\theta}\left(\frac{m+M}{m+M-\sqrt{M}-1}\right)^{s+s_L}\\ &= \chi.\end{aligned}$$

The justification of the last inequality is contained in the following three lemmas. □

Lemma 6.3.6. *If M satisfies* (6.32)*, then for all $k \geq M$*

$$\theta^{\frac{k}{2}}\frac{k}{2}\left(\frac{m+k}{m}\right)^{s+s_L} \leq \theta^{\frac{M}{2}}\frac{M}{2}\left(\frac{m+M}{m}\right)^{s+s_L}.$$

Proof. For $x > 0$, let $\varphi_1(x) := \theta^{\frac{x}{2}}x(m+x)^{s+s_L}$, whose derivative is

$$\begin{aligned}\varphi_1'(x) &= \sqrt{\theta}^x\left(\left(\ln\sqrt{\theta}\right)x(m+x)^{s+s_L}+(m+x)^{s+s_L}+(s+s_L)x\,(m+x)^{s+s_L-1}\right)\\ &= (m+x)^{s+s_L-1}\sqrt{\theta}^x\left(\left(\ln\sqrt{\theta}\right)(m+x)x+(m+x)+(s+s_L)x\right)\\ &= (m+x)^{s+s_L-1}\sqrt{\theta}^x\left(\left(\ln\sqrt{\theta}\right)x^2+\left(m\ln\sqrt{\theta}+s+s_L+1\right)x+m\right).\end{aligned}$$

For $0 < \theta < 1$, the discriminant of $\ln\sqrt{\theta}x^2 + \left(m\ln\sqrt{\theta} + s + s_L + 1\right)x + m$ given by

$$\Delta := \left(m\ln\sqrt{\theta} + s + s_L + 1\right)^2 - 4m\ln\sqrt{\theta},$$

is positive. Since M satisfies (6.32), $\varphi_1'(x) \leq 0$ for any $x \geq M$ and so $\varphi_1(k) \leq \varphi_1(M)$ for all $k \geq M$. □

Lemma 6.3.7. *If M satisfies* (6.32)*, then for all $k \geq M$,*

$$\theta^{\sqrt{k}}\frac{k}{2}2^{s+s_L} \leq \theta^{\sqrt{M}}\frac{M}{2}2^{s+s_L}.$$

Proof. Let $\varphi_2(x) := \theta^{\sqrt{x}}x$. Then

$$\varphi_2'(x) = \theta^{\sqrt{x}}\left(\frac{\ln\theta}{2\sqrt{x}}x + 1\right) = \frac{\theta^{\sqrt{x}}}{2}\left(\sqrt{x}\ln\theta + 2\right).$$

Hence, for $x \geq \dfrac{4}{(\ln\theta)^2}$, $\varphi_2'(x) \leq 0$ and so $\varphi_2(k) \leq \varphi_2(M)$ for all $k \geq M$. □

Lemma 6.3.8. *If M satisfies* (6.32)*, then for all $k \geq M$,*

$$\frac{1}{1-\theta}\left(\frac{m+k}{m+k-\sqrt{k}-1}\right)^{s+s_L} \leq \frac{1}{1-\theta}\left(\frac{m+M}{m+M-\sqrt{M}-1}\right)^{s+s_L}.$$

Proof. Let $\varphi_3(x) := \dfrac{m+x}{m+x-\sqrt{x}-1}$. Then

$$\varphi_3'(x) = \frac{m+x-\sqrt{x}-1-(m+x)\left(1-\frac{1}{2\sqrt{x}}\right)}{(m+x-\sqrt{x}-1)^2} = -\frac{x+2\sqrt{x}-m}{2\sqrt{x}\,(m+x-\sqrt{x}-1)^2}.$$

Hence, for $x \geq m$, $\varphi_3'(x) \leq 0$ and $\varphi_3(k) \leq \varphi_3(M)$ for all $k \geq M$. □

Finally, we will need to bound the error made by using $\tilde{w}$ instead of $(w_I)_0$ for the definition (6.21) of A.

Lemma 6.3.9. *Assume that $L \geq k_0$ and $\delta < \frac{1}{2}$. Then*

$$|(w_I)_0 - \tilde{w}| \leq \frac{\theta^{2L}}{|\mu_m|\,(1-\theta^2)}. \tag{6.35}$$

Proof. Using (6.5) together with the sequence (u_l) introduced in the proof of Lemma 6.2.1, we get

$$\begin{aligned}
|(w_I)_0 - \tilde{w}| &\leq \sum_{l=L}^{\infty}\frac{|\delta_0|^2}{|\delta_l|\,|\delta_{l+1}|}\,|c_1|\dots|c_l|\;|a_2|\dots|a_{l+1}| \\
&\leq \frac{|\delta_0|}{|\delta_1|}\sum_{l=L}^{\infty}\delta^{2l}\left(\frac{1}{u_1}\cdots\frac{1}{u_l}\right)\left(\frac{1}{u_2}\cdots\frac{1}{u_{l+1}}\right) \\
&\leq \frac{1}{|\mu_m|}\sum_{l=L}^{\infty}\theta^{2l} \\
&= \frac{\theta^{2L}}{|\mu_m|\,(1-\theta^2)}.
\end{aligned}$$

□

6.3.2 Computation of the Y bounds

From now on, we shall assume for the sake of clarity that the nonlinearity N of f in (6.1) is a polynomial of degree two. The generalization to a polynomial nonlinearity of higher degree could be obtained thanks to the use of the estimates developed in [118] in order to bound terms like

$$\left(x^1 * \ldots * x^p\right)_n$$

where $x^1, \ldots, x^p \in B_0(r)$. Moreover, as long as one is interested in problems with nonlinearities built from elementary functions of mathematical physics (powers, exponential, trigonometric functions, rational, Bessel, elliptic integrals, etc.), our method is applicable. Indeed, since these nonlinearities are themselves solutions of low order linear or polynomial ODEs, they can be appended to the original problem of interest in order to obtain polynomial nonlinearities, albeit in a higher number of variables. This standard trick is explained in more details in [145], and is used in [154] to prove existence of periodic solutions in the planar circular restricted three body problem.

With this in mind, we are ready to compute the bound Y appearing in Theorem 6.3.1. In everything that follows, $|\cdot|$, when applied to vectors or matrices (even infinite dimensional), must be understood component-wise.

The main estimate of this subsection, that is the bound on Y, is presented in the following Proposition:

Proposition 6.3.10. *Consider an integer M such that*

$$M \geq \max\left(\frac{-s}{\ln\theta} - m, m-2\right), \tag{6.36}$$

and define $Y = (Y_k)_{k\geq 0}$ component-wise by

$$Y_F := \left|A_m\left(f(\bar{x})\right)_F\right| + \left|\beta_{m-1}\right| \eta \left(\sum_{l=0}^{m-2} \theta^l \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|}\right) \left|\left(A_m\right)_{c_{m-1}}\right|, \tag{6.37}$$

$$\begin{aligned} Y_{m+k} := &\left(\left|\left(A_m\right)_{r_{m-1}} f(\bar{x})_F\right| + \left|\beta_{m-1}\left(A_m\right)_{m-1,m-1}\right| \eta \left(\sum_{l=0}^{m-2} \theta^l \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|}\right)\right) \eta\theta^k \frac{|\lambda_m|}{|\mu_m|} \\ &+ \eta \sum_{l=0}^{k} \theta^{k-l} \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|} + \eta \sum_{l=k+1}^{m-2} \theta^{l-k} \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|}, \quad \forall\ 0 \leq k \leq m-3, \end{aligned} \tag{6.38}$$

$$\begin{aligned} Y_{m+k} := &\left(\left|\left(A_m\right)_{r_{m-1}} f(\bar{x})_F\right| + \left|\beta_{m-1}\left(A_m\right)_{m-1,m-1}\right| \eta \left(\sum_{l=0}^{m-2} \theta^l \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|}\right)\right) \eta\theta^k \frac{|\lambda_m|}{|\mu_m|} \\ &+ \eta\theta^k \sum_{l=0}^{m-2} \frac{|f(\bar{x})|_{m+l}}{\theta^l |\mu_{m+l}|}, \quad \forall m-2 \leq k \leq M, \end{aligned} \tag{6.39}$$

and

$$Y_{m+k} := Y_{m+M} \frac{\omega_{m+M}^s}{\omega_{m+k}^s}, \quad \forall\ k > M. \tag{6.40}$$

Then

$$|T(\bar{x}) - \bar{x}| \leq Y.$$

Proof. By definition of T,

$$|T(\bar{x}) - \bar{x}| = |Af(\bar{x})|\,.$$

Note that since we suppose that f is at most quadratic, and since $\bar{x}$ is constructed in such a way that $\bar{x}_k = 0$ for all $k \geq m$, we get the identity $(f(\bar{x}))_{m+k} = 0$ for all $k \geq m-1$. Thanks to (6.21),

$$|(Af(\bar{x}))_F| \leq |A_m\,(f(\bar{x}))_F| + |\beta_{m-1}|\left|\left(U_I^{-1}L_I^{-1}\,(f(\bar{x}))_I\right)_0\right|\left|(A_m)_{c_{m-1}}\right|,$$

so that using (6.30) with $n = m-1$ and $k = 0$, we get

$$|(Af(\bar{x}))_F| \leq |A_m\,(f(\bar{x}))_F| + |\beta_{m-1}|\,\eta\left(\sum_{l=0}^{m-2}\theta^l\frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|}\right)\left|(A_m)_{c_{m-1}}\right|,$$

which provides the bound (6.37).

Using (6.21) again,

$$\begin{aligned}|(Af(\bar{x}))_I| \leq{}& |\lambda_m|\left(\left|(A_m)_{r_{m-1}}\,f(\bar{x})_F\right| + \left|\beta_{m-1}\,(A_m)_{m-1,m-1}\left(U_I^{-1}L_I^{-1}f(\bar{x})_I\right)_0\right|\right)|w_I| \\ &+ \left|U_I^{-1}L_I^{-1}f(\bar{x})_I\right|,\end{aligned}$$

so using (6.29), (6.30) and (6.31) (again with $n = m-1$), we get

$$\begin{aligned}\left|(Af(\bar{x}))_{m+k}\right| \leq{}& \left(\left|(A_m)_{r_{m-1}}\,f(\bar{x})_F\right| + \left|\beta_{m-1}\,(A_m)_{m-1,m-1}\right|\eta\left(\sum_{l=0}^{m-2}\theta^l\frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|}\right)\right)\eta\theta^k\frac{|\lambda_m|}{|\mu_m|} \\ &+ \eta\sum_{l=0}^{k}\theta^{k-l}\frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|} + \eta\sum_{l=k+1}^{m-2}\theta^{l-k}\frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|}, \quad \forall\ 0 \leq k \leq m-3,\end{aligned}$$

which provides the bound (6.38), and

$$\begin{aligned}\left|(Af(\bar{x}))_{m+k}\right| \leq{}& \left(\left|(A_m)_{r_{m-1}}\,f(\bar{x})_F\right| + \left|\beta_{m-1}\,(A_m)_{m-1,m-1}\right|\eta\left(\sum_{l=0}^{m-2}\theta^l\frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|}\right)\right)\eta\theta^k\frac{|\lambda_m|}{|\mu_m|} \\ &+ \eta\theta^k\sum_{l=0}^{m-2}\frac{|f(\bar{x})|_{m+l}}{\theta^l\,|\mu_{m+l}|}, \quad \forall k \geq m-2,\end{aligned}$$

which provides the bound (6.39). Finally, by (6.36), $\theta^k(m+k)^s \leq \theta^M(m+M)^s$ for all $k > M$, and we obtain the bound (6.40). □

We present in Section 6.3.4 the rationale behind the definition of Y_{m+k} for $k > M$.

6.3.3 Computation of the Z bounds

In order to compute the Z bounds from Theorem 6.3.1, we need to estimate the quantity

$$DT\,(\bar{x}+y)\,z = (I - ADf\,(\bar{x}+y))\,z = \left(I - AA^\dagger\right)z - A\left(Df\,(\bar{x}+y) - A^\dagger\right)z$$

for all $y, z \in B_0(r)$. We are going to bound each term separately in the next two Sub-subsections. We introduce the notation

$$W_F^s := \left(\frac{1}{\omega_0^s}, \ldots, \frac{1}{\omega_{m-1}^s}\right)^T. \tag{6.41}$$

Estimates for $(I - AA^\dagger)z$

In this Sub-subsection, we present the bound on $(I-AA^\dagger)z$, which constitutes the first part of a bound for Z.

Proposition 6.3.11. *Let M be an integer satisfying* (6.36). *We define $Z^1 = (Z^1_k)_{k\geq 0}$ componentwise by*

$$Z^1_F := \left(\left|I - A_m\tilde{K}\right| W^s_F + \frac{|\beta_{m-1}|\,|\lambda_m|\,\theta^{2L}}{|\mu_m|\,\omega^s_{m-1}(1-\theta^2)}\,|A_m|_{c_{m-1}}\right) r, \tag{6.42}$$

$$Z^1_{m+k} := \left(\left|I - A_m\tilde{K}\right|_{r_{m-1}} W^s_F + \frac{|\beta_{m-1}|\,|\lambda_m|\,\theta^{2L}}{|\mu_m|\,\omega^s_{m-1}(1-\theta^2)}\,|A_m|_{m-1,m-1}\right)\eta\theta^k\frac{|\lambda_m|}{|\mu_m|}r, \tag{6.43}$$
$$\forall\ 0\leq k\leq M,$$

and

$$Z^1_{m+k} := Z^1_{m+M}\frac{\omega^s_{m+M}}{\omega^s_{m+k}}, \quad \forall\ k > M. \tag{6.44}$$

Then for all $z\in B_0(r)$,

$$\left|\left(I - AA^\dagger\right)z\right| \leq Z^1.$$

Proof. Thanks to (6.6) and (6.21),

$$\begin{aligned}
\left(AA^\dagger z\right)_F &= A_m\left(Dz_F + \begin{pmatrix}0\\ \vdots\\ \beta_{m-1}z_m\end{pmatrix}\right)\\
&\quad - \beta_{m-1}\left(U_I^{-1}L_I^{-1}\left(Tz_I + \lambda_m z_{m-1}\mathbf{e_1}\right)\right)_0 (A_m)_{c_{m-1}}\\
&= A_mDz_F + \beta_{m-1}z_m\,(A_m)_{c_{m-1}} - \beta_{m-1}\left(z_m + \lambda_m z_{m-1}\,(w_I)_0\right)(A_m)_{c_{m-1}}\\
&= A_m\tilde{K}z_F + \beta_{m-1}\lambda_m\left(\tilde{w} - (w_I)_0\right)z_{m-1}\,(A_m)_{c_{m-1}},
\end{aligned}$$

and so

$$\left(\left(I - AA^\dagger\right)z\right)_F = \left(I - A_m\tilde{K}\right)z_F + \beta_{m-1}\lambda_m\left(\tilde{w} - (w_I)_0\right)z_{m-1}\,(A_m)_{c_{m-1}}.$$

For $z\in B_0(r)$ we have, using (6.35),

$$\left|\left(I - AA^\dagger\right)z\right|_F \leq \left(\left|I - A_m\tilde{K}\right| W^s_F + \frac{|\beta_{m-1}|\,|\lambda_m|\,\theta^{2L}}{|\mu_m|\,\omega^s_{m-1}(1-\theta^2)}\,|A_m|_{c_{m-1}}\right) r,$$

which provides the bound (6.42).

Using again (6.6) and (6.21), we get

$$\begin{aligned}
\left(AA^\dagger z\right)_I &= -\lambda_m\,(A_m)_{r_{m-1}}\left(Dz_F + \begin{pmatrix}0\\ \vdots\\ \beta_{m-1}z_m\end{pmatrix}\right)w_I + \left(U_I^{-1}L_I^{-1} + \Lambda\right)\left(Tz_I + \lambda_m z_{m-1}\mathbf{e_1}\right)\\
&= z_I + \lambda_m w_I\left(-\,(A_m)_{r_{m-1}}Dz_F - \beta_{m-1}\,(A_m)_{m-1,m-1}\,z_m + z_{m-1}\right.\\
&\quad \left.+\beta_{m-1}\,(A_m)_{m-1,m-1}\,(z_I + \lambda_m z_{m-1}w_I)_0\right)\\
&= z_I + \lambda_m\left(-\,(A_m)_{r_{m-1}}Dz_F + z_{m-1} + \beta_{m-1}\lambda_m\,(A_m)_{m-1,m-1}\,z_{m-1}\,(w_I)_0\right)w_I\\
&= z_I + \lambda_m\left(z_{m-1} - (A_m)_{r_{m-1}}\tilde{K}z_F + \beta_{m-1}\lambda_m\,(A_m)_{m-1,m-1}\,z_{m-1}\left(\tilde{w} - (w_I)_0\right)\right)w_I\\
&= z_I + \lambda_m\left(\left(I - A_m\tilde{K}\right)_{r_{m-1}}z_F + \beta_{m-1}\lambda_m\,(A_m)_{m-1,m-1}\,z_{m-1}\left(\tilde{w} - (w_I)_0\right)\right)w_I,
\end{aligned}$$

and so

$$\left(\left(I - AA^\dagger\right) z\right)_I = -\lambda_m \left((I - A_m K)_{r_{m-1}} z_F + \beta_{m-1}\lambda_m \left(A_m\right)_{m-1,m-1} z_{m-1} \left(\tilde{w} - (w_I)_0\right)\right) w_I.$$

For $z \in B_0(r)$ we have, using (6.29) and (6.35),

$$\begin{aligned}\left|\left(I - AA^\dagger\right) z\right|_{m+k} \leq \Big(& \left|I - A_m \tilde{K}\right|_{r_{m-1}} W_F^s \\ & + \frac{|\beta_{m-1}|\,|\lambda_m|\,\theta^{2L}}{|\mu_m|\,\omega_{m-1}^s (1-\theta^2)} \left|A_m\right|_{m-1,m-1} \Big) \eta\theta^k \frac{|\lambda_m|}{|\mu_m|} r, \quad \forall\ k \geq 0,\end{aligned}$$

which gives (6.43), as well as (6.44) thanks to (6.36). □

Estimates for $A\left(Df\left(\bar{x}+y\right) - A^\dagger\right) z$

This Sub-subsection is devoted to the exposition of a bound for $A\left(Df\left(\bar{x}+y\right) - A^\dagger\right) z$, which constitutes the second (and last) part of a bound for Z. This bound is detailed in Proposition 6.3.18.

Recall the assumption that the nonlinear part N is polynomial of degree 2. Hence, $Df\left(\bar{x}+y\right)$ can be written as a finite Taylor expansion

$$Df\left(\bar{x}+y\right) = Df\left(\bar{x}\right) + D^2 f\left(\bar{x}\right)(y),$$

and

$$\left(Df\left(\bar{x}+y\right) - A^\dagger\right) z = \left(Df\left(\bar{x}\right) - A^\dagger\right) z + D^2 f\left(\bar{x}\right)(y,z). \tag{6.45}$$

We are going to bound the two terms of (6.45) separately. Let us denote by σ the coefficient of degree 2 of f, that is $D^2 f\left(\bar{x}\right)(y,z) = 2\sigma(y * z)$. We bound this convolution product thanks to the following result:

Lemma 6.3.12. *Let $s \geq 2$ be an algebraic decay rate and $n \geq 6$, let $L \geq 1$ be computational parameters. For $x, y \in \Omega^s$ and for any $k \geq 0$,*

$$|(x*y)_k| \leq \alpha_k^s(n) \frac{\|x\|_s \|y\|_s}{\omega_k^s},$$

where

$$\alpha_k^s(n) := \begin{cases} 1 + 2\displaystyle\sum_{l=1}^{L} \frac{1}{l^s} + \frac{2}{(s-1)L^{s-1}}, & k = 0, \\ 2 + 2\displaystyle\sum_{l=1}^{L} \frac{1}{l^s} + \frac{2}{(s-1)L^{s-1}} + \sum_{l=1}^{k-1} \frac{k^s}{l^s (k-l)^s}, & 1 \leq k < n, \\ 2 + 2\displaystyle\sum_{l=1}^{L} \frac{1}{l^s} + \frac{2}{(s-1)L^{s-1}} + 2\left(\frac{n}{n-1}\right)^s \\ \quad + \left(\dfrac{4\ln(n-2)}{n} + \dfrac{\pi^2 - 6}{3}\right)\left(\dfrac{2}{n} + \dfrac{1}{2}\right)^s, & k \geq n. \end{cases}$$

Proof. See [121] for a proof of this bound and [79] for a similar bound for $1 < s < 2$. □

Remark 6.3.13. *It is important to notice here that $\alpha_k^s(n) = \alpha_n^s(n)$ for all $k \geq n$. From now on, we assume that m is taken larger or equal to 6, which will allow us to use Lemma 6.3.12 with $n = m$. Note that this condition is not stringent, since in practice more than 6 modes are usually needed in order to get a good numerical solution $\bar{x}$.*

We begin by bounding the first term of (6.45).

Proposition 6.3.14. *Define* $C^1 = C^1(\bar{x}) = \left(C^1_k(\bar{x})\right)_{k\geq 0}$ *component-wise by*

$$C^1_0(\bar{x}) := 0, \quad C^1_k(\bar{x}) := 2\,|\sigma| \sum_{l=m-k}^{m-1} \frac{|\bar{x}_l|}{\omega^s_{k+l}},\ \forall\ 1 \leq k \leq m-1,$$

and

$$C^1_{m+k}(\bar{x}) := \frac{2\,|\sigma|\,\alpha^s_m(m)\,\|\bar{x}\|_s}{\omega^s_{m+k}}, \quad \forall\ k \geq 0.$$

Then for all $z \in B_0(r)$

$$\left|\left(Df(\bar{x}) - A^\dagger\right) z\right| \leq C^1(\bar{x})r.$$

Proof. According to the definition of $A^\dagger$ in (6.6), we see that

$$\left(\left(Df(\bar{x}) - A^\dagger\right) z\right)_F = (Df(\bar{x})z)_F - Df^{(m)}(\bar{x})z_F - \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ 0 \\ \beta_{m-1}\, z_m \end{pmatrix}$$

$$= 2\sigma\left((\bar{x} * z)_F - (\bar{x} * z_F)_F\right),$$

where in the convolution product, z_F must be understood as the infinite vector $(z_F, 0, \ldots, 0, \ldots)^T$. Therefore, $\left(\left(Df(\bar{x}) - A^\dagger\right) z\right)_0 = 0$, and for all $z \in B_0(r)$,

$$\left|\left(Df(\bar{x}) - A^\dagger\right) z\right|_k \leq 2\,|\sigma|\, r \sum_{l=m-k}^{m-1} \frac{|\bar{x}_l|}{\omega^s_{k+l}}, \quad \forall\ \ 1 \leq k \leq m-1.$$

Then, remembering that $Df(\bar{x}) = \mathcal{L} + DN(\bar{x})$ and (6.6), we see that

$$\left(\left(Df(\bar{x}) - A^\dagger\right) z\right)_I = (DN(\bar{x})z)_I = 2\sigma\,(\bar{x} * z)_I\,,$$

so that using Lemma 6.3.12, for all $z \in B_0(r)$, we end up with the bound

$$\left|\left(Df(\bar{x}) - A^\dagger\right) z\right|_{m+k} \leq \frac{2\,|\sigma|\,\alpha^s_{m+k}(m)\,\|\bar{x}\|_s}{\omega^s_{m+k}} r, \quad \forall\ k \geq 0. \qquad \square$$

We now bound the second term of (6.45).

Proposition 6.3.15. *Recall* (6.41) *and define* $C^2 = (C^2_k)_{k\geq 0}$ *component-wise by*

$$C^2_k := \frac{2\,|\sigma|\,\alpha^s_k(m)}{\omega^s_k}, \quad \forall\ k \geq 0.$$

Then for all $y, z \in B_0(r)$

$$\left|D^2 f\,(\bar{x})\,(y,z)\right| \leq C^2 r^2.$$

Proof. Remembering that $D^2 f\,(\bar{x})\,(y,z) = 2\sigma(y * z)$, this is a consequence of Lemma 6.3.12. $\square$

Finally,

$$\left|A\left(Df\,(\bar{x}+y) - A^\dagger\right) z\right| \leq |A|\left(C^1(\bar{x})r + C^2 r^2\right),$$

and we are left to bound $|A|\,C^1(\bar{x})$ and $|A|\,C^2$.

Proposition 6.3.16. *Let M be an integer satisfying* (6.32) *and* (6.36). *We define $D^1 = (D^1_k)_{k\geq 0}$ component-wise by*

$$D^1_F(\bar{x}) := |A_m|\, C^1_F(\bar{x}) + \frac{2\,|\beta_{m-1}|\,\eta\,|\sigma|\,\alpha^s_m(m)\|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}}\,|A_m|_{c_{m-1}}\,, \tag{6.46}$$

$$\begin{aligned} D^1_{m+k}(\bar{x}) := &\left(|A_m|_{r_{m-1}}\, C^1_F(\bar{x}) + \frac{2\,|\beta_{m-1}|\,|A_m|_{m-1,m-1}\,\eta\,|\sigma|\,\alpha^s_m(m)\|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}}\right)\eta\frac{|\lambda_m|}{|\mu_m|}\theta^k \\ &+ \frac{2\eta\,|\sigma|\,\alpha^s_m(m)\|\bar{x}\|_s}{C_1\omega_{m+k}^{s+s_L}}\left(\sum_{l=0}^{k}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L} + \frac{\theta}{1-\theta}\right), \quad \forall\ 0\leq k<M, \end{aligned} \tag{6.47}$$

$$\begin{aligned} D^1_{m+M}(\bar{x}) := &\left(|A_m|_{r_{m-1}}\, C^1_F(\bar{x}) + \frac{2\,|\beta_{m-1}|\,|A_m|_{m-1,m-1}\,\eta\,|\sigma|\,\alpha^s_m(m)\|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}}\right)\eta\frac{|\lambda_m|}{|\mu_m|}\theta^M \\ &+ \frac{2\eta\,|\sigma|\,\alpha^s_m(m)\|\bar{x}\|_s}{C_1\omega_{m+M}^{s+s_L}}\left(\chi + \frac{\theta}{1-\theta}\right), \end{aligned} \tag{6.48}$$

and

$$D^1_{m+k}(\bar{x}) := D^1_{m+M}(\bar{x})\frac{\omega^s_{m+M}}{\omega^s_{m+k}}, \quad \forall\ k>M. \tag{6.49}$$

Then

$$|A|\, C^1(\bar{x}) \leq D^1(\bar{x}).$$

Proof. Thanks to (6.21),

$$\left(|A|\, C^1(\bar{x})\right)_F \leq |A_m|\, C^1_F(\bar{x}) + |\beta_{m-1}|\left|U_I^{-1}L_I^{-1}C^1_I(\bar{x})\right|_0 |A_m|_{c_{m-1}}\,,$$

and using (6.33)

$$\left|U_I^{-1}L_I^{-1}C^1_I(\bar{x})\right|_0 \leq \frac{\eta\|C^1_I(\bar{x})\|_s}{C_1(1-\theta)\omega_m^{s+s_L}} \leq \frac{2\eta\,|\sigma|\,\alpha^s_m(m)\|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}},$$

so that (6.46) holds. Still thanks to (6.21),

$$\begin{aligned} \left(|A|\, C^1(\bar{x})\right)_I &\leq |\lambda_m|\,|A_m|_{r_{m-1}}\, C^1_F(\bar{x})\,|w_I| + \left|U_I^{-1}L_I^{-1}C^1_I(\bar{x})\right| \\ &\quad + |\lambda_m|\,|\beta_{m-1}|\,|A_m|_{m-1,m-1}\left|U_I^{-1}L_I^{-1}C^1_I(\bar{x})\right|_0 |w_I| \\ &\leq |\lambda_m|\left(|A_m|_{r_{m-1}}\, C^1_F(\bar{x}) + \frac{2\,|\beta_{m-1}|\,|A_m|_{m-1,m-1}\,\eta\,|\sigma|\,\alpha^s_m(m)\|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}}\right)|w_I| \\ &\quad + \left|U_I^{-1}L_I^{-1}C^1_I(\bar{x})\right|. \end{aligned}$$

Using (6.29) and (6.33), we get

$$\begin{aligned} \left(|A|\, C^1(\bar{x})\right)_{m+k} \leq &\left(|A_m|_{r_{m-1}}\, C^1_F(\bar{x}) + \frac{2\,|\beta_{m-1}|\,|A_m|_{m-1,m-1}\,\eta\,|\sigma|\,\alpha^s_m(m)\|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}}\right)\eta\frac{|\lambda_m|}{|\mu_m|}\theta^k \\ &+ \frac{2\eta\,|\sigma|\,\alpha^s_m(m)\|\bar{x}\|_s}{C_1\omega_{m+k}^{s+s_L}}\left(\sum_{l=0}^{k}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L} + \frac{\theta}{1-\theta}\right), \quad \forall\ 0\leq k<M, \end{aligned}$$

so that (6.47) holds, and using (6.29) and (6.34), we get

$$\begin{aligned} \left(|A|\, C^1(\bar{x})\right)_{m+M} \leq &\left(|A_m|_{r_{m-1}}\, C^1_F(\bar{x}) + \frac{2\,|\beta_{m-1}|\,|A_m|_{m-1,m-1}\,\eta\,|\sigma|\,\alpha^s_m(m)\|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}}\right)\eta\frac{|\lambda_m|}{|\mu_m|}\theta^M \\ &+ \frac{2\eta\,|\sigma|\,\alpha^s_m(m)\|\bar{x}\|_s}{C_1\omega_{m+M}^{s+s_L}}\left(\chi + \frac{\theta}{1-\theta}\right), \end{aligned}$$

so that (6.48) holds. As before, (6.49) follows from (6.36). □

We get similar results for the second order term.

Proposition 6.3.17. *Let M be an integer satisfying* (6.32) *and* (6.36). *Define $D^2 = \left(D_k^2\right)_{k\geq 0}$ component-wise by*

$$D_F^2 := |A_m|\, C_F^2 + \frac{2\,|\beta_{m-1}|\,\eta\,|\sigma|\,\alpha_m^s(m)}{C_1(1-\theta)\omega_m^{s+s_L}}\,|A_m|_{c_{m-1}}\,,$$

$$\begin{aligned} D_{m+k}^2 := &\left(|A_m|_{r_{m-1}}\, C_F^2 + \frac{2\,|\beta_{m-1}|\,|A_m|_{m-1,m-1}\,\eta\,|\sigma|\,\alpha_m^s(m)}{C_1(1-\theta)\omega_m^{s+s_L}}\right)\eta\frac{|\lambda_m|}{|\mu_m|}\theta^k \\ &+ \frac{2\eta\,|\sigma|\,\alpha_m^s(m)}{C_1\omega_{m+k}^{s+s_L}}\left(\sum_{l=0}^{k}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L} + \frac{\theta}{1-\theta}\right), \quad \forall\ 0\leq k<M, \end{aligned}$$

$$\begin{aligned} D_{m+M}^2 := &\left(|A_m|_{r_{m-1}}\, C_F^2 + \frac{2\,|\beta_{m-1}|\,|A_m|_{m-1,m-1}\,\eta\,|\sigma|\,\alpha_m^s(m)}{C_1(1-\theta)\omega_m^{s+s_L}}\right)\eta\frac{|\lambda_m|}{|\mu_m|}\theta^M \\ &+ \frac{2\eta\,|\sigma|\,\alpha_m^s(m)}{C_1\omega_{m+M}^{s+s_L}}\left(\chi + \frac{\theta}{1-\theta}\right), \end{aligned}$$

and

$$D_{m+k}^2 := D_{m+M}^2\frac{\omega_{m+M}^s}{\omega_{m+k}^s}, \quad \forall\ k>M.$$

Then

$$|A|\, C^2 \leq D^2.$$

Finally we can sum up all the computations of this Sub-subsection and state the following result:

Proposition 6.3.18. *Let M be an integer satisfying* (6.32) *and* (6.36). *We define D^1 (resp. D^2) as in Proposition 6.3.16 (resp. Proposition 6.3.17) and let*

$$Z^2(r) := D^1(\bar{x})r + D^2 r^2.$$

Then for all $y, z \in B_0(r)$

$$A\left(Df\left(\bar{x}+y\right) - A^\dagger\right)z \leq Z^2(r).$$

Putting this together with Proposition 6.3.11, we end up with the following result:

Proposition 6.3.19. *Let M be an integer satisfying* (6.32) *and* (6.36). *Let*

$$Z(r) := Z^1(r) + Z^2(r).$$

Then for all $y, z \in B_0(r)$,

$$\left|DT\left(\bar{x}+y\right)z\right| \leq Z(r).$$

6.3.4 The radii polynomials and interval arithmetics

All the work done up to now in Sections 6.2 and 6.3 can be summarized in the following statement:

Theorem 6.3.20. *Let $s > 1$, and $s_L > 0$. Assume that f is a map from Ω^s to Ω^{s-s_L} of the form $f = \mathcal{L} + N$, where $\mathcal{L}$ is a tridiagonal operator satisfying* (6.3), (6.4) *and* (6.5)*, and where the nonlinear part N is quadratic. Assume that for some $m \geq 6$ we have computed an approximate zero of f, of the form $\bar{x} = (\bar{x}_0, \ldots, \bar{x}_{m-1}, 0, \ldots, 0, \ldots)$, and D an approximate inverse of $Df^{(m)}(\bar{x})$. Consider*

$$T : \begin{cases} \Omega^s \to \Omega^s, \\ x \mapsto x - Af(x), \end{cases}$$

where A is defined as in (6.21)*. Take M satisfying* (6.32) *and* (6.36) *and $L \geq 0$ a computational parameter. Then the bound Y defined in Proposition 6.3.10 satisfies* (6.26) *and for all $r > 0$, the bound $Z(r)$ defined in Proposition 6.3.19 satisfies* (6.27)*.*

Now that we have found bounds Y and $Z(r)$ that satisfy (6.26) and (6.27), we must find a radius $r > 0$ such that $\|Y + Z(r)\|_s < r$ in order to apply Theorem 6.3.1. By definition of the norm $\|\cdot\|_s$, it amounts to find an $r > 0$ such that, for every $k \geq 0$, the *radii polynomial* $P_k(r)$ satisfies

$$P_k(r) := Y_k + Z_k(r) - \frac{r}{\omega_k^s} < 0.$$

Note that since we constructed Y and Z in such a way that for every $k \geq M$,

$$Y_{m+k} = Y_{m+M} \frac{\omega_{m+M}^s}{\omega_{m+k}^s} \quad \text{and} \quad Z_{m+k} = Z_{m+M} \frac{\omega_{m+M}^s}{\omega_{m+k}^s},$$

it is enough to find an $r > 0$ such that for all $0 \leq k \leq m + M$, $P_k(r) < 0$. In order to do so, we numerically compute, for each $0 \leq k \leq m + M$,

$$I_k := \{r > 0 \mid P_k(r) < 0\},$$

and

$$I := \bigcap_{k=0}^{m+M} I_k.$$

If I is empty, then the proof fails, and we should try again with some larger parameters m and M. If I is non empty, we pick an $r \in I$ and check rigorously, using the interval arithmetics package INTLAB [190], that for all $0 \leq k \leq m + M$, $P_k(r) < 0$, which according to Theorem 6.3.1, proves that T defined in (6.25) is a contraction on $B_s(\bar{x}, r)$, thus yielding the existence of a unique solution of $f(x) = 0$ in $B_s(\bar{x}, r)$.

6.4 An example of application

We present in this Section an example of equation, for which it is possible to apply the method developed in this paper. We first explain the link between the equation that we study (cf. (6.50) below) and the tridiagonal operator defined in Section 6.2. Then, we explain what are in this example the values of the various constants and parameters of our method.

Equations of the following form:

$$\begin{aligned} &-(2 + \cos\xi)u''(\xi) + u(\xi) = -\sigma u(\xi)^2 + g(\xi), \\ &u'(0) = u'(\pi) = 0, \end{aligned} \tag{6.50}$$

where g is a 2π-periodic even smooth function, fall into the framework developed in Section 6.2. Consider indeed the cosine Fourier expansions of u and g:

$$u(\xi) = \sum_{k \in \mathbb{Z}} x_k \cos(k\xi), \qquad g(\xi) = \sum_{k \in \mathbb{Z}} g_k \cos(k\xi).$$

Then, (6.50) can be rewritten as $f(x) = 0$, where

$$f_0(x) := x_0 + x_1 + \sigma\,(x * x)_0 - g_0,$$

and for all $k \geq 1$,

$$f_k(x) := \frac{1}{2}(k-1)^2 x_{k-1} + (1+2k^2)x_k + \frac{1}{2}(k+1)^2 x_{k+1} + \sigma\,(x * x)_k - g_k. \tag{6.51}$$

We see that the linear part of (6.51) is, as in (6.3), given by

$$\mathcal{L}_k(x) = \lambda_k x_{k-1} + \mu_k x_k + \beta_k x_{k+1},$$

with

$$\mu_0 := 1, \quad \beta_0 := 1,$$

and for all $k \geq 1$,

$$\lambda_k := \frac{1}{2}(k-1)^2, \quad \mu_k := (1+2k^2) \quad \text{and} \quad \beta_k := \frac{1}{2}(k+1)^2.$$

Let us fix some $m \geq 2$. With

$$C_1 = 2, \quad C_2 = 3 \quad \text{and} \quad \delta = \frac{1}{4}\frac{(m+1)^2}{m^2 + \frac{1}{2}},$$

we get

$$\forall\ k \geq 1, \quad \left|\frac{\lambda_k}{k^2}\right|, \left|\frac{\mu_k}{k^2}\right|, \left|\frac{\beta_k}{k^2}\right| \leq C_2,$$

together with

$$\forall\ k \geq m, \quad C_1 \leq \left|\frac{\mu_k}{k^2}\right| \quad \text{and} \quad \left|\frac{\lambda_k}{\mu_k}\right|, \left|\frac{\beta_k}{\mu_k}\right| \leq \delta.$$

We now focus on the example when

$$g(\xi) := \frac{1}{2} + 3\cos(\xi) + \frac{1}{2}\cos(2\xi),$$

so that $u(\xi) = \cos(\xi)$ is a trivial solution for $\sigma = 0$. We are going to use rigorous computations in order to prove the existence of solutions for $\sigma \neq 0$, and to compute these solutions.

Starting from $\sigma = 0$, we first use standard pseudo-arclength continuation techniques to numerically get some nontrivial approximate solutions for $\sigma \neq 0$. We computed 1250 different solutions (675 for $\sigma > 0$ and 675 for $\sigma < 0$). See Figure 6.2 for a diagram summing up those computations, where each point represents a solution of (6.50).

Then we use the rigorous computation method described in this paper to prove, for each numerical solution, the existence of a true solution in a small neighbourhood of the numerical approximation. We keep $m = 20$ Fourier coefficients for the numerical computation, and use $M = 20$ and the decay rate $s = 2$ for the proof. The bounds of Lemma 6.3.12 as well as the error on $\tilde{\omega}$ (6.35) are computed with $L = 100$. For each numerical solution, the proof is successful. The set I defined in Section 6.3.4 on which all radii polynomials should be negative always contains $[4 \times 10^{-11}, 10^{-4}]$, and we rigorously prove using interval arithmetics that they are indeed all negative for $r = 10^{-10}$. Hence the assumptions of Theorem 6.3.1 hold and as a consequence, within a ball of radius $r = 10^{-10}$ in Ω^s centered on the numerical approximation, there exists a unique solution to (6.50). Therefore the existence of the solutions represented in Figure 6.2 is rigorously proven, within a margin of error that is too small to be depicted. The codes used to perform the proofs can be found in [74].

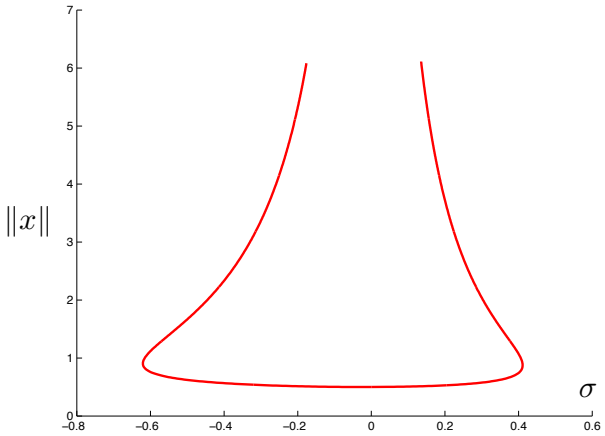


Figure 6.2 – Branch of solutions of (6.50).

Notice that existence of solutions of (6.50) could certainly have been obtained in different and more classical ways, for example using perturbative methods when σ is close to 0, or using a variational approach (that is, considering (6.50) as the Euler-Lagrange equation related to the critical points of a functional), or even using topological tools such as the Leray-Schauder theory. The advantage of our method is that it gives us more quantitative information than those approaches: indeed it enables to provide more than one solution for some values of σ, and, maybe more importantly, it gives a very precise localization of this (or these) solution(s) in terms of Fourier coefficients (something that looks very hard to obtain with qualitative PDEs methods).

6.5 Conclusion and Perspectives

A first interesting future direction of research would consist in adapting our approach to the rigorous computation connecting orbits of ODEs (using spectral methods). For instance, we would like to investigate the possibility of combining Hermite spectral methods with our approach to compute homoclinic orbits (e.g. see [148, 149]). Since the differential operator in frequency space of the Hermite functions is tridiagonal, adapting our method to this class of operator could lead to a new rigorous numerical method for connecting orbits.

It would also be interesting to adapt our method to the case of solutions belonging to the sequence space

$$\ell^1_\nu = \{x = (x_k)_{k\geq 0} : \|x\|_\nu := \sum_{k\geq 0} |x_k|\nu^k < \infty\}$$

for some $\nu \geq 1$. With this choice of Banach space, we could use the fact that ℓ^1_ν is naturally a Banach algebra under discrete convolutions. This could greatly simplify the nonlinear analysis.

Note that assumption (6.5) requires the tridiagonal operator to have symmetric ratios between the diagonal terms and the upper and lower diagonal terms. This is a restriction that could hopefully be relaxed. Since many interesting problems involve tridiagonal operators with non symmetric ratios (as in the case of differentiation in frequency space of the Hermite functions), we believe that this is a promising route to follow.

Finally, generalizing our approach to problems with block-tridiagonal structures could also be a valuable project.

Chapter 7

Efficient usage of the parameterization method to compute and validate local (un)stable manifolds

Abstract

This chapter is taken from [77]. We develop some automatic procedures for computing high order polynomial expansions of local (un)stable manifolds for equilibria of differential equations. Our method incorporates validated truncation error bounds, and maximizes the size of the image of the polynomial approximation relative to some specified constraints. More precisely we use that the manifold computations depend heavily on the scalings of the eigenvectors: indeed we study the precise effects of these scalings on the estimates which determine the validated error bounds. This relationship between the eigenvectors scalings and the error estimates plays a central role in our automatic procedures. In order to illustrate the utility of these methods we present several applications, including visualization of invariant manifolds in the Lorenz and FitzHugh-Nagumo systems and an automatic continuation scheme for (un)stable manifolds in a suspension bridge problem. In the present work we treat explicitly the case where the eigenvalues satisfy a certain non-resonance condition.

7.1 Introduction

Invariant sets are fundamental objects of study in dynamical systems theory. Sometimes we are interested in an invariant set which is a smooth manifold, and we seek a representation of a chart patch as the graph of a function or as the image of a chart map. Semi-numerical methods providing high order formal expansions of invariant manifolds have a long history in dynamical systems theory. We refer to the lecture notes of Simó [196], the historical remarks in Appendix B of the paper by Cabré, Fontich, and de la Llave [84], the manuscript of Haro [129], as well as the book by Meyer and Hall [164] for more complete discussion of this literature.

The present work is concerned with algorithms for computing local stable/unstable manifolds of equilibria solutions of differential equations, with validated error bounds. The methods employed here have some free computational parameters and we are especially interested in choosing these in an automatic way. We employ the parameterization method of [82, 83, 84] in our computations. This method provides powerful functional analytic tools for studying in-

variant manifolds. The core of the parameterization method is an invariance equation which conjugates a chart map for the local stable/unstable manifold to the linear dynamics given by the eigenvalues (see for example (7.3) in Section 7.2). Expanding the invariance equation as a formal series and matching like powers leads to homological equations for the coefficients of the series. These homological equations are solved to any desired order, yielding a finite approximation.

Given a finite approximate parameterization we would like to evaluate the associated truncation error. An important feature of the parameterization method is that there is a natural notion of a posteriori error, i.e. one can "plug" the approximate solution back into the invariance equation and measure the distance from zero in an appropriate norm. Further analysis is of course necessary in order to obtain validated error bounds, as small defects need not imply small truncation errors. When the invariance equation is formulated on a regular enough function space it is possible to apply a Newton-Kantorovich argument to get the desired bounds.

A uniqueness result for the parameterization method states that the power series coefficients are unique up to the choice of the scalings of the (un)stable eigenvectors [82]. This freedom in the choice of scaling can be exploited in order to control the numerical properties of the scheme. For example by increasing or decreasing the length of the eigenvectors it is possible to manipulate the decay rates of the power series coefficients, and thus influence the numerical stability of the scheme.

One of the main findings of the present work is that the bounds required in the Newton-Kantorovich argument (see the definition of the *radii polynomials* bounds in (7.19)) depend in an explicit way on the choice of the eigenvector scalings. This result leads to algorithms for optimizing the choice of eigenvectors scalings under some fixed constraints. The algorithms developed in the present work complement similar automatic schemes developed in [136] (for computer-assisted study of periodic orbits) and are especially valuable in continuation arguments, where one wants to compute the invariant manifolds over a large range of parameter values in an automatic way.

Remark 7.1.1. *The optimization constraints referred to above can be chosen in different ways depending on ones goals. For example when the goal of the computation is visualization of the manifold it is desirable to choose scalings which maximize the "extent" of the manifold in phase space (i.e. maximize the surface measure of the patch). On the other hand when the eigenvalues have different magnitudes then it may be desirable to maximize the image of the manifold under the constraint that the ratios of the scalings of the eigenvectors are fixed (this is especially useful in "fast-slow" systems). In other situations one might want to optimize some other quantity all together. Whatever constraints one chooses, we always want to optimize while holding the error of the computation below some specified tolerance. The main point of the present work is that whatever the desired constraints, the explicit dependency of the bounds on the scaling facilitates the design of algorithms which respect the specified error tolerance.*

Remark 7.1.2. *We fix the domain of our approximate parameterization to be the unit ball in* $\mathbb{C}^m$ *(where* m *is the number of (un)stable eigenvalues, i.e. the dimension of the manifold) and vary the scalings of the eigenvectors in order to optimize with respect to the constraints. Another (theoretically equivalent approach) would be to fix the scalings of the eigenvectors and vary the size of the domain. However the scalings of the eigenvectors determine the decay rates of the power series coefficients, and working with analytic functions of fast decay seems to stabilize the problem numerically.*

Remark 7.1.3. *In many previous applications of the parameterization method the free constants were selected by some "numerical experimentation." See for example the introduction and discussion in Section* 5 *of [60], Remark* 3.6 *of [65], Remark* 2.18 *and* 2.20 *of [94], the discussion of Example* 5.2 *in [171], Remark* 2.4 *of [172], and the discussion in Sections* 4.2 *and* 6 *of [172]. This motivates the need for systematic procedures developed here.*

Remark 7.1.4. *The algorithms developed here facilitate the computation of local stable/unstable manifolds. Once the local computations have been optimized one could extend or "grow" larger patches of the local manifold using adaptive integration/continuation techniques. This is a topic of substantial research and we refer the interested reader to the survey article [151]. See also the works of [150, 182, 183, 127, 212, 209, 210] and the references therein. Combining these integration/continuation algorithms with the methods of the present work could be an interesting topic for future research.*

Remark 7.1.5. *In the present work we employ a functional analytic style of a-posteriori analysis in conjunction with the parameterization method of [82, 83, 84]. Moreover the arguments are framed in classical weighted sequences spaces following the work of [152, 113]. There are in the literature many other methods for obtaining rigorous computer-assisted error bounds on numerical approximations of invariant manifolds. The interested reader should consult the works of [65, 172, 209, 210, 178, 89, 90, 87, 207, 203, 88, 198, 205, 49, 54, 204] for other approaches and results.*

Remark 7.1.6. *In recent years a number of authors have developed numerical methods based on the parameterization method in order to compute invariant manifolds of fixed and equilibrium points (e.g. see [171, 168, 169] for more discussion). The parameterization method can also be used to compute stable/unstable manifolds associated with periodic orbits of differential equations [94, 128, 134], as well as stable/unstable manifolds associated with invariant circles/tori [131, 135]. Indeed the parameterization method can be extended in order to compute the invariant tori themselves [115], leading to a KAM theory "without action angle coordinates". For more complete discussion of numerical methods based on the parameterization method we refer to the upcoming book [130]. For the moment we remark that the optimization algorithms developed in the present work could be adapted to these more general settings.*

Our paper is organized as follows. In Section 7.2 we present briefly the parameterization method and discuss its behaviour with respect to some specific changes of variable. In Section 7.3 we give a way to numerically compute an approximate parameterization and then address the issue of finding a rescaling that maximize the image of the parameterization, while verifying some *a posteriori* bounds that ensure (in some sense) the validity of the approximate parameterization. One possible way of proving the validity of the approximation is to use the ideas of rigorous computation, which we detail in Section 7.4. We conclude in Section 7.5 by presenting the results obtained with our method to compute maximal patches of local manifolds for several examples. The codes for all the examples can be found at [78].

7.2 The parameterization method

In this section, we introduce the parameterization method for the stable manifold of an equilibrium solution of a vector field. The unstable manifold is obtained by time reversal.

7.2.1 Invariance equation for stable manifolds of equilibria of vector fields

We consider an ordinary differential equation (ODE) of the form

$$y' = g(y), \tag{7.1}$$

where $g : \mathbb{R}^n \to \mathbb{R}^n$ is analytic. Assume that $p \in \mathbb{R}^n$ is an equilibrium point, i.e. $g(p) = 0$, and assume that the dimension of the stable manifold at p is given by $n_s \leq n$. Denote (λ_k, V_k), $1 \leq k \leq n_s$ the stable eigenvalues (that is $\Re(\lambda_k) < 0$, for $k = 1, \ldots, n_s$) together with associated eigenvectors, and denote $\Lambda = diag(\lambda_1, \ldots, \lambda_{n_s})$.

We want to find an analytic parameterization of the local stable manifold at p. So we look for a power series representation

$$f(\theta) = \sum_{|\alpha| \geq 0} a_\alpha \theta^\alpha, \quad \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_{n_s} \end{pmatrix} \in \mathbb{R}^{n_s}, \; a_\alpha = \begin{pmatrix} a_\alpha^{(1)} \\ \vdots \\ a_\alpha^{(n)} \end{pmatrix} \in \mathbb{R}^n, \tag{7.2}$$

with the classical multi-indexes notations $|\alpha| = \alpha_1 + \ldots + \alpha_{n_s}$ and $\theta^\alpha = \theta_1^{\alpha_1} \ldots \theta_{n_s}^{\alpha_{n_s}}$, and assume that the parameterization f conjugates the flow φ induced by g to the linear flow induced by Λ, that is

$$f\left(e^{\Lambda t}\theta\right) = \varphi(t, f(\theta)).$$

Differentiating with respect to t and taking $t = 0$, we get that f satisfies the invariance equation

$$Df(\theta)\Lambda\theta = g(f(\theta)), \tag{7.3}$$

and to get a well-posed problem we add the following constraints

$$f(0) = p, \quad Df(0) = \begin{pmatrix} V_1 & \ldots & V_{n_s} \end{pmatrix}. \tag{7.4}$$

Endow $\mathbb{C}^{n_s}$ with norm $\|\theta\|_{\mathbb{C}^{n_s}} = \max\{|\theta_k| : k = 1, \ldots, n_s\}$, where $|\cdot|$ denotes the complex modulus, and using that norm, denote by $B_\nu \subset \mathbb{C}^{n_s}$ the closed ball of radius ν centered at 0. We look for a parameterization f which is analytic on a ball $B_\nu \subset \mathbb{C}^{n_s}$ with $\nu > 0$. We call the image $f[B_\nu]$ a *patch* of the local invariant manifold.

Remark 7.2.1. *If some of the eigenvalues happen to be complex-conjugate, say $\overline{\lambda_1} = \lambda_2, \ldots, \overline{\lambda_{2m-1}} = \lambda_{2m}$, it is easier to consider a power series f with complex coefficients (i.e. with $a_\alpha \in \mathbb{C}^n$) and acting on $\theta \in \mathbb{C}^{n_s}$. We can then recover the real parameterization by considering, for $\theta \in \mathbb{R}^{n_s}$,*

$$f_{real}(\theta_1, \ldots, \theta_{n_s}) = f(\theta_1 + i\theta_2, \theta_1 - i\theta_2, \ldots, \theta_{2m-1} + i\theta_{2m}, \theta_{2m-1} - i\theta_{2m}, \theta_{2m+1}, \ldots, \theta_{n_s}).$$

See [155] for a more detailed explanation of this fact. To be general in the sequel of our presentation, we will assume that f is a complex power series.

We say that there is a resonance of order α between the stable eigenvalues if

$$\alpha_1\lambda_1 + \ldots + \alpha_{n_s}\lambda_{n_s} = \lambda_j \tag{7.5}$$

for some $1 \leq j \leq n_s$. If there is no resonance for any $\alpha \in \mathbb{N}^{n_s}$ then we say that the stable eigenvalues are non-resonant. Note that if $|\alpha|$ is large enough then a resonance is impossible.

It is shown in [82] that if g is analytic then (7.3) has an analytic solution f as long as the eigenvalues are non-resonant. Moreover the power series coefficients of f are uniquely determined up to the choice of the scalings of the eigenvectors. This abstract result does not however provide explicit bounds on the size of the domain of analyticity B_ν for the parameterization: hence the need for a-posteriori validation of our numerical computations. We also note that if there is a resonance then the invariance equation can be modified so that we conjugate to a polynomial (instead of linear) dynamical system [82, 66], and that the later work just cited implements computer-assisted error bounds for the resonant case using the radii polynomial approach. Adapting the methods of the present work to the resonant case will be the topic of a future study. It is clear from the work of [82] that in the resonant case the Taylor coefficients of the parameterization are unique up to the choice of the eigenvector scalings. What remains to be checked is that in the resonant case the eigenvector scalings appear in the radii polynomials in an explicit way (as is the case in for non-resonant eigenvalues, see Section 7.4).

7.2.2 Change of coordinates

Assume that f is a power series of the form (7.2) satisfying (7.3) and (7.4) (therefore it is a local parameterization of the stable manifold at p). Now consider a change of coordinates in $\mathbb{C}^{n_s}$, defined by some invertible matrix $\Gamma \in M_{n_s}(\mathbb{C})$, and the new power series

$$\tilde{f}(\theta) = f(\Gamma\theta).$$

On one side we have

$$D\tilde{f}(\theta)\Lambda\theta = Df(\Gamma\theta)\Gamma\Lambda\theta,$$

and on the other side, thanks to (7.3) we get

$$g(\tilde{f}(\theta)) = g(f(\Gamma\theta)) = Df(\Gamma\theta)\Lambda\Gamma\theta.$$

Therefore, if Γ is such that $\Gamma\Lambda = \Lambda\Gamma$, $\tilde{f}$ also satisfies the invariance equation (7.3), together with the slightly modified conditions

$$\tilde{f}(0) = p, \quad D\tilde{f}(0) = \Gamma \begin{pmatrix} V_1 & \dots & V_{n_s} \end{pmatrix}. \tag{7.6}$$

Remark 7.2.2. *From now on we assume that* $\Gamma = \operatorname{diag}(\gamma_1, \dots, \gamma_{n_s})$*, which is sufficient to have* $\Gamma\Lambda = \Lambda\Gamma$ *(it is also necessary if the* λ_k *are pairwise distinct). We also assume that the* γ_i *are all real positive numbers and that coefficients* γ_i *corresponding to two complex conjugates eigenvalues are equal. Taking* γ_i *real is natural if all the eigenvalues are real (and* f *is therefore a real power series). On the other hand if there are some complex-conjugate eigenvalues, say* $\lambda_1 = \overline{\lambda_2}$*, then the recovery of a real parameterization as explained in Remark 7.2.1 uses the fact that the corresponding eigenvectors* V_1 *and* V_2 *also are complex-conjugate, and that this property is propagated to all the coefficients of the parameterization when recursively solving the invariance equation* (7.8)*. By taking* γ_1 *and* γ_2 *real and equal, we ensure that this property is conserved after the rescaling (namely* $\gamma_1 V_1 = \overline{\gamma_2 V_2}$*), so that we can still easily recover a real parameterization. Admittedly, we could relax this hypothesis and only assume that* γ_1 *and* γ_2 *themselves are complex-conjugate, but we will not consider this possibility here.*

As announced, we now consider $\Gamma = \operatorname{diag}(\gamma_1, \dots, \gamma_{n_s})$, where $\gamma_i \in \mathbb{R}_+$ for all $1 \leq i \leq n_s$, and $\tilde{f}$ defined as $\tilde{f}(\theta) = f(\Gamma\theta)$. The above discussion shows that $\tilde{f}$ is a new parameterization of the local manifold, since it satisfies (7.3) and (7.6). Besides, the Taylor expansion of $\tilde{f}$ can be easily expressed in terms of the Taylor expansion of f. Indeed if we write $\tilde{f}$ as

$$\tilde{f}(\theta) = \sum_{|\alpha| \geq 0} \tilde{a}_\alpha \theta^\alpha,$$

then the coefficients are given by

$$\tilde{a}_\alpha = a_\alpha \gamma^\alpha, \tag{7.7}$$

where $\gamma = (\gamma_1, \dots, \gamma_{n_s})$ and again standard multi-indexes notations. Therefore it is enough to find one parameterization f of the local manifold (or more precisely its coefficients a_α) to get all the re-parameterizations $\tilde{f}$ (at least those given by a diagonal matrix Γ) without further work. Let us introduce an operator acting on sequences to express this rescaling in a condensed way.

Definition 7.2.3. *Given* $\gamma = (\gamma_1, \dots, \gamma_{n_s})$*, we define* $\mathcal{L}$ *(acting on* a*) component-wise by*

$$\mathcal{L}_\alpha(a) = \gamma^\alpha a_\alpha, \quad \forall |\alpha| \geq 0.$$

Therefore, if a is the sequence of coefficients of the parameterization f, the sequence of coefficients of the parameterization $\tilde{f}$ defined as above is given by $\mathcal{L}(a)$.

7.3 How to compute f and maximize the local manifold patch

In this section we present a method to compute numerically a parameterization of the manifold (that is the coefficients (a_α)) and then choose a proper rescaling γ to maximize the corresponding image. We assume in the sequel that the nonlinearities in g are polynomials. Note that this not so restrictive as it might first seems, as techniques of automatic differentiation can be used in order to efficiently compute the (Taylor/Fourier/Chebyshev) series expansions of compositions with elementary functions. The authors first learned these techniques from Chapter 4.7 of [145], but the interested reader should also refer to the discussion and references in [129, 139].

Automatic differentiation is also a valuable tool for validated numerics, as polynomial nonlinearities are often more convenient to work with than transcendental ones. Since elementary functions of mathematical physics (powers, exponential, trigonometric functions, rational, Bessel, elliptic integrals, etc.) are themselves solutions of ODEs, these ODEs can be appended to the original problem of interest in order to obtain a new problem with only polynomial nonlinearities (but with more variables and more equations). Moreover, in many computer-assisted proofs it is the dimension of the underlying invariant object, and not the dimension of the embedding space, that informs the difficulty of the problem. We refer for example the book of [201] for a much more complete discussion of these matters. We also mention that automatic differentiation has been combined with the radii polynomial approach in [154] in order to compute periodic orbits of some celestial mechanics applications.

Of course automatic differentiation is not the only method which can be used in order to replace a transcendental vector field with a polynomial one. Any method of polynomials approximation can be used. A detailed survey of the interpolation literature is far beyond the scope of the present work, however we mention the works of [52, 108] where one can find implementation details and fuller discussion of the literature surrounding the use of Chebyshev polynomials to expand transcendental nonlinearities and obtain computer assisted error bounds. We also note that general purpose software exists for carrying out these kinds of manipulations, even with mathematically rigorous error bounds [69, 208, 147].

7.3.1 Computation of the approximate parameterization

Let f be a power series as in (7.2), assume g is a polynomial vector field of degree d given by

$$g(y) = \sum_{|\beta|\leq d} b_\beta y^\beta, \qquad b_\beta \in \mathbb{R}^n$$

and plug it into the invariance equation (7.3). We obtain

$$\sum_{|\alpha|\geq 0} \left(\alpha_1\lambda_1 + \ldots + \alpha_{n_s}\lambda_{n_s}\right) a_\alpha \theta^\alpha = \sum_{|\beta|\leq d} b_\beta \left(\sum_{|\alpha|\geq 0} a_\alpha \theta^\alpha\right)^\beta = \sum_{|\alpha|\geq 0}\sum_{|\beta|\leq d} b_\beta \left(a^\beta\right)_\alpha \theta^\alpha, \tag{7.8}$$

where again we use multi-indexes notations, $a^\beta = \left(a^{(1)}\right)^{\beta_1} * \ldots * \left(a^{(n)}\right)^{\beta_n}$, and $*$ denotes the Cauchy product. Notice that the two conditions in (7.4) imply that the coefficients of order 0 and 1 are the same on both sides of (7.8). There are several ways to obtain an approximation of the coefficients $(a_\alpha)_{|\alpha|\geq 2}$ so that (7.8) is satisfied, one of them being to compute them recursively for increasing $|\alpha|$. Here we present another method, which fits naturally with the ideas of rigorous computations exposed later in the paper. We define the infinite dimensional vector $a = (a_\alpha)_{|\alpha|\geq 0}$

and the operator F, acting on a component-wise by

$$F_\alpha(a) = \begin{cases} a_0 - p, & \text{if } \alpha = 0, \\ a_{e_i} - V_i, & \text{if } \alpha = e_i,\ \forall\ 1 \le i \le n_s \\ (\alpha_1\lambda_1 + \ldots + \alpha_{n_s}\lambda_{n_s})\, a_\alpha - \sum_{|\beta|\le d} b_\beta \left(a^\beta\right)_\alpha, & \forall\ |\alpha| \ge 2. \end{cases}$$

Finding a solving (7.8) and the additional conditions (7.4) is equivalent to solve

$$F(a) = \{F_\alpha(a)\}_{|\alpha|\ge 0} = 0. \tag{7.9}$$

Given $\gamma = (\gamma_1, \ldots, \gamma_{n_s})$, finding a rescaled parameterization (that is solving (7.3) and (7.6)) can also be expressed as finding the zero of the function $\tilde{F}$, which is defined the same way as F except for the indices $|\alpha| = 1$:

$$\tilde{F}_\alpha(a) = \begin{cases} a_0 - p, & \text{if } \alpha = 0, \\ a_{e_i} - \gamma_i V_i, & \text{if } \alpha = e_i,\ \forall\ 1 \le i \le n_s \\ (\alpha_1\lambda_1 + \ldots + \alpha_{n_s}\lambda_{n_s})\, a_\alpha - \sum_{|\beta|\le d} b_\beta \left(a^\beta\right)_\alpha, & \forall\ |\alpha| \ge 2. \end{cases} \tag{7.10}$$

Notice that the discussion in Section 7.2.2 shows that $F(a) = 0$ if and only if $\tilde{F}(\mathcal{L}(a)) = 0$.

Remark 7.3.1. *Since a_0 and the a_{e_i} are fixed by the additional conditions (7.4), we could also consider them as parameters and define F as $(F_\alpha)_{|\alpha|\ge 2}$, acting only on $(a_\alpha)_{|\alpha|\ge 2}$. We do this for the examples of Sections 7.5.1 and 7.5.2, but we keep the above definition of F and $\tilde{F}$ when we use rigorous computation (Section 7.4 and example in Section 7.5.3), because it allows for a simpler presentation.*

Now we fix an integer N and define the truncated operator $F^{[N]} = (F_\alpha)_{|\alpha|<N}$, acting on a truncated sequence $a^{[N]} = (a_\alpha)_{|\alpha|<N}$, by

$$F_\alpha(a^{[N]}) = \begin{cases} a_0 - p, & \text{if } \alpha = 0, \\ a_{e_i} - V_i, & \text{if } \alpha = e_i,\ \forall\ 1 \le i \le n_s \\ (\alpha_1\lambda_1 + \ldots + \alpha_{n_s}\lambda_{n_s})\, a_\alpha - \sum_{|\beta|\le d} b_\beta \left(a^\beta\right)_\alpha, & \forall\ 2 \le |\alpha| < N. \end{cases}$$

Since the problem is now finite dimensional, we can use Newton's method to compute an approximate zero of $F^{[N]}$. In the rest of this paper, $\bar{a}$ will denote such an approximate solution completed with 0 for $|\alpha| \ge N$. See Section 7.5 for explicit examples. Also note that the only property that really matters concerning the approximate parameterization $\bar{a}$ is that $\bar{a}_\alpha = 0$ for all $|\alpha| \ge N$. As long as it satisfies this property, everything in the sequel will work, even if $\bar{a}$ was obtained in a different fashion than the one we just presented (for instance by solving inductively a finite number of homological equations).

Remark 7.3.2. *Taking N larger leads to a better approximation but at the expense of computational cost, so its choice depends on how precise an approximation you need, and how much computational resources you have.*

7.3.2 Maximizing the image of the parameterization

Now that we have an approximate parameterization, we focus on maximizing the image of the corresponding manifold, while checking that our approximation is still valid. The power series f given by (7.2) is now considered as

$$f : B_\nu \to \mathbb{C}^n$$

for some $\nu > 0$. One approach in getting the largest possible image of f would be to maximize the ν for which (7.3) is *valid* on B_ν. We give in Definition 7.3.6 and Definition 7.3.9 two different definitions of parameterization validity.

Remark 7.3.3. *For reasons of numerical stability, we always consider the parameter space B_ν for $\nu = 1$ and instead use the γ introduced in the reparameterization of Section 7.2.2 as a parameter. Indeed, assume that the parameterization f is valid on B_{ν_1} for some ν_1, then proving that it still is on B_{ν_2} for a different ν_2 is equivalent to prove that $\tilde{f}(\theta) = f(\Gamma\theta) = f(\gamma_1\theta_1, \ldots, \gamma_{n_s}\theta_{n_s})$ is valid on B_{ν_1}, with $\gamma_k = \frac{\nu_2}{\nu_1}$ for all k. So we can always keep $\nu = 1$ and rather try to maximize the γ_k for which $\tilde{f}$ is valid on B_1.*

Based on the previous remark, for the applications presented in this paper we always fix $\nu = 1$. However, in the sequel we keep a parameter $\nu > 0$ in the theoretical estimates, so that they can be reused without modifications even in situations where ν has to be chosen different from 1.

Remark 7.3.4. *If the eigenvalues are real and not all equal to the same value, it may be useful to consider different scalings for each direction, that is to take $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_{n_s})$ with different γ_k rather than $\Gamma = \mathrm{diag}(\gamma, \ldots, \gamma)$. Indeed in this work we aim at maximizing the surface of the manifold patch, but for some specific problem (a fast-slow system for instance), you may rather want to enlarge the manifold in one precise direction, in which case you should definitely consider different γ_k for each k.*

In this paper we will use two different criteria to say that our parameterization is valid on B_ν. The first one is a numerical a posteriori estimate and the second is a rigorous validation. In order to measure the validity of a parameterization, we need to compute the norm of a sequence $a = \{a_\alpha\}_{|\alpha|\geq 0}$ with $a_\alpha \in \mathbb{C}^n$. For this, let us introduce (for some $\nu > 0$) the space

$$\ell^1_\nu := \left\{ u = \{u_\alpha\}_{|\alpha|\geq 0} \mid u_\alpha \in \mathbb{C} \text{ and } \|u\|_{\ell^1_\nu} := \sum_{|\alpha|\geq 0} |u_\alpha| \nu^{|\alpha|} < \infty \right\}.$$

Given $a = (a_\alpha)_{|\alpha|\geq 0}$, with $a_\alpha = \begin{pmatrix} a^{(1)}_\alpha \\ \vdots \\ a^{(n)}_\alpha \end{pmatrix} \in \mathbb{C}^n$, denote $a^{(i)} = \left(a^{(i)}_\alpha\right)_{|\alpha|\geq 0}$. Then, consider the product space

$$X := \left(\ell^1_\nu\right)^n := \left\{ a = (a_\alpha)_{|\alpha|\geq 0} \mid \|a\|_X := \max_{1\leq i\leq n} \left\|a^{(i)}\right\|_{\ell^1_\nu} < \infty \right\}.$$

Remark 7.3.5. *It will be usefull to represent linear operators acting on elements of X with (infinite) matrix/vector notations. To prevent any future ambiguity, let us precise the ordering we use in this paper for those vectors and matrices. Given $a \in X$, we represent it as the (infinite) vector $(a_\alpha)_{|\alpha|\geq 0}$ where the a_α are ordered by growing $|\alpha|$, and by lexicographical order within the coefficients with same $|\alpha|$. For instance, if $n_s = 2$,*

$$a = (a_{0,0}, a_{1,0}, a_{0,1}, a_{2,0}, a_{1,1}, a_{0,2}, \ldots)^T.$$

Notice that each a_α is himself a vector of $\mathbb{R}^n$. For an (infinite) matrix $M = (M_{\alpha,\beta})_{|\alpha|,|\beta|\geq 0}$ representing a linear operator on X, we use the same order for the rows and columns. Notice that each coefficient $M_{\alpha,\beta}$ is in fact a n by n matrix whose coefficient on row i and column j will be denoted as $M^{(i,j)}_{\alpha,\beta}$, so that

$$(Ma)^{(i)}_\alpha = \sum_{|\beta|\geq 0} \sum_{j=1}^{n} M^{(i,j)}_{\alpha,\beta} a^{(j)}_\beta.$$

We now give the two announced criteria to measure the validity of a parameterization.

Definition 7.3.6. *Fix a* defect *threshold $\varepsilon_{max} > 0$, a truncation dimension N and an approximate solution $\bar{a}^{[N]}$ computed using the method of Section 7.3.1. Denote $\bar{a} = \bar{a}^{[N]}$. We say that*

$$f(\theta) := \sum_{|\alpha|<N} \bar{a}_\alpha \theta^\alpha \tag{7.11}$$

is defect-valid *on B_ν if*

$$\|F(\bar{a})\|_X < \varepsilon_{max}. \tag{7.12}$$

Equivalently, we say that $\bar{a}$ is defect-valid *on B_ν if* (7.12) *holds. Given $\gamma = (\gamma_1, \ldots, \gamma_{n_s})$, we also say that the rescaled parameterization $\mathcal{L}(\bar{a})$ is* defect-valid *on B_ν if*

$$\|\tilde{F}(\mathcal{L}(\bar{a}))\|_X < \varepsilon_{max}. \tag{7.13}$$

Remember that g is assumed to be polynomial, and so F is also polynomial, say of degree d. Since $\bar{a}_\alpha = 0$ for $|\alpha| \geq N$, then $F_\alpha(\bar{a}) = 0$ for all $|\alpha| \geq d(N-1)+1$. Thus the quantity $\|F(\bar{a})\|_X$ in (7.12) is only a finite sum and can be computed explicitly.

Assume now that we have computed all the $F_\alpha(\bar{a})$ for $|\alpha| \leq d(N-1)$ (which can be quite long because of the Cauchy products coming from the nonlinearities). When we then consider some $\gamma = (\gamma_1, \ldots, \gamma_{n_s})$ and the rescaled parameterization $\mathcal{L}(\bar{a})$, we get (using the fact that the nonlinearities are polynomial and the definition of the Cauchy product) that for all $|\alpha| \geq 0$,

$$\tilde{F}_\alpha(\mathcal{L}(\bar{a})) = \gamma^\alpha F_\alpha(\bar{a}). \tag{7.14}$$

This way, the evaluation of $\|\tilde{F}(\mathcal{L}(\bar{a}))\|_X$ for any rescaling is computationally cheap and thus it is rather straightforward to find the γ for which the re-parameterization $\mathcal{L}(\bar{a})$ gives the largest image of the manifold, while being defect-valid. Let us be a little more precise about this. Depending on our goal we use two different approaches.

Method 1: We look for eigenvector scalings which maximize the surface measure, subject to the restriction that the rescaled parameterization is defect-valid. Therefore we find numerically a mesh of the compact set

$$\left\{\gamma \in \mathbb{R}_+^{n_s} \mid \|\tilde{F}(\mathcal{L}(\bar{a}))\|_X = \varepsilon_{max}\right\}$$

and then approximately compute the surface area of the image for each point of the mesh. We refer to Sections 7.5.1 and 7.5.2 for explicit examples in dimension 2.

Method 2: We want to emphasize some specific directions when computing the manifold. Therefore we fix some weights $\omega_1, \ldots, \omega_{n_s}$ and consider only rescalings of the form

$$\gamma = \gamma(t) = (t\omega_1, \ldots, t\omega_{n_s}).$$

We then look for the largest t such that the rescaled parameterization is defect-valid. By doing so we obtain a manifold that stretches more in the directions with the largest weights. We refer to Sections 7.5.1 and 7.5.2 for explicit examples in dimension 2 where we stretch the manifolds in the slow direction.

Remark 7.3.7. *When there is only one stable/unstable eigenvalue (or a single pair of complex conjugate eigenvalues) then Method 2 reduces to choosing the largest possible scaling for the eigenvector (or for the complex conjugate pair of eigenvectors) so that the rescaled parameterization is defect-valid.*

Now we would like to present a different definition of validity of a parameterization, inspired by the field of rigorous computing. For this, we briefly review the ideas of rigorous computation. The idea is to reformulate the problem $F(a) = 0$ given in (7.9) and to look for a fixed point of a Newton-like equation of the form

$$T(a) = a - AF(a)$$

where A is an approximate inverse of $DF(\bar{a})$, and $\bar{a}$ is a numerical approximation obtained by computing a finite dimensional projection of F (in our case we called it $F^{[N]}$). Let us explain how we construct A. Remembering that

$$F_\alpha(a) = (\alpha_1\lambda_1 + \ldots + \alpha_{n_s}\lambda_{n_s})\, a_\alpha - \sum_{|\beta|\leq d} b_\beta \left(a^\beta\right)_\alpha, \quad \forall\ |\alpha| \geq 2$$

we consider the following approximation for $DF(\bar{a})$

$$A^\dagger = \begin{pmatrix} DF^{[N]}(\bar{a}) & & 0 & \\ & A_N^\dagger & & \\ 0 & & A_{N+1}^\dagger & \\ & & & \ddots \end{pmatrix}$$

where for each $k \geq N$, $A_k^\dagger$ is a finite bloc diagonal matrix, each of its diagonal block being of size n and of the form $(\alpha_1\lambda_1 + \ldots + \alpha_{n_s}\lambda_{n_s})\, I_n$, where $|\alpha| = k$ and I_n is the n by n identity matrix. In other words (see Remark 7.3.5)

$$A_k^\dagger\,(a_\alpha)_{|\alpha|=k} = ((\alpha_1\lambda_1 + \ldots + \alpha_{n_s}\lambda_{n_s})\, a_\alpha)_{|\alpha|=k}\,.$$

We then define an approximate inverse A of $DF(\bar{a})$ as

$$A := \begin{pmatrix} A^{[N]} & & 0 & \\ & A_N & & \\ 0 & & A_{N+1} & \\ & & & \ddots \end{pmatrix}, \tag{7.15}$$

where $A^{[N]}$ is a numerical approximation of $DF^{[N]}(\bar{a})^{-1}$ while the $A_k := \left(A_k^\dagger\right)^{-1}$ are the exact inverses. We then prove the existence of a zero of F by using a contraction argument yielding the existence of a fixed point of T. A precise theorem is stated below, but just before that we need (given $\gamma = (\gamma_1, \ldots, \gamma_{n_s})$) to define a *rescaled* operator

$$\tilde{T} := I - \tilde{A}\tilde{F} \tag{7.16}$$

that we can use in a similar fashion to prove the existence of a zero of $\tilde{F}$. Remembering that

$$\tilde{F}(\mathcal{L}(a)) = \mathcal{L}F(a)$$

we have

$$D\tilde{F}(\mathcal{L}(a)) = \mathcal{L}DF(a)\mathcal{L}^{-1}$$

and therefore we consider

$$\tilde{A}^\dagger := \mathcal{L}A^\dagger\mathcal{L}^{-1} \quad \text{and} \quad \tilde{A} := \mathcal{L}A\mathcal{L}^{-1} \tag{7.17}$$

as approximations for $D\tilde{F}(\mathcal{L}(\bar{a}))$ and $\left(D\tilde{F}(\mathcal{L}(\bar{a}))\right)^{-1}$ respectively.

The rigorous enclosure of a solution follows by verifying the hypothesis of the following Newton-Kantorovich type argument. Our method, often called the *radii polynomial approach*, was originally developed to study equilibria of PDEs [109] and was strongly influenced by the work of Yamamoto [214]. The differences between the radii polynomial approach and the standard Newton-Kantorovich approach are mainly twofold. First, the map $\tilde{F}$ under study is not required to map the Banach space X into itself. This is often the case when the map $\tilde{F}$ comes from a differential equation and results in a loss of regularity of the function it maps. Second, the approach does not require controlling the exact inverse of the derivative, but rather only an approximate inverse. This can be advantageous as controlling exact inverses of infinite dimensional linear operator can be challenging. For more details on the radii polynomial approach for rigorous computations of stable and unstable manifolds of equilibria, we refer to [66]. Given $r > 0$, denote by $B_r(a) \subset X = \left(\ell^1_\nu\right)^n$ the ball centered at $a \in X$ of radius r.

Theorem 7.3.8. *Let $\gamma = (\gamma_1, \dots, \gamma_{n_s}) \in \mathbb{R}^{n_s}_+$. Assume that the linear operator A in* (7.15) *is injective. For each $i = 1, \dots, n$, assume the existence of bounds $\tilde{Y} = \left(\tilde{Y}^{(1)}, \dots, \tilde{Y}^{(n)}\right)$ and $\tilde{Z}(r) = \left(\tilde{Z}^{(1)}(r), \dots, \tilde{Z}^{(n)}(r)\right)$ such that*

$$\left\| \left(\tilde{T}(\mathcal{L}(\bar{a})) - \mathcal{L}(\bar{a}) \right)^{(i)} \right\|_{\ell^1_\nu} \leq \tilde{Y}^{(i)} \quad \text{and} \quad \sup_{b,c \in B_r(0)} \left\| \left(D\tilde{T}(\mathcal{L}(\bar{a}) + b)c \right)^{(i)} \right\|_{\ell^1_\nu} \leq \tilde{Z}^{(i)}(r). \tag{7.18}$$

If there exists $r > 0$ such that

$$\tilde{P}^{(i)}(r) := \tilde{Y}^{(i)} + \tilde{Z}^{(i)}(r) - r < 0, \quad \text{for all } i = 1, \dots, n \tag{7.19}$$

then $\tilde{T} : B_r(\mathcal{L}(\bar{a})) \to B_r(\mathcal{L}(\bar{a}))$ is a contraction. By the contraction mapping theorem, there exists a unique $a^* \in B_r(\mathcal{L}(\bar{a})) \subset X$ such that $\tilde{F}(a^*) = 0$. Moreover, $\|a^* - \mathcal{L}(\bar{a})\|_X \leq r$.

As we see in Section 7.4, the bounds $\tilde{P}^{(1)}(r), \dots, \tilde{P}^{(n)}(r)$ given in (7.19) can be constructed as polynomials in r and are called the *radii polynomials*.

The statement of Theorem 7.3.8 is now used to define our second definition of validity of a parameterization, which is of course more costly than the first one but provides rigorous bounds.

Definition 7.3.9. *Fix a* proof *threshold r_{max}, a truncation dimension N and an approximate solution $\bar{a}$. Given a numerical zero $\bar{a}$ of F and $\gamma = (\gamma_1, \dots, \gamma_{n_s})$, we say that the parameterization $\mathcal{L}(\bar{a})$ is* proof-valid *on B_ν if there exists $r > 0$ such that condition* (7.19) *holds for some $r \leq r_{max}$.*

In the next section we explain how the bounds $\tilde{Y}$ and $\tilde{Z}$ can be constructed so that they depend explicitly on the scaling γ. Then, as for Definition 7.3.6, you only need to do the costly computations once for $\bar{a}$ (that is for $\gamma = (1, \dots, 1)$) and then the new bounds (and thus the new radii polynomials $\tilde{P}^{(i)}$) can be computed easily for any rescaling. Therefore the process of finding the rescaling γ which maximizes the image of a manifold given by a proof-valid parameterization is also rather straightforward. We give in Section 7.5.3 an example of application where we explicitly compute the bounds $\tilde{Y}$ and $\tilde{Z}$.

7.4 Explicit dependency of the radii polynomials in the scaling γ

In this section we construct the bounds $\tilde{Y}$ and $\tilde{Z}$ satisfying (7.18) with an explicit dependency on the γ whose action is given by (7.7).

7.4.1 The bound $\tilde{Y}$

Proposition 7.4.1. *The bound $\tilde{Y} = (\tilde{Y}^{(1)}, \dots, \tilde{Y}^{(n)})$ defined component-wise by*

$$\tilde{Y}^{(i)} = \left\| \left(\mathcal{L}AF(\bar{a})\right)^{(i)} \right\|_{\ell^1_\nu}, \quad \forall\ 1 \le i \le n,$$

satisfies (7.18)*.*

Proof. By definition of $\tilde{T}$,

$$\tilde{T}(\mathcal{L}(\bar{a})) - \mathcal{L}(\bar{a}) = \mathcal{L}AF(\bar{a})$$

and we have that

$$\tilde{A}\tilde{F}(\mathcal{L}(\bar{a})) = \mathcal{L}AF(\bar{a}),$$

which yields the formula for $\tilde{Y}$. □

Remark 7.4.2. *As previously mentioned, $F_\alpha(\bar{a}) = 0$ if $|\alpha| \ge d(N-1)+1$, and since A is of the form*

$$A = \left(\begin{array}{c|c} A^{[N]} & \\ \hline & Tail \end{array}\right),$$

where $Tail$ is a diagonal matrix (see (7.15)*), then $\tilde{Y}$ can be computed as a finite sum. Moreover, the $\tilde{Y}$ bound can be expensive to evaluate, since it requires computing the Cauchy products involved in $F(\bar{a})$, the matrix D which is the numerical inverse of the full and possibly large matrix $DF^{[N]}(\bar{a})$, and the product $AF(\bar{a})$. However, once $AF(\bar{a})$ is computed, we only need to do the component-wise multiplication defined by $\mathcal{L}$ and the finite sum corresponding to the ℓ^1_ν norm to get the bound $\tilde{Y}$ for any rescaling γ. Therefore, recomputing the bound $\tilde{Y}$ for a different rescaling is cheap.*

7.4.2 The bound $\tilde{Z}$

For the clarity of the exposition, we now assume that the nonlinearity of g (and thus of F) are of degree 2. We insist that the method presented here still holds for nonlinearity of higher degree (see for instance [60, 65]). It is also worth mentioning that in the context of computing equilibria of PDEs in [109, 121, 118] and periodic orbits of delay differential equations in [143], the bounds of the radii polynomials have been derived for general polynomial problems. Here, we decided that staying fully general would only obscure the point with notations, hence our restriction to quadratic nonlinearities.

To compute the $\tilde{Z}$ bound, we split $D\tilde{T}(\mathcal{L}(\bar{a}) + b)c$ as

$$\begin{aligned} D\tilde{T}(\mathcal{L}(\bar{a}) + b)c &= \left(I - \tilde{A}\tilde{A}^\dagger\right)c + \tilde{A}\left(D\tilde{F}(\mathcal{L}(\bar{a}) + b)\tilde{A}^\dagger\right)c \\ &= \left(I - \tilde{A}\tilde{A}^\dagger\right)c + \tilde{A}\left(D\tilde{F}(\mathcal{L}(\bar{a})) - \tilde{A}^\dagger\right)c + D^2\tilde{F}(\mathcal{L}(\bar{a}))(b,c) \end{aligned}$$

and we are going to bound each term separately.

The bound $\tilde{Z}_0$

We start this section with a result providing an explicit formula for the ℓ^1_ν operator norm of a matrix.

Lemma 7.4.3. *Let $\varrho_{n,n_s,N} = n\binom{N+n_s-1}{n_s}$ and $B \in M_{\varrho_{n,n_s,N}}(\mathbb{C})$. For all $c \in (\ell^1_\nu)^n$,*

$$\left\| \left(Bc^{[N]} \right)^{(i)} \right\|_{\ell^1_\nu} \leq \sum_{j=1}^{n} K_B^{(i,j)} \left\| c^{(j)} \right\|_{\ell^1_\nu},$$

where

$$K_B^{(i,j)} = \max_{0 \leq |\beta| < N} \left(\frac{1}{\nu^{|\beta|}} \sum_{0 \leq |\alpha| < N} \left| B^{(i,j)}_{\alpha,\beta} \right| \nu^{|\alpha|} \right), \quad \forall\ 1 \leq i,j \leq n. \tag{7.20}$$

The matrix/vector product should be understood according to Remark 7.3.5 with $\varrho_{n,n_s,N}$ simply being the length of $(c_\alpha)_{|\alpha|<N}$ seen as a vector of complex numbers. Lemma 7.4.3 is just the computation of the matrix norm associated to the weighted vector norm defined on ℓ^1_ν.

Proposition 7.4.4. *Let $B := I_{\frac{nN(N+1)}{2}} - A^{[N]}(DF^{[N]}(\bar{a}))$ and*

$$\tilde{B} := \mathcal{L}^{[N]} B \left(\mathcal{L}^{[N]} \right)^{-1}. \tag{7.21}$$

Let the bound $\tilde{Z}_0 = (\tilde{Z}_0^{(1)}, \ldots, \tilde{Z}_0^{(n)})$ defined component-wise by

$$\tilde{Z}_0^{(i)} := \sum_{j=1}^{n} K_{\tilde{B}}^{(i,j)}, \quad \forall\ 1 \leq i \leq n.$$

Then

$$\left\| \left(\left(I - \tilde{A}\tilde{A}^\dagger \right) c \right)^{(i)} \right\|_{\ell^1_\nu} \leq \tilde{Z}_0^{(i)}, \quad \forall\ 1 \leq i \leq n,$$

for all c such that $\|c\|_X \leq 1$.

Remark 7.4.5. *This bound can also be quite costly, because of the matrix-matrix multiplication required to get B. But again, once B has been computed, we only need to do the multiplication by the diagonal matrices associated do $\mathcal{L}^{[N]}$ and $\left(\mathcal{L}^{[N]}\right)^{-1}$ to get $\tilde{B}$ and then to compute the quantities $K_{\tilde{B}}^{(i,j)}$ to get the new bound for any rescaling.*

Proof. We start by noticing that

$$I - \tilde{A}\tilde{A}^\dagger = \mathcal{L}\left(I - AA^\dagger \right) \mathcal{L}^{-1}.$$

Then by definition of $A^\dagger$ and A, $\left(\left(I - AA^\dagger \right) c \right)_\alpha = 0$ for all $|\alpha| \geq N$ and we have

$$\left\| \left(\left(I - \tilde{A}\tilde{A}^\dagger \right) c \right)^{(i)} \right\|_{\ell^1_\nu} = \left\| \left(\mathcal{L}^{[N]} \left(I_{\frac{nN(N+1)}{2}} - D\left(DF^{[N]}(\bar{a}) \right) \right) \left(\mathcal{L}^{[N]} \right)^{-1} c^{[N]} \right)^{(i)} \right\|_{\ell^1_\nu},$$

and Lemma 7.4.3 yields the formula for $\tilde{Z}_0$. □

The bound $\tilde{Z}_1$

In this section we will need two additional results. The first one is a quantitative statement that ℓ^1_ν is a Banach algebra and allows us to bound the nonlinear terms.

Definition 7.4.6. *Let $u, v \in \ell^1_\nu$. We denote by $u * v$ the Cauchy product of u and v, namely*

$$(u * v)_\alpha = \sum_{0 \leq \beta \leq \alpha} u_{\alpha-\beta} v_\beta, \qquad \forall\ |\alpha| \geq 0,$$

where $\beta \leq \alpha$ means $\beta_i \leq \alpha_i$ for all $1 \leq i \leq n_s$ and $(\alpha - \beta)_i = \alpha_i - \beta_i$ for all $1 \leq i \leq n_s$.

Lemma 7.4.7.

$$\forall\ u, v \in \ell^1_\nu, \quad \|u * v\|_{\ell^1_\nu} \leq \|u\|_{\ell^1_\nu} \|v\|_{\ell^1_\nu}.$$

The second one bounds the action of the (infinite) diagonal part of A.

Lemma 7.4.8. *Let $d \in X = \left(\ell^1_\nu\right)^n$, such that $d_\alpha = 0$ for all $|\alpha| < N$. Then*

$$\left\|(Ad)^{(i)}\right\|_{\ell^1_\nu} \leq \frac{1}{N \min\limits_{1\leq l\leq n_s} |\Re(\lambda_l)|} \left\|d^{(i)}\right\|_{\ell^1_\nu}, \quad \forall\ 1 \leq i \leq n.$$

These two lemma allow us to get the Z_1 bound.

Proposition 7.4.9. *The bound $\tilde{Z}_1 = \left(\tilde{Z}_1^{(1)}, \ldots, \tilde{Z}_1^{(n)}\right)$ defined component-wise by*

$$\tilde{Z}_1^{(k)} = \frac{\sum_{1\leq i\leq n} \left|b^{(k)}_{\beta_i}\right| + \sum_{1\leq i,j\leq n} \left|b^{(k)}_{\beta_{i,j}}\right| \left\|(\mathcal{L}(\bar{a}))^{(i)}\right\|_{\ell^1_\nu}}{N \min\limits_{1\leq i\leq n_s} |\Re(\lambda_i)|}, \quad \forall\ 1 \leq k \leq n,$$

satisfies

$$\left\|\left(\tilde{A}\left(D\tilde{F}(\mathcal{L}(\bar{a})) - \tilde{A}^\dagger\right)c\right)^{(i)}\right\|_{\ell^1_\nu} \leq \tilde{Z}_1^{(i)}, \quad \forall\ 1 \leq i \leq n,$$

for all c such that $\|c\|_X \leq 1$.

Remark 7.4.10. *This bound is not costly, as we only need to get $\mathcal{L}(\bar{a})$ from $\bar{a}$ (a component-wise multiplication) and then to evaluate a finite sum to get the ℓ^1_ν norm of $\mathcal{L}(\bar{a})$.*

Proof. We first prove the bound without rescaling (that is for $\gamma = (1, \ldots, 1)$). By definition of $A^\dagger$, $\left(\left(DF(\bar{a}) - A^\dagger\right)c\right)_\alpha = 0$ for all $|\alpha| < N$. For $|\alpha| \geq N$, remember that the general expression for F is (for quadratic linearity)

$$F_\alpha(a) = (\alpha_1\lambda_1 + \ldots + \alpha_{n_s}\lambda_{n_s})\, a_\alpha - \sum_{|\beta|\leq 2} b_\beta \left(a^\beta\right)_\alpha, \quad \forall\ |\alpha| \geq 2.$$

Then, again by definition of $A^\dagger$, the $(\alpha_1\lambda_1 + \ldots + \alpha_{n_s}\lambda_{n_s})$ term cancels out in $\left(\left(DF(\bar{a}) - A^\dagger\right)c\right)_\alpha$ and what is left is

$$\left(\left(DF(\bar{a}) - A^\dagger\right)c\right)_\alpha = -\left(\sum_{1\leq i\leq n} b_{\beta_i} c^{(i)}_\alpha + \sum_{1\leq i,j\leq n} b_{\beta_{i,j}} \left(\bar{a}^{(i)} * c^{(j)}\right)_\alpha\right), \quad \forall\ |\alpha| \geq N, \tag{7.22}$$

where β_i must be understood as the multi-index with 1 at index i and 0 elsewhere, and $\beta_{i,j}$ as the multi-index with 1 at indexes i and j, and 0 elsewhere. We then use Lemma 7.4.7 to get

$$\left\|\left(\left(DF(\bar{a}) - A^\dagger\right)c\right)^{(k)}\right\|_{\ell^1_\nu} \leq \sum_{1\leq i\leq n} \left|b^{(k)}_{\beta_i}\right| \left\|c^{(i)}\right\|_{\ell^1_\nu} + \sum_{1\leq i,j\leq n} \left|b^{(k)}_{\beta_{i,j}}\right| \left\|\bar{a}^{(i)}\right\|_{\ell^1_\nu} \left\|c^{(j)}\right\|_{\ell^1_\nu}.$$

We now use Lemma 7.4.8 which yields

$$\left\|\left(A\left(DF(\bar{a}) - A^\dagger\right)c\right)^{(k)}\right\|_{\ell^1_\nu} \leq \frac{\sum_{1\leq i\leq n} \left|b^{(k)}_{\beta_i}\right| \left\|c^{(i)}\right\|_{\ell^1_\nu} + \sum_{1\leq i,j\leq n} \left|b^{(k)}_{\beta_{i,j}}\right| \left\|\bar{a}^{(i)}\right\|_{\ell^1_\nu} \left\|c^{(j)}\right\|_{\ell^1_\nu}}{N \min\limits_{1\leq l\leq n_s} |\Re(\lambda_l)|},$$

and the formula for Z_1 follows (in the particular case when $\gamma = (1, \ldots, 1)$), since we assumed that $\|c\|_X \leq 1$. Now we want to get the general bound. First notice that

$$\tilde{A}\left(D\tilde{F}(\mathcal{L}(\bar{a})) - \tilde{A}^\dagger\right)c = \mathcal{L}A\left(DF(\bar{a}) - A^\dagger\right)\mathcal{L}^{-1}c. \tag{7.23}$$

Then, going back to (7.22) and using that $\bar{a} * \mathcal{L}^{-1}c = \mathcal{L}^{-1}\left(\mathcal{L}(\bar{a}) * c\right)$, we get for all $|\alpha| \geq N$ that

$$\left(\left(DF(\bar{a}) - A^{\dagger}\right)\mathcal{L}^{-1}c\right)_{\alpha} = -\left(\sum_{1\leq i\leq n} b_{\beta_i}\left(\mathcal{L}^{-1}c\right)^{(i)}_{\alpha} + \sum_{1\leq i,j\leq n} b_{\beta_{i,j}}\left(\mathcal{L}^{-1}\left((\mathcal{L}(\bar{a}))^{(i)} * c^{(j)}\right)\right)_{\alpha}\right). \tag{7.24}$$

Then, since we only need to consider the action of the diagonal part of A (that is for $|\alpha| \geq N$) we can commute A and $\mathcal{L}$ in (7.23). Finally, applying $\mathcal{L}$ to (7.24) the $\mathcal{L}$ and $\mathcal{L}^{-1}$ cancel out and using again Lemma 7.4.8 we get the announced formula for $\tilde{Z}_1$. □

The bound $\tilde{Z}_2$

To get the last bound we need a last lemma, which is a combination of Lemma 7.4.3 and Lemma 7.4.8 and thus provides a bound on the full action of A.

Lemma 7.4.11. *For any $d \in \left(\ell^1_\nu\right)^n$ and for all $1 \leq i \leq n$,*

$$\left\|(Ad)^{(i)}\right\|_{\ell^1_\nu} \leq \max\left(\frac{1}{N\min\limits_{1\leq l\leq n_s}|\Re(\lambda_l)|}, K^{(i,i)}_{A^{[N]}}\right)\left\|d^{(i)}\right\|_{\ell^1_\nu} + \sum_{j\neq i} K^{(i,j)}_{A^{[N]}}\left\|d^{(j)}\right\|_{\ell^1_\nu}$$

Proposition 7.4.12. *The bound $\tilde{Z}_2 = \left(\tilde{Z}_2^{(1)}, \ldots, \tilde{Z}_2^{(n)}\right)$ defined component-wise by*

$$\tilde{Z}_2^{(k)} = \max\left(\frac{1}{N\min\limits_{1\leq i\leq n_s}|\Re(\lambda_i)|}, K^{(k,k)}_{\tilde{A}^{[N]}}\right)\sum_{1\leq i,j\leq n}\left|b^{(k)}_{\beta_{i,j}}\right| + \sum_{l\neq k} K^{(k,l)}_{\tilde{A}^{[N]}}\sum_{1\leq i,j\leq n}\left|b^{(l)}_{\beta_{i,j}}\right|, \quad \forall\ 1 \leq k \leq n,$$

where

$$\tilde{A}^{[N]} = \mathcal{L}^{[N]}A^{[N]}\left(\mathcal{L}^{[N]}\right)^{-1},$$

satisfies

$$\left\|\left(\tilde{A}D^2\tilde{F}(\mathcal{L}(\bar{a}))(b,c)\right)^{(i)}\right\|_{\ell^1_\nu} \leq \tilde{Z}_2^{(i)},$$

for all b and c such that $\|b\|_X \leq 1$ and $\|c\|_X \leq 1$.

Remark 7.4.13. The only costly part in this bound is to get $\tilde{D}$ (and the quantities $K^{(k,l)}_{\tilde{D}}$), but we already needed to compute $\tilde{D}$ for the $\tilde{Y}$ bound.

Proof. Again we prove the bound without rescaling first (that is for $\gamma = (1, \ldots, 1)$). Since we assume that F is quadratic, we get that

$$D^2F(\bar{a})(b,c) = -\sum_{1\leq i,j\leq n} b_{\beta_{i,j}} b^{(i)} * c^{(j)}, \tag{7.25}$$

with the same conventions as in Section 7.4.2 for the $\beta_{i,j}$. Therefore, using Lemma 7.4.7 and since $\|b\|_X \leq 1$ and $\|c\|_X \leq 1$,

$$\left\|\left(D^2F(\bar{a})(b,c)\right)^{(k)}\right\|_{\ell^1_\nu} \leq \sum_{1\leq i,j\leq n}\left|b^{(k)}_{\beta_{i,j}}\right|.$$

Lemma 7.4.11 then yields the formula for Z_2 (in the particular case when $\gamma = (1, \ldots, 1)$). To get the general formula, we can compute

$$\begin{aligned}\tilde{A}D^2\tilde{F}(\mathcal{L}(\bar{a}))(b,c) &= \mathcal{L}AD^2F(\bar{a})(\mathcal{L}^{-1}b, \mathcal{L}^{-1}c)\\ &= \mathcal{L}A\mathcal{L}^{-1}D^2F(\bar{a})(b,c),\end{aligned}$$

where we used $\left(\mathcal{L}^{-1}b\right) * \left(\mathcal{L}^{-1}c\right) = \mathcal{L}^{-1}\left(b * c\right)$ in (7.25). The infinite part of $\mathcal{L}A\mathcal{L}^{-1}$ (for $|\alpha| \geq N$) is the same as the one of A since the infinite part of A is diagonal. The only difference is that $\left(\mathcal{L}A\mathcal{L}^{-1}\right)^{[N]} = \mathcal{L}^{[N]}D\left(\mathcal{L}^{[N]}\right)^{-1} = \tilde{D}$, which yields the formula for $\tilde{Z}_2$. □

7.4.3 Radii polynomials

Let us sum up the results of the previous sections.

Proposition 7.4.14. *Given $\gamma = (\gamma_1, \dots, \gamma_{n_s})$, we consider $\tilde{F}$ defined as in* (7.10)*. We also consider $\bar{a}$ an element of $X = \left(\ell^1_\nu\right)^n$ such that $(\bar{a}_\alpha)_\alpha = 0$ for all $|\alpha| \geq N$ (in practice a numerical approximate zero of F) and the operator $\tilde{T}$ defined by* (7.16)*,* (7.17) *and* (7.15)*. Then the bound $\tilde{Y}$ defined in Proposition 7.4.1, and the bound*

$$\tilde{Z}(r) = (\tilde{Z}_0 + \tilde{Z}_1)r + \tilde{Z}_2 r^2,$$

where $\tilde{Z}_0$, $\tilde{Z}_1$ and $\tilde{Z}_2$ are defined in Propositions 7.4.4, 7.4.9 and 7.4.12 respectively, satisfy the hypothesis (7.18) *of Theorem 7.3.8.*

Then, for each $1 \leq i \leq n$, $\tilde{P}^{(i)}$ defined in Theorem 7.3.8 is a quadratic polynomial. If there exists $r^* > 0$ such that $\tilde{P}^{(i)}(r^*) < 0$ for all $1 \leq i \leq n$, then there exists an interval $\mathcal{I} = (r_0, r_1)$ such that $\tilde{P}^{(i)}(r) < 0$ for all $1 \leq i \leq n$ and for all $r \in \mathcal{I}$. By Theorem 7.3.8, we know that for all $r \in \mathcal{I}$, within a ball of radius r centered in $\mathcal{L}(\bar{a})$ their exists a unique local parameterization of the manifold. Moreover, if one wants to make this fully rigorous, a final step consists of computing the bounds $\tilde{Y}$ and $\tilde{Z}$ with interval arithmetic and then check, still with interval arithmetic, that $\tilde{P}^{(i)}(r)$ is negative.

Finally, if the goal is to get a proof-valid parameterization while having the largest possible image, we process as follows. We start by computing the bounds (and the associated radii polynomials) without rescaling. Then if $\mathcal{I}$ is empty, or if $r_{max} < r_0$, we can rescale $\bar{a}$ to $\mathcal{L}(\bar{a})$ by some γ and then compute the interval $\mathcal{I}$ associated to the rescaled polynomials $\tilde{P}^{(i)}$ (of course one should choose $\gamma_k < 1$) but this time the computation of the coefficients of the polynomials, namely $\tilde{Y}$, $\tilde{Z}_0$, $\tilde{Z}_1$ and $\tilde{Z}_2$, are much faster thanks to the formulas of the previous sections. Conversely, if r_0 is small compared to r_{max}, we can rescale $\bar{a}$ to $\mathcal{L}(\bar{a})$ by some γ, this time with $\gamma_k > 1$ larger and larger, which will give a larger and larger manifold patch associated to the rescaled parameterization, until we reach the limit of $r_0 = r_{max}$. We explain more in detail how we do this on an example in Section 7.5.3.

7.5 Examples

7.5.1 Defect-valid parameterizations for the Lorenz system

As a first example, we consider the Lorenz system, given by the vector field

$$g(x, y, z) = \begin{pmatrix} \sigma(y - x) \\ \rho x - y - xz \\ xy - \beta z \end{pmatrix},$$

with standard parameter values : $\sigma = 10$, $\beta = \frac{8}{3}$ and $\rho = 28$. In this case it is well known that the origin has a two dimensional stable manifold. We detail on this example the method presented in Sections 7.2 and 7.3 to automatically compute a maximal patch of the local stable manifold at $p = 0$.

We start by recalling that the stable eigenvalues are

$$\lambda_1 = -\frac{1}{2}\left(\sigma + 1 + \sqrt{(\sigma - 1)^2 + 4\sigma\rho}\right) \quad \text{and} \quad \lambda_2 = -\beta,$$

together with the stable eigenvectors

$$V_1 = \begin{pmatrix} \frac{\sigma}{\lambda_1 + \sigma} \\ 1 \\ 0 \end{pmatrix} \quad \text{and} \quad V_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

As explained in Section 7.2, we look for a parameteriztion of the local stable manifold in the form of a power series f, which should satisfy the invariance equation

$$Df(\theta)\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}\theta = g(f(\theta)). \tag{7.26}$$

together with the condition conditions

$$f(0) = p \quad \text{and} \quad Df(0) = \begin{pmatrix}V_1 & V_2\end{pmatrix}.$$

Notice that in this case the two stable eigenvalues are real and therefore we can directly work with a real power series defined on $[-1,1]^2$. Expanding f into a power series, (7.26) rewrites as

$$\sum_{|\alpha|\geq 2}(\alpha_1\lambda_1+\alpha_2\lambda_2)a_\alpha\theta^\alpha = \sum_{|\alpha|\geq 2}\begin{pmatrix}\sigma\left(a_\alpha^{(2)}-a_\alpha^{(1)}\right)\\ \rho a_\alpha^{(1)}-a_\alpha^{(2)}-\left(a^{(1)}*a^{(3)}\right)_\alpha\\ \left(a^{(1)}*a^{(2)}\right)_\alpha-\beta a_\alpha^{(3)}\end{pmatrix}\theta^\alpha,$$

where

$$a_\alpha = \begin{pmatrix}a_\alpha^{(1)}\\ a_\alpha^{(2)}\\ a_\alpha^{(3)}\end{pmatrix}.$$

So we set $a_{0,0}=p$, $a_{1,0}=V_1$, $a_{0,1}=V_2$ and define $F=(F_\alpha)_{|\alpha|\geq 2}$, acting on $a=(a_\alpha)_{|\alpha|\geq 2}$, by

$$F_\alpha(a) = (\alpha_1\lambda_1+\alpha_2\lambda_2)a_\alpha - \begin{pmatrix}\sigma\left(a_\alpha^{(2)}-a_\alpha^{(1)}\right)\\ \rho a_\alpha^{(1)}-a_\alpha^{(2)}-\left(a^{(1)}*a^{(3)}\right)_\alpha\\ \left(a^{(1)}*a^{(2)}\right)_\alpha-\beta a_\alpha^{(3)}\end{pmatrix}, \quad \forall\ |\alpha|\geq 2.$$

Our goal is now to find a numerical zero $\bar{a}$ and then the rescaling $\gamma=(\gamma_1,\gamma_2)$ so that the parameterization $\tilde{f}$ defined as

$$\tilde{f}(\theta) = \sum_{|\alpha|\geq 0}\mathcal{L}_\alpha(\bar{a})\theta^\alpha, \quad \forall\ \theta\in[-1,1]^2,$$

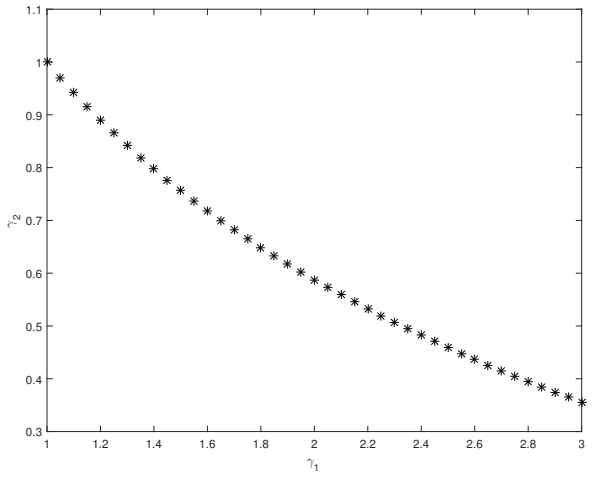
gives us the maximal patch of manifold, while checking (according to Definition 7.3.6) that $\|\tilde{F}(\mathcal{L}(\bar{a}))\|_X < \varepsilon_{max}$, which will ensure that $\mathcal{L}(\bar{a})$ is a good approximate parameterization.

First we fix an integer N and consider a truncated version of F, that is

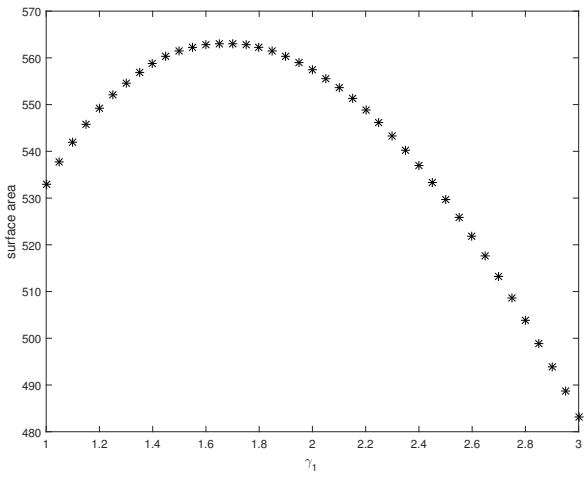
$$F^{[N]} = (F_\alpha)_{2\leq|\alpha|<N},$$

for which we can numerically compute a zero $\bar{a}$ with Newton's method. Then we fix an ε_{max} and use Method 1 described in Section 7.3. First we compute $F(\bar{a})$, which can be done explicitly because by construction $\bar{a}_\alpha = 0$ for any $|\alpha|\geq N$, so for $i=1$, $F^{(i)}(\bar{a})=0$ for any $|\alpha|\geq N$ and for $i\in\{2,3\}$, $F^{(2)}(\bar{a})=0$ for any $|\alpha|\geq 2N-1$ (because of the quadratic terms). Then we find numerically the curve in the plane (γ_1,γ_2) that corresponds to $\|\tilde{F}(\mathcal{L}(\bar{a}))\|_X=\varepsilon_{max}$. In our case, we took a sample of values of γ_1 and for each we looked for the largest γ_2 for which $\|\tilde{F}(\mathcal{L}(\bar{a}))\|_X<\varepsilon_{max}$ (as explained in Section 7.3 this doesn't require much computations since the coefficient of $F(\bar{a})$ are already known). Finally we compute the surface of the corresponding patch of the manifold along this sample and find its maximum. The results are displayed in Figure 7.1, along with the results of similar computations for the unstable manifolds of the nontrivial equilibria, or "eyes," of the attractor.

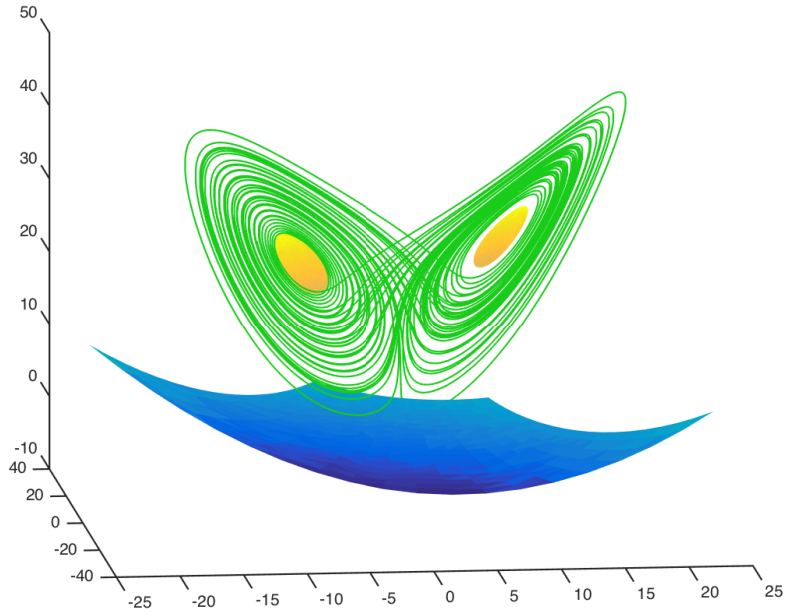
By way of contrast we consider another parameterization of the local stable manifold at p but focusing on the slow direction given by V_2. Therefore we apply Method 2 described in Section 7.3: we define the ratio $\varrho := \left|\frac{\lambda_1}{\lambda_2}\right|$ and only consider rescalings of the form $\gamma=(\gamma_1,\varrho\gamma_1)$. Then we simply find numerically the largest γ_1 such that the rescaled parameterization $\mathcal{L}(\bar{a})$ is defect valid, and obtain the results displayed in Figure 7.2.



(a) The curve of (γ_1, γ_2) for which $\|\tilde{F}(\mathcal{L}(\bar{a}))\|_X = \varepsilon_{max}$ for the local stable manifold of the origin.



(b) The corresponding values of the surface area (again for the local stable manifold of the origin).



(c) Lorenz System: local stable manifold of the origin (with the rescaling maximizing the surface area, i.e. $\gamma_1 = 1.7$ and $\gamma_2 = 0.68$) and local unstable manifolds of the eyes. The unstable manifolds have complex conjugate eigenvalues, so we simply maximize the length of the eigenvectors.

Figure 7.1 – For each manifold we take a defect constraint of $\varepsilon_{max} = 10^{-5}$. The order of the parameterizations is $N = 50$ for the eyes and $N = 30$ for the stable local manifold of the origin.

7.5.2 Defect-valid parameterizations for the FitzHugh-Nagumo equations

We consider the vector field given by

$$g(u, v, w) = \begin{pmatrix} v \\ \frac{1}{\Delta}\left(sv + w - q + u^3 - (1+\sigma)u^2 + \sigma u\right) \\ \frac{\varepsilon}{s}\left(u - \zeta w\right) \end{pmatrix},$$

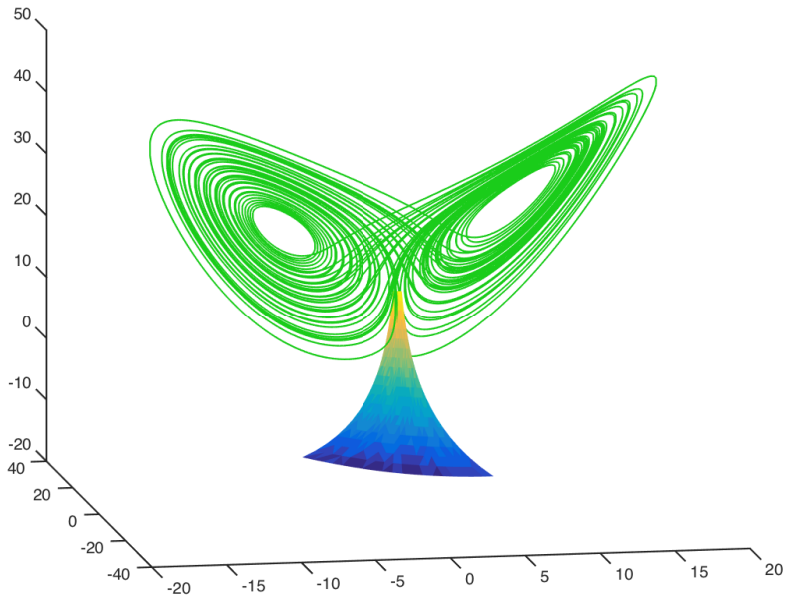


Figure 7.2 – Lorenz System: the figure illustrates the results of maximizing the lengths of the stable eigenvectors of the origin subject to the constraint that the slow eigenvector is $\varrho = |\lambda_1|/|\lambda_2|$ times longer than the fast eigenvector and that the defect is less than $\varepsilon_{max} = 10^{-5}$. The order of this parameterization is $N = 50$. Note that the resulting patch covers more of the slow stable manifold than the patch shown in Figure 7.1, however the surface area is smaller.

where

$$\sigma = \frac{1}{10},\ s = 1.37,\ \Delta = 1,\ q = 0.001,\ \varepsilon = 0.15 \text{ and } \zeta = 5.$$

There are trivial zeros given by $v = 0$, $w = \frac{u}{\zeta}$ and u solution of the cubic equation

$$u^3 - (1+\sigma)u^2 + \left(\sigma + \frac{1}{\zeta}\right)u - q = 0.$$

We want to compute the stable local manifold at one of them:

$$p \simeq \begin{pmatrix} 0.003374970076610 \\ 0 \\ 0.000674994015322 \end{pmatrix}.$$

With the selected values of the parameters we have two real stable eigenvalues at this point p:

$$\lambda_1 \simeq -0.662724919921474 \quad \text{and} \quad \lambda_2 \simeq -0.184083645070452,$$

with associated eigenvectors

$$V_1 \simeq \begin{pmatrix} -0.576099055982516 \\ 0.381795200742850 \\ -0.722732524787547 \end{pmatrix} \quad \text{and} \quad V_2 \simeq \begin{pmatrix} -0.966141520359494 \\ 0.177850852721684 \\ -0.186921472344981 \end{pmatrix}.$$

In this case we also want to compute a parameterization of the local stable manifold at p focusing more on the slow direction given by V_2. Therefore we again apply Method 2 and obtain the results displayed in Figure 7.3.

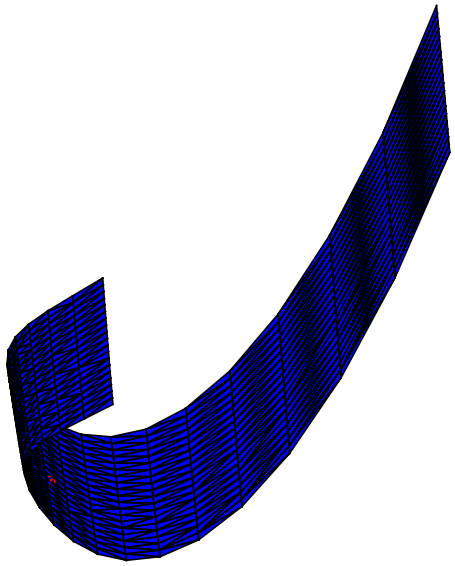


Figure 7.3 – FitzHugh-Nagumo System: the figure illustrates the results of maximizing the lengths of the stable eigenvectors of the origin subject to the constraint that the slow eigenvector is $\varrho = |\lambda_1|/|\lambda_2|$ times longer than the fast eigenvector and that the defect is less than 10^{-5}. The order of this parameterization is $N = 30$. The red star indicates the location of the equilibrium. The local manifold illustrated here is not the graph of any function over the stable eigenspace, i.e. the parameterization follows a fold in the manifold. Note that the triangulation in the figure is an artifact of the plotting procedure for the manifolds, and not part of the parameterization computation. If a finer mesh is desired we simply evaluate the polynomial approximation at more points. It is not necessary to re-compute the parameterization.

7.5.3 Proof-valid parameterizations for the suspension bridge equation

We consider the vector field

$$g(v) = \begin{pmatrix} v_2 + v_1 v_2 \\ v_3 \\ v_4 \\ -\beta v_3 - v_1 \end{pmatrix},$$

which is obtained after a change of variable when one looks for travelling waves in the suspension bridge equation (e.g. see [98, 80])

$$\frac{\partial^2 u}{\partial t^2} = -\frac{\partial^4 u}{\partial x^4} - (e^u - 1)\,.$$

We are going to rigorously compute the stable manifold at 0 (for a given $\beta \in (0, 2)$), which is two-dimensional. The stable eigenvalues are λ and $\overline{\lambda}$, where

$$\lambda = -\frac{1}{2}\sqrt{2-\beta} + i\frac{1}{2}\sqrt{2+\beta}, \tag{7.27}$$

and associated eigenvectors are given by

$$V_1 = \begin{pmatrix} 1 \\ \lambda \\ \lambda^2 \\ \lambda^3 \end{pmatrix} \quad \text{and} \quad V_2 = \overline{V}_1.$$

We then define $F = (F_\alpha)_{|\alpha|\geq 0}$, acting on $a = (a_\alpha)_{|\alpha|\geq 0} \in \left(\ell^1\right)^4$, by

$$F_\alpha(a) = \begin{cases} a_0 - 0, & \text{if } \alpha = 0, \\ a_{1,0} - V_1, & \text{if } \alpha = (1,0) \\ a_{0,1} - V_2, & \text{if } \alpha = (0,1) \\ \\ (\alpha_1\lambda + \alpha_2\overline{\lambda})a_\alpha - \begin{pmatrix} a_\alpha^{(2)} + (a^{(1)} * a^{(2)})_\alpha \\ a_\alpha^{(3)} \\ a_\alpha^{(4)} \\ -a_\alpha^{(1)} - \beta a_\alpha^{(3)} \end{pmatrix}, & \forall\ |\alpha| \geq 2. \end{cases}$$

This time since the eigenvalues are not real, we consider complex parameterization a, i.e. $a_\alpha \in \mathbb{C}^4$ for all α. Then we compute a numerical zero $\bar{a}$ with the method described in Section 7.3.1. To rigorously prove the existence of a nearby solution a we follow the ideas exposed in Section 7.3 and consider an operator T of the form

$$T : a \mapsto a - AF(a).$$

The following infinite matrix should be a good approximation of $DF(\bar{a})$ (at least for N large enough)

$$A^\dagger = \begin{pmatrix} DF^{[N]}(\bar{a}) & & 0 & \\ & A_N & & \\ 0 & & A_{N+1} & \\ & & & \ddots \end{pmatrix},$$

where for each $k \geq N$, A_k is a $4(k+1)$ by $4(k+1)$ bloc diagonal matrix defined as

$$A_k = \begin{pmatrix} k\lambda I_4 & & 0 & \\ & ((k-1)\lambda + \overline{\lambda})I_4 & & \\ 0 & & \ddots & \\ & & & k\overline{\lambda}I_4 \end{pmatrix},$$

with I_4 the 4 by 4 identity matrix. Therefore we define

$$A = \begin{pmatrix} D & & 0 & \\ & M_N & & \\ 0 & & M_{N+1} & \\ & & & \ddots \end{pmatrix},$$

where D is a numerical approximation of $DF^{[N]}(\bar{a})^{-1}$ while the $M_k = A_k^{-1}$ are exact inverses.

We are now ready to compute the bounds $\tilde{Y}$ and $\tilde{Z}$ defined in Section 7.3 in order to apply Theorem 7.3.8 an prove the existence of a true parameterization near $\bar{a}$. In practice, we first compute the bounds without rescaling (that is for $\gamma = (1,1)$) and denote them simply Y and Z, and then we find the largest rescaling for which the parameterization is still proof valid.

Computation of the bounds Y and Z, and of the radii polynomials

Concerning the bounds Y and Z_0, there is nothing to add or to specify to what was said in Section 7.4. We set, for $1 \leq i \leq 4$

$$Y^{(i)} = \left\| \left(AF(\bar{a})\right)^{(i)} \right\|_{\ell^1_\nu},$$

and

$$Z_0^{(i)} = \sum_{j=1}^{4} K_B^{(i,j)},$$

where the $K_B^{(i,j)}$ are defined as in Section 7.4.2. For Z_1 and Z_2 we can specify the bounds of Section 7.4, because we now work with a specific nonlinearity. We get

$$Z_1^{(1)} = \frac{1 + \left\|\bar{a}^{(1)}\right\|_{\ell^1_\nu} + \left\|\bar{a}^{(2)}\right\|_{\ell^1_\nu}}{N|\Re(\lambda)|}, \quad Z_1^{(2)} = \frac{1}{N|\Re(\lambda)|}, \quad Z_1^{(3)} = \frac{1}{N|\Re(\lambda)|}, \quad Z_1^{(4)} = \frac{1+\beta}{N|\Re(\lambda)|},$$

and

$$Z_2^{(1)} = 2\max\left(K_D^{(1,1)}, \frac{1}{|\Re(\lambda)|N}\right), \quad Z_2^{(2)} = 2K_D^{(2,1)}, \quad Z_2^{(3)} = 2K_D^{(3,1)}, \quad Z_2^{(4)} = 2K_D^{(4,1)}.$$

Now we can consider, for all $1 \leq i \leq 4$, the radii polynomial defined by

$$P^{(i)}(r) = Y^{(i)} + (Z_0^{(i)} + Z_1^{(i)} - 1)r + Z_2^{(i)} r^2$$

and we can try and look for a positive r such that $P^{(i)}(r) < 0$ for all $1 \leq i \leq 4$.

Remark 7.5.1. *Y should be very small if $\bar{a}$ is a good approximative zero of F. Z_0 should also be very small because $B = I_{\frac{N(N+1)}{2}} - D(DF^{[N]}(\bar{a}))$ and D is a numerical inverse of $(DF^{[N]}(\bar{a}))$. Finally Z_1 can be made very small by choosing N large enough. Therefore the radii polynomials are of the form*

$$P^{(i)}(r) = \varepsilon - (1-\eta)r + Z_2^{(i)} r^2$$

where ε could be made arbitrarily small if we could get an arbitrarily good approximation $\bar{a}$ and η could be made arbitrarily small if we could take with an arbitrarily large N (and if we could numerically compute inverses of matrices with sufficient accuracy). So up to having sufficient computational precision we should always be able to find a positive r such that $P^{(i)}(r) < 0$.

Results

For this problem we are interested in proving (rigorously and with and error bound r smaller than r_{max}) the largest possible patch of the stable manifold, for values of β between 0.5 and 2. Since we already computed the bounds Y, Z_0, Z_1 and Z_2 without rescaling, we can now easily compute the radii polynomial $\tilde{P}$ for any rescaling, and so we look by dichotomy for the largest γ such that the rescaled radii polynomial $\tilde{P}$ has a positive root r_0 which is less or equal to r_{max}. Notice that the eigenvalues are complex conjugated for this problem and that is why we only consider uniform rescaling (i.e. $\gamma_1 = \gamma_2$).

When β goes to 2, the real part of λ goes to 0 (remember (7.27)) so we expect it to be harder and harder to compute the manifold when β goes to 2. Indeed we observe that the largest γ for which we are able to do the proof becomes smaller and smaller when β goes to 2 (see Figure 7.4). The computations were made with $N = 30$, $\nu = 1$ and $r_{max} = 10^{-5}$ for the proof.

Remark 7.5.2. *Another interesting point here is that a closer look at the bound Z_0 shows why it is better to take $\nu = 1$. The matrix B is supposed to be approximatively 0, and we want the terms $K_B^{(i,j)}$ of Lemma 7.4.3 to be as small as possible, but their definition*

$$K_B^{(i,j)} = \max_{|\beta|<N} \left(\frac{1}{\nu^{|\beta|}} \sum_{|\alpha|<N} \left| B_{\alpha,\beta}^{(i,j)} \right| \nu^{|\alpha|} \right).$$

show that there is a risk of numerical errors if ν is too small or too large, hence our choice of always considering $\nu = 1$.

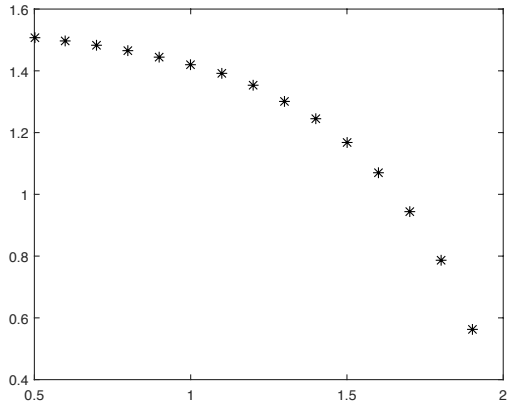


Figure 7.4 – Maximal value of γ for which we can still do the proof with $r \leq r_{max}$, for different values of β. The manifold computations are completely automated.

To speed up the process of redoing the proof after a rescaling, we kept track of the γ dependency in the bound Y and Z, and constructed the rescaled bound $\tilde{Y}$ and $\tilde{Z}$ based on the original ones. However by doing things this way we introduce in the $\tilde{Z}_0$ bound the same kind of instability that comes with taking $\nu \neq 1$ (see (7.21)). If the $\tilde{Z}_0$ bound becomes too big, we could deal with it (at the expense of speed), by recomputing all the bounds without using the fact that they came from a rescaling and thus eliminating this numerical instability issue.

Chapter 8

Polynomial interpolation and *a priori bootstrap* for validated numerics

Abstract

This chapter is taken from [75]. We introduce a method based on piecewise polynomial interpolation to enclose rigorously solutions of nonlinear ODEs. Using a technique which we call *a priori bootstrap*, we transform the problem of solving the ODE into one of looking for a fixed point of a high order smoothing Picard-like operator. We then develop a rigorous computational method based on a Newton-Kantorovich type argument (the radii polynomial approach) to prove existence of a fixed point of the Picard-like operator. We present all necessary estimates in full generality and for any nonlinearities. With our approach, we study two systems of nonlinear equations: the Lorenz system and the ABC flow. For the Lorenz system, we solve Cauchy problems and prove existence of periodic and connecting orbits at the classical parameters, and for ABC flows, we prove existence of ballistic spiral orbits.

8.1 Introduction

This paper introduces an approach based on polynomial interpolation to obtain mathematically rigorous results about existence of solutions of nonlinear ordinary differential equations (ODEs). Our motivation for the present work is threefold. First, we believe that polynomial interpolation techniques are versatile and can lead to efficient and general computational methods to approximate solutions of ODEs with complicated (non polynomial) nonlinearities. Second, while polynomial interpolation techniques have be used to produce computer-assisted proofs in ODEs, their applicability to produce proofs is sometimes limited by the formulation of the problem itself. More precisely, a standard way to prove (both theoretically and computationally) existence of solutions of systems of ODEs is to reformulate the problem into an integral equation (often in the form of a Picard operator) and then to apply the contraction mapping theorem to get existence. If one is interested to produce computer-assisted proofs using that approach, the analytic estimates required to perform the proofs depend on the amount of regularity gained by applying the integral operator. This observation motivated developing what we call the *a priori bootstrap*, which consists of reformulating the original ODE problem into one of looking for the fixed point of a higher order smoothing Picard-like operator. Third, we believe (and hope) that our proposed method can be adapted to study infinite dimensional continuous dynamical

systems (e.g. partial differential equations and delay differential equations) for which spectral methods may sometimes be difficult to apply (for instance because of the shape of the spatial domains or because the differential operators are difficult to invert in a given spectral basis).

It is important to realize that computer-assisted arguments to study differential equations are by now standard, and that providing a complete overview of the literature would require a tremendous effort and is outside the scope of this paper. However, we encourage the reader to read the survey papers [159, 147, 191, 173, 177, 63], the books [201, 130] and to consult the webpage of the CAPD group [86] to get a flavour of the extraordinary recent advances made in the field.

More closely related to the present work are methods based on the contraction mapping theorem using the *radii polynomial approach* (first introduced in [109]), which has been developed in the last decade to study fixed points, periodic orbits, invariant manifolds and connecting orbits of ODEs, partial differential equations and delay differential equations (see for instance [60, 67, 77, 95, 122, 153]). The numerics and a posteriori analysis in those works mainly use spectral methods like Fourier and Chebyshev series, and Taylor methods. First order polynomial (piecewise linear) approximations were also used using the radii polynomial approach (see [61, 65, 155]), but more seldom, mainly because the numerical cost was higher and the accuracy was lower than for spectral methods. The computational cost of these low order methods is essentially due to the above mentioned low gain of regularity of the Picard operators chosen to perform the computer-assisted proofs.

In an attempt to address the low gain of regularity problem, we present here a new technique that we call *a priori bootstrap* which, when combined with the use of higher order interpolation, significantly improves the efficiency of computer-assisted proofs with polynomial interpolation methods. We stress that the limitations that affected the previous works using interpolation were not solely due to the use of first order methods, and that the *a priori bootstrap* is crucial (that is, just increasing the order of the interpolation does not significantly improve the results in those previous works). This point is illustrated in Section 8.6.

While we believe that one of the advantage of our proposed method is the versatility of the polynomial interpolations to tackle problems with complicated (non polynomial) nonlinearities, we hasten to mention the existence of previous powerful methods which have been developed in rigorous computing to study such problems. For instance, *automatic differentiation* (AD) techniques provide a beautiful and efficient means of computing solutions of nonlinear problems (e.g. see [201, 145, 139]) and are often combined with Taylor series expansions to prove existence of solutions of differential equations with non polynomial nonlinearities (e.g. see [159, 69, 210, 218, 207, 217, 190]). Also, in the recent work [154], the ideas of AD are combined with Fourier series to prove existence of periodic orbits in non polynomial vector fields. Independently of AD techniques, a method involving Chebyshev series to approximate the nonlinearities have been proposed recently in [108]. Finally, the fast Fourier transform (FFT) algorithm is used in [116] to control general nonlinearities in the context of computer-assisted proofs in KAM theory.

In this paper we consider $\phi : \mathbb{R}^n \to \mathbb{R}^n$ a $C^{\boldsymbol{r}}$ vector field (not necessarily polynomial) with $\boldsymbol{r} \geq 1$, and we present a rigorous numerical procedure to study problems of the form

$$\begin{cases} \dfrac{du}{dt}(t) = \phi(u(t)), \quad t \in [0, \tau], \\ BV(u(0), u(\tau)) = 0. \end{cases} \tag{8.1}$$

We treat three special cases for BV, corresponding to an initial value problem, a problem of looking for periodic orbits and a problem of looking for connecting orbits. We also note that, as for the already existing spectral methods, the technique presented here extends easily to treat parameter dependent versions of (8.1) (e.g. using ideas from [64, 79, 123]). For the sake

of simplicity, we expose all the general arguments for the initial value problem only, that is for

$$\begin{cases} \dfrac{du}{dt}(t) = \phi(u(t)), \quad t \in [0,\tau], \\ u(0) = u_0, \end{cases} \tag{8.2}$$

given an integration time $\tau > 0$ and an initial condition $u_0 \in \mathbb{R}^n$. We explain how (8.2) needs to be modified for different problems as we introduce them in Sections 8.6 and 8.7.

Our paper is organized as follows. In Section 8.2, we start by presenting our *a priori bootstrap* technique, together with a *piecewise* reformulation of the operator that we use throughout this work. We then recall in Section 8.3 some definitions and error estimates about polynomial interpolation, and explain how to combine them with our *a priori bootstrap* formulation to get computer-assisted proofs. The precise estimates needed for the proofs are then derived in Section 8.4, and their dependency with respect to the *a priori bootstrap* and to the parameters of the polynomial interpolation is commented in Section 8.5. This discussion is complemented by several examples in Section 8.6, where we apply our technique to validate solutions for the Lorenz system. Finally we give another example of application in Section 8.7, where we prove the existence of some specific orbits for ABC flows.

8.2 Reformulations of the Cauchy problem

8.2.1 A priori bootstrap

One of the usual strategies used to study (8.2), both theoretically and numerically, is to recast it as a fixed point problem, as in the following lemma.

Lemma 8.2.1. *Consider the standard Picard operator*

$$f : \begin{cases} \mathcal{C}^0([0,1],\mathbb{R}^n) \to \mathcal{C}^1([0,1],\mathbb{R}^n) \\ u \mapsto f(u), \end{cases}$$

where

$$f(u)(t) := u_0 + \tau \int_0^t \phi(u(s))ds, \quad \textit{for all } t \in [0,1]. \tag{8.3}$$

Then u is a fixed point of f if and only if $v : t \mapsto u(\frac{t}{\tau})$ is a solution of (8.2).

In previous works using this reformulation, the limiting factor in the estimates needed to apply the contraction mapping theorem was a consequence of the fact that f only gains one order of regularity, that is maps $\mathcal{C}^0$ into $\mathcal{C}^1$. This fact will be made precise in Section 8.4 where we derive the estimates in question and in Section 8.5 where we discuss how those estimates affect the effectiveness of our technique.

To circumvent this limitation, we propose a different reformulation that we call *a priori bootstrap*. This approach provides operators which gain more regularity, and therefore lead to sharper analytic estimates. First we introduce some notations. The following definition allows concisely describing the higher order equations obtained by taking successive derivatives of (8.2).

Definition 8.2.2. *Consider the sequence of vector fields $\left(\phi^{[p]}\right)_{0\le p\le \boldsymbol{r}+1}$ with $\phi^{[p]} : \mathbb{R}^n \to \mathbb{R}^n$,*

$$\phi^{[0]}(u) := u \quad \textit{and} \quad \phi^{[p+1]}(u) := D\phi^{[p]}(u)\phi(u) \quad \textit{for all } u \in \mathbb{R}^n \textit{ and for all } p = 0,\dots,\boldsymbol{r}.$$

Lemma 8.2.3. *For any $1 \le p \le \boldsymbol{r}+1$, u solves* (8.2) *if and only if u solves the following Cauchy problem*

$$\begin{cases} \dfrac{d^p u}{dt^p}(t) = \phi^{[p]}(u(t)), \quad t \in [0,\tau], \\ \dfrac{d^q u}{dt^q}(0) = \phi^{[q]}(u_0), \quad \textit{for all } q = 0,\dots,p-1. \end{cases} \tag{8.4}$$

Proof. The direct implication is trivial. To prove the converse application, we consider $e := \frac{du}{dt} - \phi(u)$ and show that is solve a linear ODE of order $p-1$, with initial conditions $\frac{d^q e}{de^q}(0) = 0$ for all $q = 0, \dots, p-2$, which implies that $e \equiv 0$. □

Integrating the p^{th} order Cauchy problem (8.4) p times leads to a new fixed point operator which now maps $\mathcal{C}^0$ into $\mathcal{C}^p$.

Lemma 8.2.4. *Let $1 \leq p \leq \boldsymbol{r}+1$ and consider the Picard-like operator*

$$\tilde{g} : \begin{cases} \mathcal{C}^0([0,1],\mathbb{R}^n) \to \mathcal{C}^p([0,1],\mathbb{R}^n) \\ u \mapsto \ \tilde{g}(u), \end{cases}$$

where

$$\tilde{g}(u)(t) := \sum_{q=0}^{p-1} \tau^q \frac{t^q}{q!} \phi^{[q]}(u_0) + \tau^p \int_0^t \frac{(t-s)^{p-1}}{(p-1)!} \phi^{[p]}(u(s)) ds, \quad \textit{for all } t \in [0,1]. \tag{8.5}$$

Then u is a fixed point of $\tilde{g}$ if and only if $v : t \mapsto u(\frac{t}{\tau})$ is a solution of (8.4) (and thus of (8.2)).

Proof. If u is a fixed point of $\tilde{g}$, an elementary computation yields that v solves (8.4). Conversely, if v solves (8.4) then Taylor's formula with integral reminder shows that u is a fixed point of $\tilde{g}$. □

It is worth noting that, in the same framework of rigorous computation as the one used here, approximations using piecewise linear functions were used in [65] to prove existence of homoclinic orbits for the Gray-Scott equation. In that case the system of ODEs considered is of order 2, and therefore the equivalent integral operator is very similar to $\tilde{g}$ in (8.5) for $p = 2$. Similarly, piecewise linear functions were used in [155] to prove existence of connecting orbits in the Lorenz equations. In that case, the standard Picard operator (8.3) was used.

Now that we have an operator which provides a gain of several orders of regularity, it becomes interesting to consider polynomial interpolation of higher order, and again this will be detailed in Section 8.5 and applied in Section 8.6.

8.2.2 Piecewise reformulation of the Picard-like operator

We finish this section by a last equivalent formulation of the initial value problem (8.2), that will be the one used in the present paper to perform the computer-assisted proofs. Given $m \in \mathbb{N}$, we introduce the mesh of $[0,1]$

$$\Delta_m := \{t_0, t_1, \dots, t_m\},$$

where $t_0 = 0 < t_1 < \ldots < t_m = 1$. Then we consider $\mathcal{C}^0_{\Delta_m}([0,1],\mathbb{R}^n)$ (respectively $\mathcal{C}^k_{\Delta_m}([0,1],\mathbb{R}^n)$) the space of piecewise continuous (respectively $\mathcal{C}^k$) functions having possible discontinuities only on the mesh Δ_m. More precisely, we use the following definition.

Definition 8.2.5. *For $k \in \mathbb{N}$, we say that $u \in \mathcal{C}^k_{\Delta_m}([0,1],\mathbb{R}^n)$ if $u_{|_{(t_j,t_{j+1})}} \in \mathcal{C}^k((t_j,t_{j+1}),\mathbb{R}^n)$ and can be extended to a $\mathcal{C}^k$ function on $[t_j,t_{j+1}]$, for all $j = 0,\dots,m-1$.*

We then introduce

$$g : \begin{cases} \mathcal{C}^0_{\Delta_m}([0,1],\mathbb{R}^n) \to \mathcal{C}^p_{\Delta_m}([0,1],\mathbb{R}^n) \\ u \mapsto g(u), \end{cases}$$

defined on the interval (t_j, t_{j+1}) $(j = 0,\dots,m-1)$ by

$$g(u)(t) := \sum_{q=0}^{p-1} \tau^q \frac{(t-t_j)^q}{q!} \phi^{[q]}(u(t_j^-)) + \tau^p \int_{t_j}^t \frac{(t-s)^{p-1}}{(p-1)!} \phi^{[p]}(u(s)) ds, \tag{8.6}$$

where $u(t_j^-)$ denotes the left limit of u at t_j, and $u(t_0^-)$ must be replaced by u_0 (this last convention will be used throughout the paper).

Remark 8.2.6. *We point out that our computer-assisted proof is based on the operator g (defined in* (8.6)*), which differs slightly from the operator $\tilde{g}$ (defined in* (8.5)*), which was used in previous studies such as [61, 65, 155]. The only difference is that the integral in g is in some sense reseted at each t_j. We introduce this* piecewise *reformulation because it allows for sharper estimates (see Remark 8.4.7).*

We finally introduce $G : \mathcal{C}^0_{\Delta_m}([0,1],\mathbb{R}^n) \to \mathcal{C}^0_{\Delta_m}([0,1],\mathbb{R}^n)$ as

$$G(u) := g(u) - u.$$

Lemma 8.2.7. *Let $u \in \mathcal{C}^0_{\Delta_m}([0,1],\mathbb{R}^n)$. Then $G(u) = 0$ if and only if $v : t \mapsto u(\frac{t}{\tau})$ solves* (8.2)*.*

Proof. This result is similar to Lemma 8.2.4. The only additional property that we need to check is that, if $u \in \mathcal{C}^0_{\Delta_m}([0,1],\mathbb{R}^n)$ satisfies $G(u) = 0$, then u cannot be discontinuous. Indeed, if $G(u) = 0$ then $g(u) = u$, and for all $j \in \{1,\dots,m-1\}$ one has

$$u(t_j^-) = \lim_{t \to t_j^+} g(u)(t) = \lim_{t \to t_j^+} u(t) = u(t_j^+). \qquad \square$$

At this point, it might seems as if defining G on $\mathcal{C}^0_{\Delta_m}([0,1],\mathbb{R}^n)$ brings unnecessary complications, and that we should simply define it on $\mathcal{C}^0([0,1],\mathbb{R}^n)$. While this is indeed a possibility, it will quickly become apparent that the present choice is more convenient, both for theoretical and numerical considerations (see Remark 8.3.4).

Finding a zero of G is the formulation of our initial problem (8.2) that we are going to use in the rest of this paper.

8.3 General framework for the polynomial interpolation

8.3.1 Preliminaries

Given a mesh Δ_m as defined in Section 8.2.2, we introduce the refined mesh $\Delta_{m,k}$ where, for all $j \in \{0,\dots,m-1\}$ we add $k-1$ points between t_j and t_{j+1}. More precisely we suppose that these points are the Chebyshev points (of the second kind) between t_j and t_{j+1}, that is we add the following points:

$$t_{j,l} := t_j + \frac{x_l^k + 1}{2}(t_{j+1} - t_j), \quad \text{for } l = 1,\dots,k-1,$$

where

$$x_l^k := \cos\theta_l^k, \quad \theta_l^k := \frac{k-l}{l}\pi, \quad \text{for } l = 0,\dots,k.$$

Notice that the above definition extends to $t_{j,0} = t_j$ and $t_{j,k} = t_{j+1}$, and that $k = 1$ corresponds to the mesh used in previous studies with first order interpolation (e.g. see [65, 155, 61]).

We then introduce the subspace $S^n_{m,k} \subset \mathcal{C}^0_{\Delta_m}([0,1],\mathbb{R}^n)$ of piecewise polynomial functions of degree k on Δ_m

$$S^n_{m,k} := \left\{ u \in \mathcal{C}^0_{\Delta_m}([0,1],\mathbb{R}^n) : \ u_{|(t_j,t_{j+1})} \text{ is a polynomial of degree } k \text{ for all } j = 0,\dots,m-1 \right\}.$$

Next, we define the projection operator

$$\Pi^n_{m,k} : \begin{cases} \mathcal{C}^0_{\Delta_m}([0,1],\mathbb{R}^n) \to S^n_{m,k} \\ \qquad\qquad u \mapsto \bar{u} = \Pi^n_{m,k}(u), \end{cases}$$

where $\bar{u}$ is the function in $S^n_{m,k}$ that matches the values of u on the mesh $\Delta_{m,k}$. Notice that u can have discontinuities at the points t_j, therefore the matching of u and $\bar{u}$ at those points must be understood as

$$u(t_j^-) = \bar{u}(t_j^-) \quad \text{and} \quad u(t_j^+) = \bar{u}(t_j^+).$$

In the sequel we will need to control the error between a function u and its interpolation $\bar{u}$. This is the content of the following propositions, where $\|\cdot\|_\infty$ denotes the sup norm on $[0,1]$.

Proposition 8.3.1. *For all $u \in \mathcal{C}^{k+1}_{\Delta_m}([0,1],\mathbb{R})$,*

$$\left\|(Id - \Pi^1_{m,k})u\right\|_\infty = \|u - \bar{u}\|_\infty \leq C_k \max_{0\leq j<m} \left((t_{j+1}-t_j)^{k+1} \max_{t\in[t_j,t_{j+1}]} \left| \frac{d^{k+1}u}{dt^{k+1}}(t) \right| \right),$$

where

$$C_k := \frac{1}{(k+1)!2^{2k}}.$$

Proposition 8.3.2. *Fix $l \in \mathbb{N}$ such that $1 \leq l \leq k$. For all $u \in \mathcal{C}^l_{\Delta_m}([0,1],\mathbb{R})$,*

$$\left\|(Id - \Pi^1_{m,k})u\right\|_\infty = \|u - \bar{u}\|_\infty \leq \tilde{C}_{k,l} \max_{0\leq j<m} \left((t_{j+1}-t_j)^{l} \max_{t\in[t_j,t_{j+1}]} \left| \frac{d^{l}u}{dt^{l}}(t) \right| \right),$$

where

$$\tilde{C}_{k,l} := \min \left[(1+\Lambda_k) \left(\frac{\pi}{4}\right)^l \frac{(k+1-l)!}{(k+1)!}, \frac{1}{l!2^l} \sum_{q=0}^{\left[\frac{l-1}{2}\right]} \frac{1}{4^q} \binom{l-1}{2q} \binom{2q}{q} \right],$$

Λ_k being the Lebesgue constant (see for instance [199]), and $\left[\frac{l-1}{2}\right]$ denoting the integer part of $\frac{l-1}{2}$.

Remark 8.3.3. *More information about the Lebesgue constant, and in particular sharp computable upper bounds for it, can be found in the Appendix, together with references and proofs of the two above propositions.*

8.3.2 Finite dimensional projection

To get an approximate zero of G (and thus an approximate solution of (8.2)), we are going to look for a zero of $\bar{G} := \Pi^n_{m,k} G_{|S^n_{m,k}}$. But first, we need a convenient way to represent the elements of $S^n_{m,k}$. Here and in the sequel, we use the exponent $^{(i)}$ to denote the i-th component of a vector in $\mathbb{R}^n$, but we will work with all the components at once as often as possible to avoid burdening the notations with this exponent $^{(i)}$. Let us introduce the set of indexes

$$\mathcal{E}^n_{m,k} := \left\{ (i,j,l) \in \mathbb{N}^3,\ 1 \leq i \leq n,\ 0 \leq j \leq m-1,\ 0 \leq l \leq k \right\}.$$

Perhaps the most natural way to characterize an element $\bar{u}$ of $S^n_{m,k}$ is to give all the values $\bar{u}^{(i)}(t_{j,l})$ for $(i,j,l) \in \mathcal{E}^n_{m,k}$. However, we will also use another representation, more suited to numerical computations, which consists of decomposing $\bar{u}$ on the Chebyshev basis. That is, we write

$$\bar{u}^{(i)}(t) = \sum_{l=0}^{k} \bar{u}^{(i)}_{j,l} T_l \left(\frac{t-t_j}{t_{j+1}-t_j} - \frac{t_{j+1}-t}{t_{j+1}-t_j} \right), \quad \text{for all } j = 0,\dots,m-1 \text{ and } t \in (t_j, t_{j+1}), \quad (8.7)$$

where T_l is the l-th Chebyshev polynomial of the first kind. We can thus also describe uniquely any function u belonging to $S^n_{m,k}$ by the family of Chebyshev coefficients $\left(\bar{u}^{(i)}_{j,l}\right)_{(i,j,l)\in\mathcal{E}^n_{m,k}}$.

Remark 8.3.4. *Let us mention how considering functions with possible discontinuities on the mesh points in Δ_m comes in handy. By restricting ourselves to functions in $\mathcal{C}^0([0,1],\mathbb{R}^n)$, we would need additional constraints on the Chebyshev coefficients to impose the continuity at each of the mesh point t_j ($j=1,\dots,m-1$) and keep track of them in all computations. Instead, the choice of working with $\mathcal{C}^0_{\Delta_m}([0,1],\mathbb{R}^n)$ allows avoiding these additional constraints.*

For $u \in \mathcal{C}^0_{\Delta_m}([0,1],\mathbb{R}^n)$ we have

$$G(u)\left(t_{j,l}\right) = \sum_{q=0}^{p-1} \tau^q \frac{(t_{j,l}-t_j)^q}{q!}\phi^{[q]}(u(t_j^-)) + \tau^p \int_{t_j}^{t_{j,l}} \frac{(t_{j,l}-s)^{p-1}}{(p-1)!}\phi^{[p]}(u(s))ds - u(t_{j,l}),$$
$$\text{for all } j=0,\dots,m-1 \text{ and } l=0,\dots,k. \tag{8.8}$$

We recall that all the values $G^{(i)}(u)\left(t_{j,l}\right)$ for $(i,j,l) \in \mathcal{E}^n_{k,m}$ uniquely characterize $\Pi^n_{m,k}G(u)$.

Using the isomorphisms $\bar{u} \mapsto \left(\bar{u}^{(i)}(t_{j,l})\right)_{(i,j,l)\in\mathcal{E}^n_{m,k}}$ and $\bar{u} \mapsto \left(\bar{u}^{(i)}_{j,l}\right)_{(i,j,l)\in\mathcal{E}^n_{m,k}}$ to identify $S^n_{m,k}$ and $\mathbb{R}^{nm(k+1)}$, we can in fact see $\bar{G} := \Pi^n_{m,k}G_{|S^n_{m,k}}$ as a function from $\mathbb{R}^{nm(k+1)}$ to itself, that associates to the coefficients $\left(\bar{u}^{(i)}_{j,l}\right)_{(i,j,l)\in\mathcal{E}^n_{m,k}}$ the values $\left(G^{(i)}(\bar{u})(t_{j,l})\right)_{(i,j,l)\in\mathcal{E}^n_{m,k}}$. Thus we can numerically find a zero $\bar{u}$ of $\bar{G}$, which is going to be our approximate solution. We note that we use these identifications between $S^n_{m,k}$ and $\mathbb{R}^{nm(k+1)}$ throughout the present work. Our objective is now to *validate* this numerical solution $\bar{u}$, that is to prove that within a given neighbourhood of $\bar{u}$ lies a true zero u of G.

8.3.3 Back to a fixed point formulation

We consider the space $\mathcal{X}^n := \mathcal{C}^0_{\Delta_m}([0,1],\mathbb{R}^n)$ and its decomposition $\mathcal{X}^n = \mathcal{X}^n_{m,k} \oplus \mathcal{X}^n_\infty$, where

$$\mathcal{X}^n_{m,k} := S^n_{m,k} \quad \text{and} \quad \mathcal{X}^n_\infty := (Id - \Pi^n_{m,k})\mathcal{C}^0_{\Delta_m}([0,1],\mathbb{R}^n).$$

We already have a projection onto $\mathcal{X}^n_{m,k}$

$$\Pi^n_{m,k} : \begin{cases} \mathcal{X}^n \to \mathcal{X}^n_{m,k} \\ \quad u \mapsto \bar{u} = \Pi^n_{m,k}(u), \end{cases}$$

and we also define its complementary

$$\Pi^n_\infty : \begin{cases} \mathcal{X}^n \to \mathcal{X}^n_\infty \\ \quad u \mapsto \Pi^n_\infty(u) = u - \bar{u} = (Id - \Pi^n_{m,k})(u). \end{cases}$$

We then define the norms

$$\left\|\Pi^n_{m,k}(u)\right\|_{\mathcal{X}^n_{m,k}} := \max_{(i,j,l)\in\mathcal{E}^n_{m,k}} \left|\bar{u}^{(i)}(t_{j,l})\right| \quad \text{and} \quad \|\Pi^n_\infty(u)\|_{\mathcal{X}^n_\infty} := \max_{1\le i\le n} \left\|u^{(i)} - \bar{u}^{(i)}\right\|_\infty.$$

On $\mathcal{X}^n$ we consider the norm

$$\|u\|_{\mathcal{X}^n} := \max\left(\left\|\Pi^n_{m,k}(u)\right\|_{\mathcal{X}^n_{m,k}}, \frac{1}{r_\infty}\|\Pi^n_\infty(u)\|_{\mathcal{X}^n_\infty}\right),$$

where r_∞ is a positive parameter. Notice that for all $r_\infty > 0$, $(\mathcal{X}^n, \|\cdot\|_{\mathcal{X}^n})$ is a Banach space. For any $r, r_\infty > 0$, we denote by $B_{\mathcal{X}^n}(r, r_\infty)$ the closed neighbourhood of 0 defined as

$$B_{\mathcal{X}^n}(r, r_\infty) = \left\{u \in \mathcal{X}, \|u\|_{\mathcal{X}^n} \le r\right\}.$$

Suppose that we now have computed a numerical zero $\bar{u}$ of $\bar{G}$. We define $A^{\dagger}_{m,k} = D\bar{G}(\bar{u})$ and consider $A_{m,k}$ an injective numerical inverse of $A^{\dagger}_{m,k}$. Finally, we introduce the *Newton-like* operator $T : \mathcal{X}^n \to \mathcal{X}^n$ defined by

$$T(u) := \left(\Pi^n_{m,k} - A_{m,k}\Pi^n_{m,k}G\right) u + \Pi^n_{\infty}\left(G(u) + u\right).$$

Notice that the fixed points of T are in one-to-one correspondence with the zeros of G. We now give a finite set of sufficient conditions, that can be rigorously checked on a computer using interval arithmetic, to ensure that T is a contraction on a given ball around $\bar{u}$. If those conditions are satisfied, the Banach fixed point theorem then yields the existence and local uniqueness of a zero of G. This is the content of the following statement (based on [214], see also [109] for a detailed proof).

Theorem 8.3.5. *Let*

$$y := T(\bar{u}) - \bar{u}, \tag{8.9}$$

and

$$z = z(u_1, u_2) := DT(\bar{u} + u_1)u_2, \quad \textit{for all } u_1, u_2 \in B_{\mathcal{X}^n}(r, r_\infty). \tag{8.10}$$

Assume that we have bounds Y and $Z(r, r_\infty)$ satisfying

$$\left|\left(\Pi^n_{m,k}y\right)^{(i)}_{j,l}\right| \leq Y^{(i)}_{j,l}, \quad \textit{for all } (i,j,l) \in \mathcal{E}^n_{m,k}, \tag{8.11}$$

$$\left\|(\Pi^n_{\infty}y)^{(i)}\right\|_{\infty} \leq Y^{(i)}_{\infty}, \quad \textit{for all } 1 \leq i \leq n, \tag{8.12}$$

$$\sup_{u_1,u_2 \in B_{\mathcal{X}^n}(r,r_\infty)} \left|\left(\Pi^n_{m,k}z(u_1,u_2)\right)^{(i)}_{j,l}\right| \leq Z^{(i)}_{j,l}(r, r_\infty), \quad \textit{for all } (i,j,l) \in \mathcal{E}^n_{m,k}, \tag{8.13}$$

and

$$\sup_{u_1,u_2 \in B_{\mathcal{X}^n}(r,r_\infty)} \left\|(\Pi^n_{\infty}z(u_1,u_2))^{(i)}\right\|_{\infty} \leq Z^{(i)}_{\infty}(r, r_\infty), \quad \textit{for all } 1 \leq i \leq n. \tag{8.14}$$

If there exist $r, r_\infty > 0$ such that

$$p^{(i)}_{j,l}(r, r_\infty) := Y^{(i)}_{j,l} + Z^{(i)}_{j,l}(r, r_\infty) - r < 0, \quad \textit{for all } (i,j,l) \in \mathcal{E}^n_{m,k} \tag{8.15}$$

and

$$p^{(i)}_{\infty}(r, r_\infty) := Y^{(i)}_{\infty} + Z^{(i)}_{\infty}(r, r_\infty) - r_\infty r < 0, \quad \textit{for all } 1 \leq i \leq n, \tag{8.16}$$

then there exists a unique zero of G within the set $\bar{u} + B_{\mathcal{X}^n}(r, r_\infty) \subset \mathcal{X}^n$.

The quantities $p^{(i)}_{j,l}(r, r_\infty)$ and $p^{(i)}_{\infty}(r, r_\infty)$ given respectively in (8.15) and (8.16) are called the *radii polynomials*.

In the next section, we show how to obtain bounds Y and Z satisfying (8.11)-(8.14). Before doing so, let us make a quick remark about the different representations and norms we can use on $\mathcal{X}^n_{m,k}$.

Remark 8.3.6. *As explained in Section 8.3.2, in practice we will mostly work with $\bar{u} \in \mathcal{X}^n_{m,k}$ represented by its Chebyshev coefficients as in* (8.7)*. However, there are going to be instances where the values $\bar{u}^{(i)}(t_{j,l})$ are needed, for instance to compute $\|\bar{u}\|_{\mathcal{X}^n_{m,k}}$. We point out that numerically, passing from one representation to the other can be done easily by using the Fast Fourier Transform.*

One other important point is that, at some point in the next section we are going to need upper bounds for $\left\|\bar{u}^{(i)}\right\|_{\infty}$. To get such a bound from our finite dimensional data, we have two options, namely

$$\max_{t \in [t_j, t_{j+1}]} \left|\bar{u}^{(i)}(t)\right| \leq \sum_{l=0}^{k} \left|\bar{u}^{(i)}_{j,l}\right|, \quad \textit{for all } j = 0, \ldots, m-1, \tag{8.17}$$

or

$$\max_{t\in[t_j,t_{j+1}]}\left|\bar{u}^{(i)}(t)\right| \leq \Lambda_k \max_{0\leq l\leq k}\left|\bar{u}^{(i)}(t_{j,l})\right|, \quad \textit{for all } j=0,\dots,m-1. \tag{8.18}$$

If $\bar{u}$ is given, then (8.17) *is usually better, whereas* (8.18) *is better if $\bar{u}$ is any function in a given ball of $\mathcal{X}^n_{m,k}$. Notice that* (8.17) *simply follows from the fact that the Chebyshev polynomials satisfy $|T_l(t)|\leq 1$ for all $t\in[-1,1]$ and all $l\in\mathbb{N}$. For more information about the bound* (8.18)*, see the Appendix and the references therein.*

8.4 Formula for the bounds

In this section, we give formulas for $Y^{(i)}_{j,l}$, $Y^{(i)}_{\infty}$, $Z^{(i)}_{j,l}$ and $Z^{(i)}_{\infty}$ satisfying the assumptions (8.11)-(8.14) of Theorem 8.3.5. To make the exposition clearer, we focus strictly on the derivation of the different bounds in this section. In particular, the discussion about the impact of the level of an priori bootstrap (that is the value of p) and the order of polynomial approximation (that is the value of k) is done in Section 8.5.

8.4.1 The Y bounds

In this section we derive the Y bounds, which measure the *defect* associated with a numerical solution $\bar{u}$, that is how close $G(\bar{u})$ is to 0. We start by the *finite dimensional* part.

Proposition 8.4.1. *Let y be defined as in* (8.9) *and consider*

$$Y^{(i)}_{j,l} \geq \left|\left(A_{m,k}\bar{G}(\bar{u})\right)^{(i)}_{j,l}\right|, \quad \textit{for all } (i,j,l)\in\mathcal{E}^n_{m,k},$$

where $\bar{G}(\bar{u})$ is here seen as the vector $\left(G^{(i)}(\bar{u})(t_{j,l})\right)_{(i,j,l)\in\mathcal{E}^n_{m,k}}$. Then (8.11) *holds.*

Proof. Simply notice that $\Pi^n_{m,k}y = -A_{m,k}\Pi^n_{m,k}G(\bar{u})$. □

Remark 8.4.2. *The above bound is not completely satisfactory, in the sense that is not directly implementable. Indeed, to compute $Y^{(i)}_{j,l}$ we need to evaluate (or at least to bound) the quantities $G^{(i)}(\bar{u})(t_{j,l})$. In particular (see* (8.8)*), we need to evaluate the integrals*

$$\int_{t_j}^{t_{j,l}}(t_{j,l}-s)^{p-1}\phi^{[p]}(\bar{u}(s))ds = \left(\frac{t_{j+1}-t_j}{2}\right)^p\int_{-1}^{x^k_l}(x^k_l-s)^p\Psi(s)ds,$$

where

$$\Psi(s) := \phi^{[p]}\left(\sum_{l=0}^{k}\bar{u}_{j,l}T_l(s)\right).$$

If ϕ is a non polynomial vector field, we use a Taylor approximation of order k_0 of Ψ to get an approximate value of the integral by computing

$$\sum_{l=0}^{k_0}\frac{1}{l!}\frac{d^l\Psi}{ds^l}(0)\int_{-1}^{x^k_l}(x^k_l-s)^p s^l ds.$$

Notice that this quantity can be evaluated explicitly. The error made in this approximation is then controlled as follows

$$\left\|\int_{-1}^{x^k_l}(x^k_l-s)^p\Psi(s)ds - \sum_{l=0}^{k_0}\frac{1}{l!}\frac{d^l\Psi}{ds^l}(0)\int_{-1}^{x^k_l}(x^k_l-s)^p s^l ds\right\| \leq$$

$$\frac{1}{(k_0+1)!}\max_{s\in[-1,1]}\left\|\frac{d^{k_0+1}\Psi}{ds^{k_0+1}}(s)\right\|\int_{-1}^{x^k_l}(x^k_l-s)^p|s|^{k_0+1}ds.$$

Notice that this error term is effective, since $\max_{s\in[-1,1]} \left\| \frac{d^{k_0+1}\Psi}{ds^{k_0+1}}(s) \right\|$ *can be bounded using interval arithmetic. Therefore, the quantity* $Y^{(i)}_{j,l}$ *that we end up implementing is of the form*

$$Y^{(i)}_{j,l} = \left| \left(A_{m,k} \hat{G}(\bar{u}) \right)^{(i)}_{j,l} \right| + \left| \left(|A_{m,k}|\, G_\epsilon(\bar{u}) \right)^{(i)}_{j,l} \right|,$$

where the vector $\hat{G}(\bar{u})$ *contains the approximate integrals and the vector* $G_\epsilon(\bar{u})$ *contains the errors bounds for these approximations. Here and in the sequel, absolute values applied to a matrix, like* $|A_{m,k}|$*, must be understood component-wise. We point out that, in practice, if the mesh* Δ_m *is refined enough, then* $\bar{u}$ *is not going to be varying much on each subinterval* $[t_j, t_{j+1}]$*, and thus we can get rather precise approximations even with a lower order* k_0 *for the Taylor expansion.*

We mention that when the vector field ϕ *is polynomial,* Ψ *has a finite Taylor expansion, therefore up the integrals can in fact be computed exactly (i.e. we can get* $G_\epsilon(\bar{u}) = 0$*).*

We now turn our attention to the second part of the Y bound.

Proposition 8.4.3. *Let* y *be defined as in* (8.9) *and consider*

$$Y^{(i)}_\infty \geq C_k \tau^p \max_{0\leq j<m} \left((t_{j+1}-t_j)^{k+1} \max_{t\in[t_j,t_{j+1}]} \left| \frac{d^{k+1-p}}{dt^{k+1-p}} (\phi^{[p]})^{(i)}(\bar{u}(t)) \right| \right), \quad \text{for all } 1 \leq i \leq n.$$

Then (8.12) *holds.*

Proof. We have $\Pi^n_\infty y = \Pi^n_\infty (G(\bar{u}) + \bar{u}) = \Pi^n_\infty g(\bar{u})$. Since $\frac{d^p}{dt^p} g(\bar{u}) = \tau^p \phi^{[p]}(\bar{u})$, we have that $\frac{d^{k+1}}{dt^{k+1}} g(\bar{u}) = \tau^p \frac{d^{k+1-p}}{dt^{k+1-p}} \phi^{[p]}(\bar{u})$ and Proposition 8.3.1 yields

$$\left\| (\Pi^n_\infty y)^{(i)} \right\|_\infty \leq C_k \tau^p \max_{0\leq j<m} \left((t_{j+1}-t_j)^{k+1} \max_{t\in[t_j,t_{j+1}]} \left| \frac{d^{k+1-p}}{dt^{k+1-p}} (\phi^{[p]})^{(i)}(\bar{u}(t)) \right| \right). \qquad \square$$

Remark 8.4.4. *As comment similar to the one of Remark 8.4.2 applies here. Indeed, the bound given in Proposition 8.4.3 is not directly implementable because of the term*

$$\max_{t\in[t_j,t_{j+1}]} \left| \frac{d^{k+1-p}}{dt^{k+1-p}} (\phi^{[p]})^{(i)}(\bar{u}(t)) \right|,$$

but we can again get an explicit bound for this quantity by using a low order Taylor approximation and interval arithmetic. In the particular case where the vector field ϕ *is polynomial, an explicit bound can also be obtained* via *the Chebyshev coefficients of the polynomial* $\frac{d^{k+1-p}}{dt^{k+1-p}} (\phi^{[p]})^{(i)}(\bar{u})$*, as in* (8.17)*.*

8.4.2 The *Z* bounds

In this section we derive the Z bounds, which measure the contraction rate of T on the ball of radius r around $\bar{u}$. We begin with the finite dimensional part, that is the projection on $\mathcal{X}^n_{m,k}$. Let z be defined as in (8.10). Then

$$\begin{aligned}
\Pi^n_{m,k} z &= \Pi^n_{m,k} \left(DT(\bar{u}+u_1)u_2 \right) \\
&= \Pi^n_{m,k} u_2 - A_{m,k} \Pi^n_{m,k} \left(DG(\bar{u}+u_1)u_2 \right) \\
&= \Pi^n_{m,k} u_2 - A_{m,k} D\Pi^n_{m,k} G(\bar{u}+u_1)u_2 \\
&= \left(Id - A_{m,k} A^\dagger_{m,k} \right) \Pi^n_{m,k} u_2 - A_{m,k} \left(D\Pi^n_{m,k} G(\bar{u}+u_1)u_2 - A^\dagger_{m,k} \Pi^n_{m,k} u_2 \right) \\
&= \left(Id - A_{m,k} A^\dagger_{m,k} \right) \Pi^n_{m,k} u_2 - A_{m,k} \left(D\Pi^n_{m,k} G(\bar{u})u_2 - A^\dagger_{m,k} \Pi^n_{m,k} u_2 \right) \\
&\quad - A_{m,k} \left(D\Pi^n_{m,k} G(\bar{u}+u_1) - D\Pi^n_{m,k} G(\bar{u}) \right) u_2,
\end{aligned}$$

where $A_{m,k}$ and $A^\dagger_{m,k}$ are defined as in Section 8.3.3. We are going to bound each term separately as

$$\left|\left(\Pi^n_{m,k}z\right)^{(i)}_{j,l}\right| \leq \underbrace{\left|\left(\left(Id - A_{m,k}A^\dagger_{m,k}\right)\Pi^n_{m,k}u_2\right)^{(i)}_{j,l}\right|}_{\leq (Z_0(r))^{(i)}_{j,l}} + \underbrace{\left|\left(A_{m,k}\left(D\Pi^n_{m,k}G(\bar{u})u_2 - A^\dagger_{m,k}\Pi^n_{m,k}u_2\right)\right)^{(i)}_{j,l}\right|}_{\leq (Z_1(r,r_\infty))^{(i)}_{j,l}}$$
$$+ \underbrace{\left|\left(A_{m,k}\left(D\Pi^n_{m,k}G(\bar{u}+u_1) - D\Pi^n_{m,k}G(\bar{u})\right)u_2\right)^{(i)}_{j,l}\right|}_{\leq (Z_2(r,r_\infty))^{(i)}_{j,l}}. \tag{8.19}$$

The bound $Z_0(r)$

The computation of the bounds $(Z_0(r))^{(i)}_{j,l}$ estimating the first of the terms in the splitting (8.19) is rather straightforward and is simply a control on the precision of the numerical inverse.

Proposition 8.4.5. *Let $u_2 \in B_{\mathcal{X}^n}(r, r_\infty)$, define the vector $\mathbf{1}_{\mathcal{X}^n_{m,k}} \in \mathbb{R}^{nm(k+1)}$ by $\left(\mathbf{1}_{\mathcal{X}^n_{m,k}}\right)^{(i)}_{j,l} = 1$ for all $(i,j,l) \in \mathcal{E}^n_{m,k}$ and let*

$$(Z_0(r))^{(i)}_{j,l} := \left(\left|Id - A_{m,k}A^\dagger_{m,k}\right|\mathbf{1}_{\mathcal{X}^n_{m,k}}\right)^{(i)}_{j,l} r, \quad \text{for all } (i,j,l) \in \mathcal{E}^n_{m,k}.$$

Then,

$$\left|\left(\left(Id - A_{m,k}A^\dagger_{m,k}\right)\Pi^n_{m,k}u_2\right)^{(i)}_{j,l}\right| \leq (Z_0(r))^{(i)}_{j,l}, \quad \text{for all } (i,j,l) \in \mathcal{E}^n_{m,k}.$$

The bound $Z_1(r, r_\infty)$

We now construct the bounds $(Z_1(r, r_\infty))^{(i)}_{j,l}$ estimating the second term in the splitting (8.19).

Proposition 8.4.6. *Let $u_2 \in B_{\mathcal{X}^n}(r, r_\infty)$, consider $\rho = \left(\rho^{(i)}_{j,l}\right)_{(i,j,l)\in\mathcal{E}^n_{m,k}}$ such that*

$$\rho^{(i)}_{j,l} \geq r_\infty r \frac{\tau^p}{p!}(t_{j,l} - t_j)^p \max_{s\in[t_j,t_{j+1}]} \left|D(\phi^{[p]})^{(i)}(\bar{u}(s))\right|\mathbf{1}_n, \quad \text{for all } (i,j,l) \in \mathcal{E}^n_{m,k},$$

where $\mathbf{1}_n$ is the vector of size n whose components all are equal to 1. Let

$$Z_1(r, r_\infty) := |A_{m,k}|\,\rho.$$

Then

$$\left|\left(A_{m,k}\left(D\Pi^n_{m,k}G(\bar{u})u_2 - A^\dagger_{m,k}\Pi^n_{m,k}u_2\right)\right)^{(i)}_{j,l}\right| \leq (Z_1(r, r_\infty))^{(i)}_{j,l}, \quad \text{for all } (i,j,l) \in \mathcal{E}^n_{m,k}.$$

Proof. By definition of $A^\dagger_{m,k}$ and $\bar{G}$, we have that

$$D\Pi^n_{m,k}G(\bar{u})\Pi^n_{m,k}u_2 = A^\dagger_{m,k}\Pi^n_{m,k}u_2, \quad \text{for all } u_2 \in \mathcal{X}^n.$$

Therefore, we can rewrite

$$\begin{aligned} A_{m,k}\left(D\Pi^n_{m,k}G(\bar{u})u_2 - A^\dagger_{m,k}\Pi^n_{m,k}u_2\right) &= A_{m,k}D\Pi^n_{m,k}G(\bar{u})\left(u_2 - \Pi^n_{m,k}u_2\right) \\ &= A_{m,k}D\Pi^n_{m,k}G(\bar{u})\Pi^n_\infty u_2, \end{aligned}$$

and we only need to prove that

$$\left|D\Pi^n_{m,k}G(\bar{u})\Pi^n_\infty u_2\right|^{(i)}_{j,l} \leq \rho^{(i)}_{j,l}, \quad \text{for all } (i,j,l) \in \mathcal{E}^n_{m,k}.$$

Remembering (8.8) and using that $\|\Pi^n_\infty u_2\|_{\mathcal{X}^n_\infty} \leq r_\infty r$, we estimate for all $(i,j,l) \in \mathcal{E}^n_{m,k}$,

$$\begin{aligned}\left|D\Pi^n_{m,k}G(\bar{u})\Pi^n_\infty u_2\right|^{(i)}_{j,l} &\leq r_\infty r\tau^p \int_{t_j}^{t_{j,l}} \frac{(t_{j,l}-s)^{p-1}}{(p-1)!}\left|D(\phi^{[p]})^{(i)}(\bar{u}(s))\right|\mathbf{1}_n ds \\ &\leq r_\infty r\frac{\tau^p}{p!}(t_{j,l}-t_j)^p \max_{s\in[t_j,t_{j+1}]}\left|D(\phi^{[p]})^{(i)}(\bar{u}(s))\right|\mathbf{1}_n.\end{aligned}$$

□

Notice that Remark 8.4.4 also applies here.

Remark 8.4.7. *Had we used the operator $\tilde{g}$ (see* (8.5)*) instead of g (see* (8.6)*), we would have gotten a bound like*

$$\left|D\Pi^n_{m,k}G(\bar{u})\Pi^n_\infty u_2\right|^{(i)}_{j,l} \leq r_\infty r\tau^p \int_{0}^{t_{j,l}} \frac{(t_{j,l}-s)^{p-1}}{(p-1)!}\left|D(\phi^{[p]})^{(i)}(\bar{u}(s))\right|\mathbf{1}_n ds,$$

which is obviously worst because one has to consider the whole integral from 0 to $t_{j,l}$ instead of just from t_j to $t_{j,l}$.

The bound Z_2

We finally construct the bounds $(Z_2(r,r_\infty))^{(i)}_{j,l}$ estimating the last term in the splitting (8.19).

Proposition 8.4.8. *Let $u_1, u_2 \in B_{\mathcal{X}^n}(r,r_\infty)$. consider $\varrho = \left(\varrho^{(i)}_{j,l}\right)_{(i,j,l)\in\mathcal{E}^n_{m,k}}$ such that*

$$\begin{aligned}\varrho^{(i)}_{j,l} \geq &\sum_{q=1}^{p-1}\frac{\tau^q}{q!}(t_{j,l}-t_j)^q \sum_{\delta=0}^{q(d-1)-1}\frac{1}{(1+\delta)!}\left|D^{2+\delta}(\phi^{[q]})^{(i)}(\bar{u}(t_j^-))\right|\left(\mathbf{1}_n^{2+\delta}\right) r^{2+\delta} \\ &+ \frac{\tau^p}{p!}(t_{j,l}-t_j)^p \sum_{\delta=0}^{p(d-1)-1}\frac{1}{(1+\delta)!}\max_{s\in[t_j,t_{j+1}]}\left|D^{2+\delta}(\phi^{[p]})^{(i)}(\bar{u}(s))\right|\left(\mathbf{1}_n^{2+\delta}\right)((\Lambda_k+r_\infty)r)^{2+\delta}.\end{aligned}$$

Let

$$Z_2(r,r_\infty) := |A_{m,k}|\,\varrho.$$

Then

$$\left|\left(A_{m,k}\left(D\Pi^n_{m,k}G(\bar{u}+u_1)-D\Pi^n_{m,k}G(\bar{u})\right)u_2\right)^{(i)}_{j,l}\right| \leq (Z_2(r,r_\infty))^{(i)}_{j,l}, \quad \text{for all } (i,j,l)\in\mathcal{E}^n_{m,k}.$$

Remark 8.4.9. *In the above proposition, $\left|D^{2+\delta}(\phi^{[p]})^{(i)}(\bar{u}(s))\right|\left(\mathbf{1}_n^{2+\delta}\right)$ must be understood as the evaluation of the $(2+\delta)$-linear form $\left|D^{2+\delta}(\phi^{[p]})^{(i)}(\bar{u}(s))\right|$ at the vectors $(\mathbf{1}_n,\dots,\mathbf{1}_n)$, that is*

$$\left|D^{2+\delta}(\phi^{[p]})^{(i)}(\bar{u}(s))\right|\left(\mathbf{1}_n^{2+\delta}\right) = \sum_{1\leq j_1,\dots,j_{\delta+2}\leq n}\left|\partial_{j_1\dots j_{\delta+2}}(\phi^{[p]})^{(i)}(\bar{u}(s))\right|.$$

Proof. (*of Proposition 8.4.8*) We only have to prove that

$$\left|\left(\left(D\Pi^n_{m,k}G(\bar{u}+u_1)-D\Pi^n_{m,k}G(\bar{u})\right)u_2\right)^{(i)}_{j,l}\right| \leq \varrho^{(i)}_{j,l}, \quad \text{for all } (i,j,l)\in\mathcal{E}^n_{m,k}.$$

Using (8.18) we have that

$$\|u_2^{(i)}\|_\infty \leq \|\Pi^1_{m,k} u_2^{(i)}\|_\infty + \|\Pi^1_\infty u_2^{(i)}\|_\infty \leq (\Lambda_k + r_\infty) r.$$

Then we estimate for all $(i,j,l) \in \mathcal{E}^n_{m,k}$,

$$\begin{aligned}
&\left|\left(D\Pi^n_{m,k} G(\bar{u}+u_1) - D\Pi^n_{m,k} G(\bar{u})\right) u_2\right|^{(i)}_{j,l} \\
&\leq \sum_{q=1}^{p-1} \frac{\tau^q}{q!}(t_{j,l}-t_j)^q \left|\left(D(\phi^{[q]})^{(i)}(\bar{u}(t_j^-)+u_1(t_j^-)) - D(\phi^{[q]})^{(i)}(\bar{u}(t_j^-))\right)(u_2(t_j^-))\right| \\
&\quad + \tau^p \int_{t_j}^{t_{j,l}} \frac{(t_{j,l}-s)^{p-1}}{(p-1)!} \left|\left(D(\phi^{[p]})^{(i)}(\bar{u}(s)+u_1(s)) - D(\phi^{[p]})^{(i)}(\bar{u}(s))\right)(u_2(s))\right| ds \\
&\leq \sum_{q=1}^{p-1} \frac{\tau^q}{q!}(t_{j,l}-t_j)^q \sum_{\delta=0}^{q(d-1)-1} \frac{1}{(1+\delta)!} \left|D^{2+\delta}(\phi^{[q]})^{(i)}(\bar{u}(t_j^-))\right| \left(\mathbf{1}_n^{2+\delta}\right) r^{2+\delta} \\
&\quad + \frac{\tau^p}{p!}(t_{j,l}-t_j)^p \sum_{\delta=0}^{p(d-1)-1} \frac{1}{(1+\delta)!} \max_{s\in[t_j,t_{j+1}]} \left|D^{2+\delta}(\phi^{[p]})^{(i)}(\bar{u}(s))\right| \left(\mathbf{1}_n^{2+\delta}\right) ((\Lambda_k+r_\infty)r)^{2+\delta}. \ \square
\end{aligned}$$

Notice that Remark 8.4.4 also applies here.

The Z_∞ bound

We are left with the *remainder part* of the Z bound, which we treat in this section.

Proposition 8.4.10. *Let $u_1, u_2 \in B_{\mathcal{X}^n}(r, r_\infty)$ and z as in* (8.10)*. Define for all $i \in \{1,\dots,n\}$*

$$\begin{aligned}
&Z^{(i)}_\infty(r, r_\infty) \geq \\
&\quad \tau^p C^{opt}_{k,p} \sum_{\delta=0}^{p(d-1)} \max_{0\leq j<m} \left((t_{j+1}-t_j)^p \frac{1}{\delta!} \max_{t\in[t_j,t_{j+1}]} \left|D^{1+\delta}(\phi^{[p]})^{(i)}(\bar{u}(t))\right| \left(\mathbf{1}_n^{1+\delta}\right)\right) ((\Lambda_k+r_\infty)r)^{1+\delta},
\end{aligned}$$

where $C^{opt}_{k,p}$ is one of the two constants given by Propositions 8.3.1 and 8.3.2, namely

$$C^{opt}_{k,p} = \begin{cases} C_k & \text{if } p = k+1, \\ \tilde{C}_{k,p} & \text{if } p \leq k. \end{cases}$$

Then (8.14) *holds.*

Proof. We need to estimate

$$\Pi^n_\infty z = \Pi^n_\infty \left(DT(\bar{u}+u_1)u_2\right) = \Pi^n_\infty \left(Dg(\bar{u}+u_1)u_2\right).$$

For any continuous function γ, one has

$$\frac{d^p}{dt^p} \int_{t_j}^t \frac{(t-s)^{p-1}}{(p-1)!} \gamma(s) ds = \gamma(t),$$

thus we get, for all $1 \leq i \leq n$

$$\begin{aligned}
&\left\|\Pi^n_\infty \left(Dg^{(i)}(\bar{u}+u_1)u_2\right)\right\|_\infty \\
&= \left\|\Pi^n_\infty \left(t \mapsto \tau^p \int_{t_j}^t \frac{(t-s)^{p-1}}{(p-1)!} D(\phi^{[p]})^{(i)}(\bar{u}(s)+u_1(s))u_2(s) ds\right)\right\|_\infty \\
&\leq \tau^p C^{opt}_{k,p} \max_{0\leq j<m} \left((t_{j+1}-t_j)^p \max_{t\in[t_j,t_{j+1}]} \left|D(\phi^{[p]})^{(i)}(\bar{u}(t)+u_1(t))u_2(t)\right|\right) \\
&\leq \tau^p C^{opt}_{k,p} \sum_{\delta=0}^{p(d-1)} \max_{0\leq j<m} \left((t_{j+1}-t_j)^p \frac{1}{\delta!} \max_{t\in[t_j,t_{j+1}]} \left|D^{1+\delta}(\phi^{[p]})^{(i)}(\bar{u}(t))\right| \left(\mathbf{1}_n^{1+\delta}\right)\right) ((\Lambda_k+r_\infty)r)^{1+\delta}. \square
\end{aligned}$$

Notice that Remark 8.4.4 also applies here.

8.4.3 The radii polynomials and interval arithmetics

The following proposition sums up what has been proven up to now in this section, namely that we have derived bounds that satisfy the requirements (8.11) to (8.14) from Theorem 8.3.5.

Proposition 8.4.11. *Let y and z defined as in* (8.9) *and* (8.10)*. Then, the bound defined in Proposition 8.4.1 satisfies* (8.11) *and the one from Proposition 8.4.3 satisfies* (8.12)*. Also, consider the bounds defined in Propositions 8.4.5 to 8.4.8. Then*

$$Z^{(i)}_{j,l}(r, r_\infty) = (Z_0(r))^{(i)}_{j,l} + (Z_1(r, r_\infty))^{(i)}_{j,l} + (Z_2(r, r_\infty))^{(i)}_{j,l},$$

satisfies (8.13) *and finally the bound from Proposition 8.4.10 satisfies* (8.14)*.*

Notice that, the way these bounds are defined, they are polynomials in r and r_∞, whose coefficients are all positive and can be computed explicitly with the help of the computer, since they depend on the numerical data of an approximate solution $\bar{u}$. Also, we make sure to control possible round-off errors by using interval arithmetic (in our case INTLAB [190]).

In practice, we first consider r_∞ so that it satisfies the constraint (8.20) introduced in the next section. If there does not exist such positive r_∞, we increase m and/or k and/or p and try again. Once r_∞ is fixed, we try to find a positive r such that the last conditions (8.15) and (8.16) of Theorem 8.3.5 hold. If there is no such positive r, we increase m and/or k and/or p and try again. If we finally find a positive r satisfying (8.15) and (8.16), then we have proven that Theorem 8.3.5 applies, that is there exists a unique zero of G in $B_{\mathcal{X}^n}(r, r_\infty)$.

In Sections 8.6 and 8.7, we give several examples where the procedure described just above is successfully used to validate solutions of an initial value problem, as well as periodic solutions and heteroclinic orbits. But before doing so, we discuss in the next section the role of the parameters k, m and p, and how they influence the bounds.

8.5 About the choice of the parameters

In this section, we explain how the parameters k, m and p should be chosen, and in particular we highlight how the *a priori bootstrap* (that is taking $p \geq 2$) helps improving the efficiency of the computer-assisted procedure we propose. The discussion will be rather informal, but we hope it helps the reader understand the results of the various comparisons presented in Section 8.6. Also, to make things slightly simpler we assume here that the grid Δ_m is uniform, therefore in the estimates each instance of $t_{j+1} - t_j$ can be replaced by $\frac{1}{m}$.

Our main constraint is that we want the method to be successful while minimizing the size of our numerical data, that is the dimension of our finite dimensional space $\mathcal{X}^n_{m,k}$, which is $nm(k+1)$. Since n is fixed by the dimension of the vector field ϕ, we want to minimize the product $m(k+1)$. As we see in the examples of Section 8.6, the usual limiting factor when trying to satisfy the radii polynomial inequalities (8.15) and (8.16) is to get the order one term (in r) to be negative. For the finite part (that is (8.15)), that means basically having that

$$r_\infty \frac{\alpha}{p!} \left(\frac{\tau}{m}\right)^p < 1,$$

(see Proposition 8.4.6), and for the remainder part (that is (8.16)) we get a condition like

$$C^{opt}_{k,p} \left(\frac{\tau}{m}\right)^p \beta(\Lambda_k + r_\infty) < r_\infty,$$

where α and β are constants depending on the numerical solution $\bar{u}$ and on the vector field ϕ, but not on the parameters k, m and p that we can tune. This leads to

$$\beta C_{k,p}^{opt}\Lambda_k\left(\frac{\tau}{m}\right)^p < \left(1-\beta C_{k,p}^{opt}\left(\frac{\tau}{m}\right)^p\right) r_\infty < \left(1-\beta C_{k,p}^{opt}\left(\frac{\tau}{m}\right)^p\right)\frac{p!}{\alpha}\left(\frac{m}{\tau}\right)^p. \tag{8.20}$$

We want to be able to chose a r_∞ satisfying the above inequalities, and a necessary and sufficient condition for that is

$$\beta C_{k,p}^{opt}\Lambda_k\left(\frac{\tau}{m}\right)^p < \left(1-\beta C_{k,p}^{opt}\left(\frac{\tau}{m}\right)^p\right)\frac{p!}{\alpha}\left(\frac{m}{\tau}\right)^p,$$

which we can rewrite

$$\left(\frac{\tau}{m}\right)^p C_{k,p}^{opt}\left(\beta+\frac{\alpha\beta}{p!}\Lambda_k\left(\frac{\tau}{m}\right)^p\right) < 1. \tag{8.21}$$

Remember that we want (8.21) to be satisfied, while minimizing the product $m(k+1)$. When p is fixed, and k becomes large, notice that $C_{k,p}^{opt}$ is decreasing like $\frac{\ln(k)}{k^p}$. However, satisfying (8.21) requires, roughly speaking, to decrease $(\frac{\tau}{m})^p C_{k,p}^{opt}$ as much as possible. This suggests two things, which we confirm in our explicit examples of Section 8.6. First, that it is slightly better to increase m than k (because of the $ln(k)$ factor) and second, that if we take p equal to 2 or more (that is if we use *a priori bootstrap*) then we can satisfy (8.21) while taking $m(k+1)$ much smaller than if we had p equal to 1.

Finally, we point out that taking $k=p-1$ seems optimal for the conditon (8.21) given by the order one term. Indeed, increasing k from $p-1$ to p increases the total number of coefficients, but brings no gain with respect to (8.21) since

$$C_{p-1,p}^{opt} = C_{p-1} < \tilde{C}_{p,p} = C_{p,p}^{opt}.$$

However, for the proof to succeed (that is for (8.15) and (8.16) to be satisfied) we also need small enough Y and Y_∞ bounds. Looking more precisely at Y_∞, we see that it is of the form

$$C_k\frac{1}{m^{k+1}}\gamma,$$

where γ depends on the numerical data $\bar{u}$ and also on k, but the dependency on k is way less important than in the $C_k\frac{1}{m^{k+1}}$ term, so we neglect it here. Looking back to the definition of C_k in Proposition 8.3.1, we see that the term that we want to be small is of the form

$$\frac{1}{(k+1)!}\frac{4}{(4m)^{k+1}}.$$

Therefore, if we really need to decrease the Y_∞ bound, increasing k is drastically better than increasing m. That is why, in practice we often take $k=p$, even though $k=p-1$ would be enough to satisfy (8.21). Finally we point out that, if we are not simply focused on getting an existence result, but also care about having sharp error bound, then we should definitively take care of having small Y and Y_∞ bounds, which, as we will show in the next section, can be achieved by slightly increasing k (that is taking $k>p$).

In the next section, we present several comparisons for different choices of parameters, that confirm the heuristic presented in this section.

8.6 Examples of applications for the Lorenz system

In this section, we consider the Lorenz system, that is

$$\phi(x,y,z) = \begin{pmatrix}\sigma(y-x)\\ \rho x - y - xz\\ -\beta z + xy\end{pmatrix},$$

with standard parameter values $(\sigma, \beta, \rho) = (10, \frac{8}{3}, 28)$. Here, we first consider the initial value problem (8.2), and use those bounds to try and validate orbits of various length with different parameters, to highlight the significant improvement made possible by the *a priori bootstrap* technique (that is taking $p \geq 2$). Then, we show that the *a priori bootstrap* also allows to validate more interesting solutions (from a dynamical point of view), namely periodic orbits and connecting orbits.

8.6.1 Comparisons for the initial value problem

The aim of this section is to showcase the improvements allowed by the use of *a priori bootstrap*, and to validate the heuristics made in Section 8.5. To do so, we fix an initial data (chosen close to the attractor of the Lorenz system)

$$u_0 = \begin{pmatrix} -14.68 \\ -11 \\ 37.67 \end{pmatrix}, \tag{8.22}$$

and do two kinds of comparisons. First, we try to validate the longest possible orbits for $p = 1, 2, 3$ at various values of m and k. We recall that by validating, we mean getting the existence of a true solution near a numerical one, by checking that the hypotheses of Theorem 8.3.5 hold. To make the comparison fair, we fix the total number of coefficients used for the numerical approximation, that is the dimension of $\mathcal{X}^n_{m,k}$, given by $nm(k+1)$. This quantity is usually the bottleneck of our approach, since we need to store and invert the matrix $A^\dagger_{m,k}$ which is of size $nm(k+1) \times nm(k+1)$. Here, we take $nm(k+1) = 14000$ (or as close as possible to 14000). The computations were made on a laptop with 8GB of RAM, and of course $nm(k+1)$ could be taken larger on a computer with more memory.

The first set of results are given in Table 8.6.1 (we recall that we work with the Lorenz system, therefore $n = 3$).

k	m	$nm(k+1)$	τ_{max}	r
1	2333	13998	0.69	2.3233×10^{-5}
2	1556	14004	0.64	1.1524×10^{-7}
3	1167	14004	0.58	8.6805×10^{-9}

Table 8.1 – Comparisons for $p = 1$. τ_{max} is the longest integration time for which the proof succeeds, and r is the associated validation radius, that is a bound of the distance (in $\mathcal{C}^0$ norm) between the numerical data used and the true solution.

In all cases, the proof fails for longer time τ, because (8.21) is no longer satisfied. We see here that, as announced in Section 8.5, it is better to take k as small as possible to get the longest possible orbit, but that increasing k helps reducing the Y_∞ bound, and thus the validation radius r. We see that simply increasing the order of the polynomial interpolation (given by k), allows to get better accuracy but does not really help to prove longer orbits. However, we are going to show on the next examples (see Table 8.2) that combining *a priori bootstrap* (that is taking $p \geq 2$) with higher order polynomial interpolation does allow to get much longer orbits.

First, comparing the $k = 1$ case when $p = 1$ and $p = 2$, we see that using *a priori bootstrap* allows to get a slightly longer orbit, even for linear interpolation. Also, even for the longest possible orbit in that case ($\tau = 0.97$), we still have much room to satisfy (8.21) (the quantity given by (8.21) is $\ll 1$). However, we cannot get a longer orbit in that case even with $p = 2$, because the Y_∞ bound becomes too large. This can be dealt with by increasing k, and we see that we can then get much longer orbits. To finish this set of comparisons, we show that doing

k	m	$nm(k+1)$	τ_{max}	r
1	2333	13998	0.97	1.5718×10^{-3}
2	1556	14004	5.6	8.4373×10^{-5}
3	1167	14004	5.5	8.4184×10^{-8}
4	933	13995	4.9	7.9190×10^{-9}

Table 8.2 – Comparisons for $p = 2$. τ_{max} is the longest integration time for which the proof succeeds, and r is the associated validation radius, that is a bound of the distance (in $\mathcal{C}^0$ norm) between the numerical data used and the true solution.

k	m	$nm(k+1)$	τ_{max}	r
2	1556	14004	5.6	7.6716×10^{-4}
3	1167	14004	8.1	9.3043×10^{-6}
4	933	13995	8.1	8.8204×10^{-8}
5	778	14004	8.0	1.6175×10^{-8}
9	467	14010	7.9	1.3748×10^{-8}
19	233	13980	6.9	2.2998×10^{-8}

Table 8.3 – Comparisons for $p = 3$. τ_{max} is the longest integration time for which the proof succeeds, and r is the associated validation radius, that is a bound of the distance (in $\mathcal{C}^0$ norm) between the numerical data used and the true solution.

one more iteration of the *a priori bootstrap* process (that is taking $p = 3$ instead of $p = 2$) still improves the results and allows to get longer orbits (see Table 8.6.1).

We sum up this set of comparisons by displaying the longest orbit obtained with $p = 1$, $p = 2$ and $p = 3$ (see Figure 8.1).

We then finish this section with another set of comparisons, where we now fix the length of the orbit, here $\tau = 2$, and instead look for the minimal total number of coefficients for which we can validate this orbit (for different values of p). The aim of this experiment is to show that using *a priori bootstrap* enables to validate solutions that one would not be able to validate without using it. Indeed, we are going to see that taking p greater than one allows us to use way less coefficients to validate the solutions. Thus, if for a given solution, the proof without *a priori bootstrap* requires more coefficients than what our computer can handle, one can reduce this number by using *a priori bootstrap* and then possibly validate the orbit. For instance, still with the initial condition given by (8.22), we cannot validate the orbit of length $\tau = 2$ without *a priori bootstrap* (that is with $p = 1$), at least not with less that 14000 coefficients. However, the next table of results shows that we can validate it with $p = 2$, and also using even less coefficients with $p = 3$.

$p = 2$	$k = 1$ no proof no proof	$k = 2$ $m = 416$ $nm(k+1) = 3744$	$k = 3$ $m = 415$ $nm(k+1) = 4980$	$k = 4$ $m = 377$ $nm(k+1) = 5655$
$p = 3$	$k = 2$ $m = 470$ $nm(k+1) = 4230$	$k = 3$ $m = 125$ $nm(k+1) = 1500$	$k = 4$ $m = 110$ $nm(k+1) = 1650$	$k = 5$ $m = 99$ $nm(k+1) = 1782$

Table 8.4 – Minimal number of coefficients needed to validate the orbit of length $\tau = 2$, starting from u_0 given in (8.22).

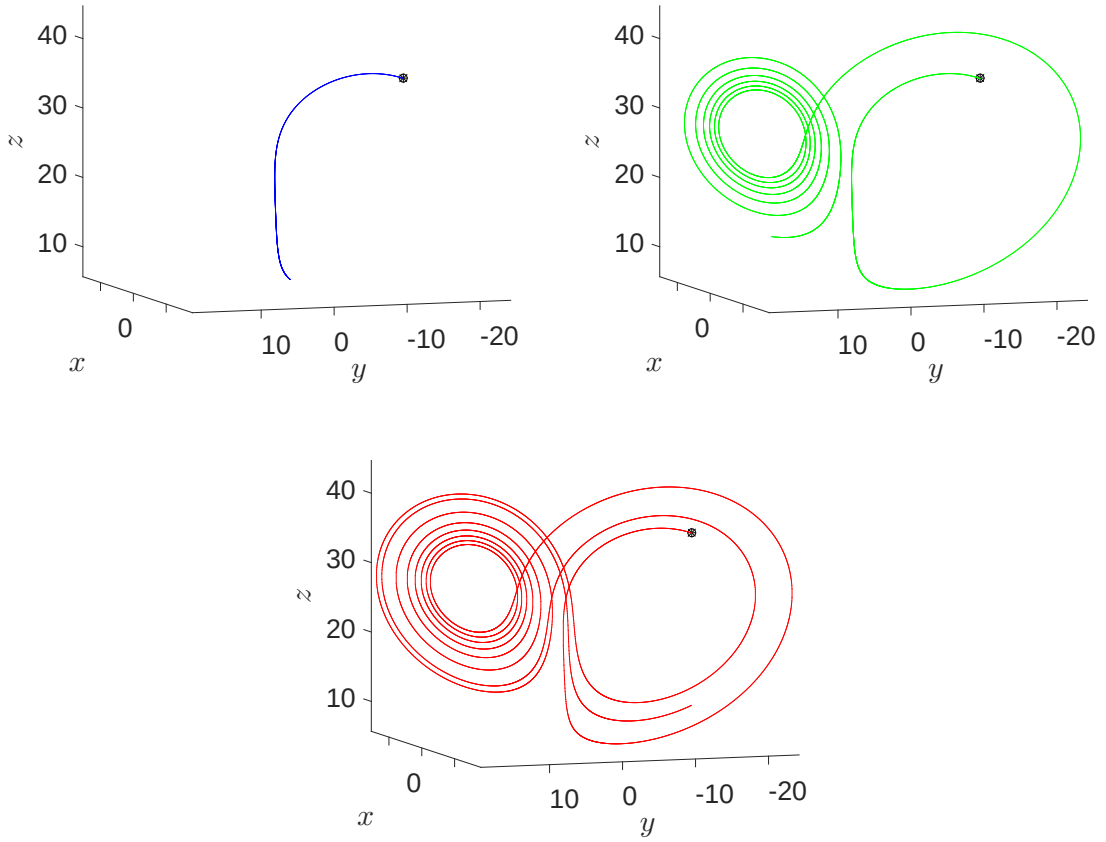


Figure 8.1 – The longest orbits we are able to validate, with a total number of coefficient of approximately 14000. In blue for $p = 1$, in green for $p = 2$ and in red for $p = 3$. The initial value is given by (8.22).

8.6.2 Validation of a periodic orbit.

To study periodic orbits, instead of an initial value problem, the system (8.1) has to be slightly modified into a boundary value problem

$$\begin{cases} u'(t) = \phi(u(t)), \quad t \in [0, \tau], \\ u(0) = u(\tau), \\ \langle u(0) - u_0, v_0 \rangle = 0, \end{cases} \tag{8.23}$$

where τ is now an unknown of the problem, and where $u_0, v_0 \in \mathbb{R}^n$. The last equation is sometimes called *Poincaré phase condition* and is here to isolate to periodic orbit.

As for the initial value problem, we can then consider an equivalent integral formulation (possibly with *a priori bootstrap*) and define an equivalent fixed point operator T very similar to the one introduced in Section 8.3. The additional phase condition and the fact that τ is now a variable only require minor modifications of T and of the bounds derived in Section 8.4 (see for intance [67]).

Using *a priori bootsrtap*, we are able to validate fairly complicated periodic orbits (see Figure 8.2).

8.6.3 Validation of a connecting orbit

In this section, we present a computer-assisted proof of existence of a connecting orbit in the Lorenz system for the standard parameter values $(\sigma, \beta, \rho) = (10, \frac{8}{3}, 28)$. It is well know that

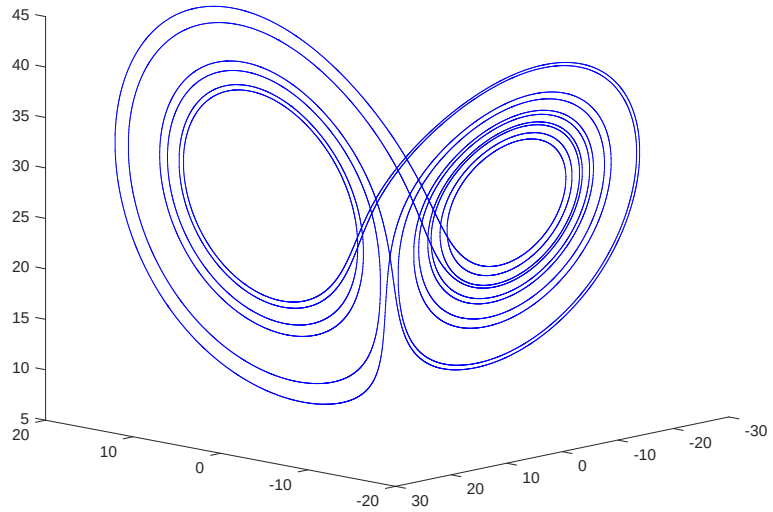


Figure 8.2 – A validated periodic orbit of the Lorenz system, whose period τ is approximately 11.9973. We used two iterations of *a priori bootstrap*, that is $p = 3$, for the validation. If we want to minimize the total number of coefficients to do the validation, we can take $k = 3$ and $m = 602$ (which makes 7225 coefficients in total), and we then get a validation radius of 1.5627×10^{-4}. It is possible to get a significantly lower validation radius, at the expense of a slight increase in the total number of coefficients: for instance with $k = 5$ and $m = 495$ (which makes 8911 coefficients in total), we get a validation radius of 4.7936×10^{-9}.

at these parameter values, the Lorenz system admits a transverse connecting orbit between $\left(\sqrt{\beta(\rho-1)}, \sqrt{\beta(\rho-1)}, \rho-1\right)$ and the origin.

While computer-assisted proofs of connecting orbits were already investigated several times using topological and analytical approaches (see [65, 155, 207, 178, 103, 104, 146, 185, 197, 125, 219, 206, 203, 181, 132]), a paper of particular relevance to the present work is [155], where a particular case of our method is developed, with only linear interpolation and no *a priori bootstrap* (that is $k = 1$ and $p = 1$). While the authors in [155] were able to validate several connecting orbits for the Lorenz system, they could not validate the aforementioned connecting orbit for the standard parameter values. In fact, one of the main motivations for the present work was to improve the setting of [155] to be able to validate more complicated orbits.

As we showcased in Section 8.6.1, using *a priori bootstrap* enables us to validate significantly more complicated orbits for the initial value problem, and this is also true for connecting orbits. Indeed we are able to validate the standard connecting orbit for the Lorenz system, with parameter values $(\sigma, \beta, \rho) = (10, \frac{8}{3}, 28)$. Before exposing the results, we briefly describe how to modify (8.2) to be able to handle connecting orbits.

Compared to an initial value problem on a given time interval, or to a periodic orbit, connecting orbits present an aditionnal difficulty which is that they are defined on an infinite time interval (from $-\infty$ to $+\infty$). To circumvent this difficulty and get back to a time interval of finite length, which is more suited to numerical computations (and to computer-assisted proofs), we are going to use local stable and unstable manifolds of the fixed points. By a computer-assisted method very similar to the one presented here, we first compute and validate local parameterization of the unstable manifold at $\left(\sqrt{\beta(\rho-1)}, \sqrt{\beta(\rho-1)}, \rho-1\right)$ and of the stable manifold at the origin. Since the main object of this work is not the rigorous computations of those manifolds, we simply assume that they are given (with validation radius) and do not give more details here. The interested reader can find more information about the computations

and validations of these parameterizations in [77, 66] and the references therein, and also more detailed examples of their usage to get connecting orbits in [65, 155, 60, 67].

We denote by P a local parameterization of the stable manifold of the origin, and by Q a local parameterization of the unstable manifold of a local parameterization of the stable manifold of $\left(\sqrt{\beta(\rho-1)}, \sqrt{\beta(\rho-1)}, \rho-1\right)$. We point out that both manifolds are two-dimensional. We then want to solve

$$\begin{cases} u'(t) = \phi(u(t)), \quad t \in [0, \tau], \\ u(0) = Q(\varphi), \\ u(\tau) = P(\theta), \end{cases} \tag{8.24}$$

where φ and θ each is a one dimensional parameter, the parameter in the other dimension being fixed to isolate the solution. Notice that τ is now an unknown of the system. As for the initial value problem, we can then consider an equivalent integral formulation (possibly with *a priori bootstrap*) and define an equivalent fixed point operator T very similar to the one introduced in Section 8.3. The additional equation $u(\tau) = P(\theta)$ and the fact that we have three additional variables τ, θ and φ only requires minor modifications of T and of the bounds derived in Section 8.4 (see for intance [67, 155]).

Using $p = 3$, $k = 3$ and $m = 1150$ (that is a total number of 13803 coefficients) we are then able to rigorously compute a solution of (8.24) (see Figure 8.3).

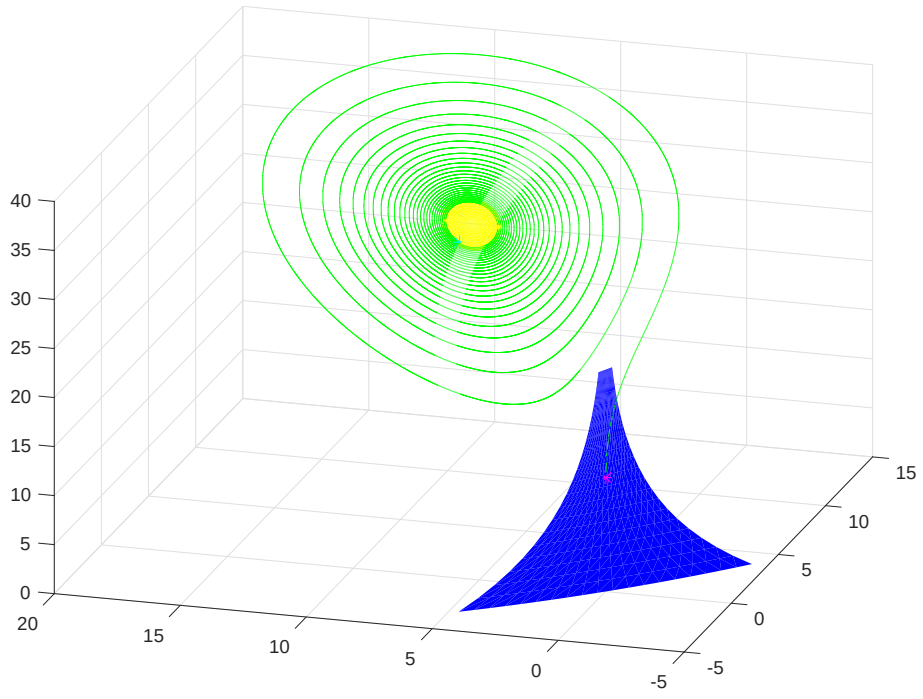


Figure 8.3 – Validated connecting orbit for the Lorenz system, with parameters $(\sigma, \beta, \rho) = (10, \frac{8}{3}, 28)$. The local stable manifold of the origin is in blue, the local unstable manifold of $\left(\sqrt{\beta(\rho-1)}, \sqrt{\beta(\rho-1)}, \rho-1\right)$ in yellow, and the green connection between them (of length $\tau \simeq 17.3$) is validated using polynomial interpolation, with *a priori bootstrap* ($p = 3$). The proof gives a validation radius of $r = 3.1340 \times 10^{-5}$.

8.7 Examples of applications for ABC flows

In this section, we apply our method to the non polynomial vector field

$$\phi_{A,B,C}(x,y,z) := \begin{pmatrix} A\sin(z) + C\cos(y) \\ B\sin(x) + A\cos(z) \\ C\sin(y) + B\cos(x) \end{pmatrix}, \quad A,B,C \in \mathbb{R}.$$

The map $\phi_{A,B,C}$ is usually referred to as the Arnold-Beltrami-Childress (ABC) vector field, and gives a prime example of complex steady incompressible periodic flows in 3D (see [55, 112, 126] and the references therein).

The main point of this section is to briefly illustrate the applicability of our technique to non polynomial vector fields. We plan on studying more thoroughly ABC flows with the help of our a posteriori validation method in a future work. Recently, the existence of orbits, that are periodic up to a shift of 2π in one coordinates, have been proven in the cases $A = B = C = 1$ and $0 < A \ll 1,\ B = C = 1$ [163, 213]. Applying the method developed in this paper, we were able to complete these results by proving the following statements.

Theorem 8.7.1. *For all $A = 0.1, 0.2, \ldots, 1$, with $B = C = 1$, there exists $\tau_A \in [\tau_A^-, \tau_A^+]$ (see Table 8.5) and a solution (x, y, z) of the ABC flow such that*

$$x(t+\tau) = x(t) + 2\pi, \quad y(t+\tau) = y(t), \quad z(t+\tau) = z(t), \quad \forall\ t \in \mathbb{R}.$$

Proof. The proof is done by running `script_proofs_A11.m` (available at [76]), which for each $A = 0.1, 0.2, \ldots, 1$ computes an approximate solution, then computes bounds satisfying (8.11)-(8.14) as described in Section 8.4, and finally finds positive r_∞ and r such that (8.15)-(8.16) holds. □

Theorem 8.7.2. *For all $A = B = C = 1$, there exists $\tau \in [7.797656, 7.797666]$ and a solution (x, y, z) of the ABC flow such that*

$$x(t+\tau) = x(t) + 4\pi, \quad y(t+\tau) = y(t), \quad z(t+\tau) = z(t), \quad \forall\ t \in \mathbb{R}$$

and $x(\cdot + \tau) \neq x(\cdot) + 2\pi$.

Proof. The proof is done by running `script_proofs_111.m` (available at [76]), which computes an approximate solution, then computes bounds satisfying (8.11)-(8.14) as described in Section 8.4, and finally finds positive r_∞ and r such that (8.15)-(8.16) holds. □

The solutions given by Theorem 8.7.1 are represented in Figure 8.4, and the solution given by Theorem 8.7.2 is represented in Figure 8.5.

A	1	0.9	0.8	0.7	0.6
τ_A^-	3.23527736	3.41779635	3.62512508	3.86405419	4.14464726
τ_A^+	3.23527746	3.41779647	3.62512521	3.86405436	4.14464749
A	0.5	0.4	0.3	0.2	0.1
τ_A^-	4.48269179	4.90491344	5.46177978	6.26680147	7.67945129
τ_A^+	4.48269213	4.90491401	5.46178092	6.26680442	7.67946552

Table 8.5 – The intervals $[\tau_A^-, \tau_A^+]$, for $A = 0.1, 0.2, \ldots, 1$ in which the *period* τ_A of the solution described in Theorem 8.7.1 is proved to be.

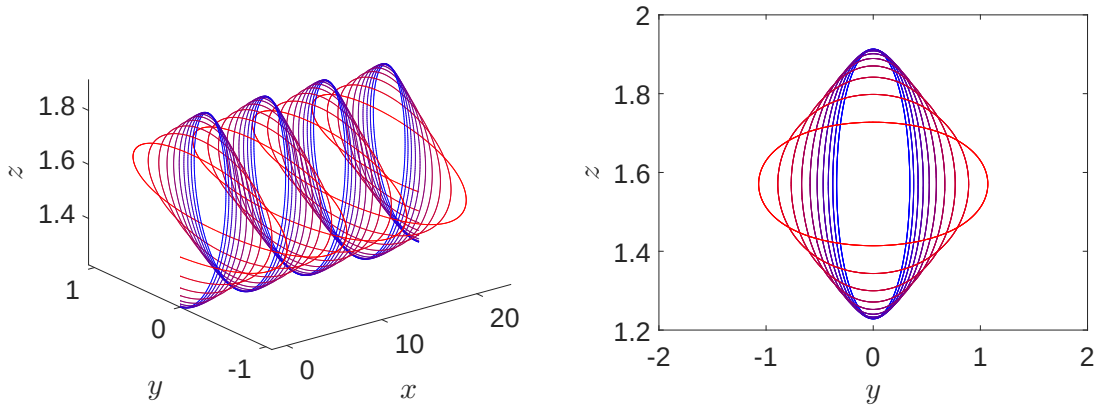


Figure 8.4 – These are the orbits that are described in Theorem 8.7.1. The color varies from blue for $A = 1$ to red for $A = 0.1$. Each proof was done with $p = 2$, $k = 2$ and $m = 50$, and gave a validation radius varying from $r = 4.8313 \times 10^{-8}$ to $r = 7.4012 \times 10^{-6}$.

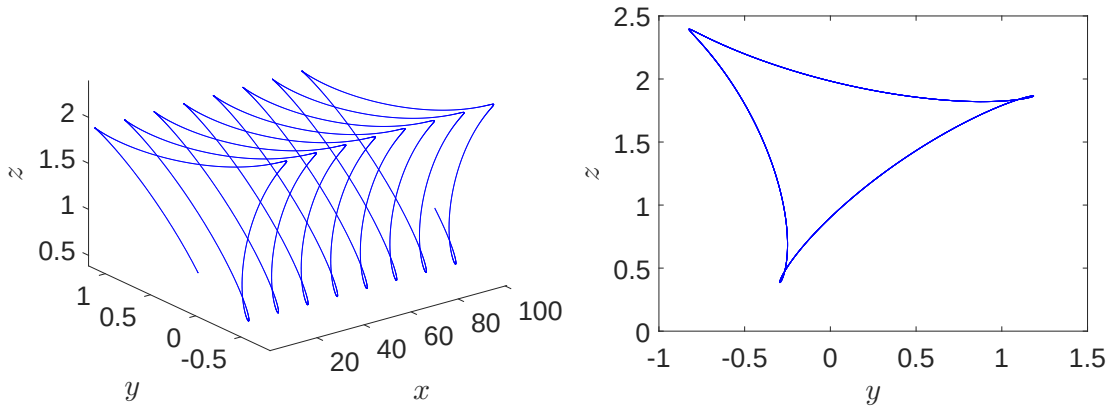


Figure 8.5 – This is the orbit that is described in Theorem 8.7.2. The proof was done with $p = 2$, $k = 2$ and $m = 300$, and gave a validation radius $r = 4.0458 \times 10^{-6}$.

Appendix

For the sake of completeness, we give here some properties of the Lebesgue constant Λ_k, as well as proofs of Proposition 8.3.1 and Proposition 8.3.2. We will assume here that u is defined and smooth on $[-1,1]$. The corresponding estimates on $[t_j, t_{j+1}]$ can the easily be deduced by rescaling.

We recall that Λ_k is defined as the norm of the interpolation operator mapping $\mathcal{C}^0([-1,1],\mathbb{R})$ to itself, and associating a continuous function u to its interpolation polynomial $P_k(u)$ of order k. Of course this operator (and its norm) depend on the interpolation points, which in this work are the Chebyshev interpolation points of the second kind

$$x_l^k = \cos\left(\frac{k-l}{k}\pi\right), \quad \text{for all } l = 0, \dots, k.$$

Introducing the basis consisting of the *Lagrange functions*

$$L_i^k(x) := \prod_{j \neq i} \frac{x - x_j^k}{x_i^k - x_j^k},$$

we have that the *Lagrange interpolation polynomial of order* k is given by

$$P_k(u)(x) = \sum_{i=0}^{k} u(x_i^k) L_i^k(x). \tag{8.25}$$

One can then show (see for instance [199]), that

$$\Lambda_k = \sup_{x \in [-1,1]} \sum_{i=0}^{k} |L_i^k(x)|, \tag{8.26}$$

and therefore we get

$$\sup_{x \in [-1,1]} |P_k(u)(x)| \leq \Lambda_k \max_{0 \leq i \leq k} |u(x_i^k)|,$$

which is exactly (8.18).

Since we used several times (8.18) and Proposition 8.3.2 in Section 8.4, the bounds that we obtained there depend on the Lebesgue constant Λ_k. Therefore we need computable (and as sharp as possible) upper bounds for Λ_k. One possibility is to use the well known bound (again see for instance [199])

$$\Lambda_k \leq 1 + \frac{2}{\pi}\ln(k+1). \tag{8.27}$$

However, we can do better, at least when k is odd. In that case, it has been shown (see [114]) that

$$\Lambda_k = \frac{1}{k} \sum_{l=0}^{k-1} \cot\left(\frac{2l+1}{4k}\pi\right),$$

and this formula can be evaluated rigorously using interval arithmetic. Unfortunately, there is no such formula for k even. For small values ($k = 2$ and $k = 4$) we computed Λ_k by hand using (8.26), and for $k \geq 6$ we used (8.27) (it is also know that $\Lambda_k \sim \frac{2}{\pi}\ln(k)$, therefore (8.27) is sharp for large k).

We now turn our attention to the interpolation estimates of Section 8.3. We point out that the analogue of Proposition 8.3.1 for the Chebyshev interpolation points of the first kind is very standard, and can be found in many textbooks. However, the case of the Chebyshev interpolation points of the second kind is seldom discussed, therefore we include a proof here for the sake of completeness (which is nothing but a slight adaptation of the *standard* proof).

Proof of Proposition 8.3.1. We consider the polynomial $W_k(x) = \prod_{l=0}^{k}(x - x_l^k)$ and use the standard interpolation error estimate for a function $u \in \mathcal{C}^{k+1}$ (see for instance [99]),

$$\|u - P_k(u)\|_\infty \leq \frac{\|W_k\|_\infty}{(k+1)!} \left\|u^{(k+1)}\right\|_\infty .$$

To prove Proposition 8.3.1, we only have to show that $\|W_k\|_\infty = \frac{1}{2^{k-1}}$ (the remaining factor $\frac{1}{2^{k+1}}$ coming from the rescaling). Introducing, for $k \in \mathbb{N}$, the k-th Chebyshev polynomial of the second kind U_k, defined by

$$U_k(\cos(\theta)) = \frac{\sin(k\theta)}{\sin(\theta)},$$

we have that

$$W_k(x) = (x-1)(x+1)\frac{U_k(x)}{2^{k-1}}. \tag{8.28}$$

Indeed, the right hand side of (8.28) is a unitary polynomial of degree $k+1$, that has the same zeros as W_k. We can then rewrite

$$\begin{aligned} W_k(x) &= \frac{1}{2^{k-1}}(x-1)(x+1)\frac{\sin(k\arccos(x))}{\sqrt{1-x^2}} \\ &= -\frac{1}{2^{k-1}}\sqrt{1-x^2}\sin(k\arccos(x)), \end{aligned}$$

and we end up with

$$\|W_k\|_\infty = \frac{1}{2^{k-1}},$$

so Proposition 8.3.1 is proven. □

Proof of Proposition 8.3.2. The first part of the bound, namely

$$(1+\Lambda_k)\left(\frac{\pi}{4}\right)^l \frac{(k+1-l)!}{(k+1)!},$$

comes from a combination of the Lebesgue constant and Jackson's Theorem, and can be found in [99]. However, it does not give a very sharp interpolation error estimate for small values of k and l, therefore we derive here the second part of the bound, namely

$$\frac{1}{l!2^l}\sum_{q=0}^{\left[\frac{l-1}{2}\right]}\frac{1}{4^q}\binom{l-1}{2q}\binom{2q}{q}$$

that can be used in those cases.

Letting $u(x) = x^p$ (with $p \in \{0,\dots,k\}$) in (8.25) leads to

$$x^p = \sum_{i=0}^{k}(x_i^k)^p L_i^k(x), \quad \text{for all } x \in \mathbb{R}. \tag{8.29}$$

We now fix a function $u \in \mathcal{C}^l$. Using (8.29) with $p = 0$ (that is $1 = \sum_{i=0}^{k} L_i^k(x)$), we get

$$P_k(u)(x) - u(x) = \sum_{i=0}^{k} u(x_i^k)L_i^k(x) - u(x)\left(\sum_{i=0}^{k} L_i^k(x)\right) = \sum_{i=0}^{k}\Big(u(x_i^k) - u(x)\Big)L_i^k(x).$$

Using Taylor's formula, we then get

$$\begin{aligned} P_k(u)(x) - u(x) &= \sum_{i=0}^{k}\left(\sum_{p=1}^{l-1}\frac{(x_i^k - x)^p}{p!}u^{(p)}(x) + \frac{(x_i^k-x)^l}{l!}u^{(l)}(y_i)\right)L_i^k(x) \\ &= \sum_{p=1}^{l-1}\frac{u^{(p)}(x)}{p!}\sum_{i=0}^{k}(x_i^k - x)^p L_i^k(x) + \sum_{i=0}^{k}\frac{(x_i^k-x)^l}{l!}u^{(l)}(y_i)L_i^k(x), \end{aligned}$$

for some y_i in $[-1,1]$. Then, expanding the $(x_i^k - x)^p$ terms and using again (8.29), we get that

$$\begin{aligned}\sum_{i=0}^{k}(x_i^k - x)^p L_i^k(x) &= \sum_{q=0}^{p}\binom{p}{q}(-x)^{p-q}\sum_{i=0}^{k}(x_i^k)^q L_i^k(x)\\ &= \sum_{q=0}^{p}\binom{p}{q}(-x)^{p-q}x^q\\ &= (x-x)^p\\ &= 0,\end{aligned}$$

and thus

$$P_k(u)(x) - u(x) = \sum_{i=0}^{k}\frac{(x_i^k - x)^l}{l!}u^{(l)}(y_i)L_i^k(x),$$

Letting

$$\lambda_i^k := \prod_{j\neq i}\frac{1}{x_i^k - x_j^k},$$

we can easily observe that $L_i^k(x) = \lambda_i^k W_k(x)/(x - x_i^k)$, and therefore

$$\begin{aligned}|P_k(u)(x) - u(x)| &\leq \frac{\left\|u^{(l)}\right\|_\infty}{l!}\sum_{i=0}^{k}|x_i^k - x|^l |L_i^k(x)|\\ &= \frac{\left\|u^{(l)}\right\|_\infty}{l!}|W_k(x)|\sum_{i=0}^{k}|\lambda_i^k||x_i^k - x|^{l-1}.\end{aligned}$$

According to [199], in case the points x_i^k are the Chebyshev interpolation points of the second kind, we have

$$\lambda_i^k = \begin{cases}(-1)^i\frac{2^{k-1}}{k}, & i = 1,\ldots,k-1,\\ (-1)^i\frac{2^{k-1}}{2k}, & i = 0,k.\end{cases}$$

Remembering that $\|W_k\|_\infty = \frac{1}{2^{k-1}}$, we get

$$|P_k(u)(x) - u(x)| \leq \frac{\left\|u^{(l)}\right\|_\infty}{l!}\frac{1}{k}\left(\frac{(1+x)^{l-1}}{2} + \frac{(1-x)^{l-1}}{2} + \sum_{i=1}^{k-1}|x_i^k - x|^{l-1}\right).$$

The function

$$x \mapsto \frac{(1+x)^{l-1}}{2} + \frac{(1-x)^{l-1}}{2} + \sum_{i=1}^{k-1}|x_i^k - x|^{l-1}$$

is even and increasing on $[0,1]$, therefore its maximum is reached at $x = 1$ and we get

$$|P_k(u)(x) - u(x)| \leq \frac{\left\|u^{(l)}\right\|_\infty}{l!}\frac{1}{k}\left(2^{l-2} + \sum_{i=1}^{k-1}\left(1 - \cos\frac{i\pi}{k}\right)^{l-1}\right).$$

Then, we compute

$$\begin{aligned}
2^{l-2}+\sum_{i=1}^{k-1}\left(1-\cos\frac{i\pi}{k}\right)^{l-1} &= \sum_{i=0}^{k}\left(1-\cos\frac{i\pi}{k}\right)^{l-1}-2^{l-2}\\
&= \sum_{i=0}^{k}\sum_{q=0}^{l-1}\binom{l-1}{q}(-1)^q\cos^q\frac{i\pi}{k}-2^{l-2}\\
&= \sum_{q=0}^{\left[\frac{l-1}{2}\right]}\binom{l-1}{2q}\sum_{i=0}^{k}\cos^{2q}\frac{i\pi}{k}-2^{l-2}\\
&= \sum_{q=0}^{\left[\frac{l-1}{2}\right]}\binom{l-1}{2q}\sum_{i=0}^{k}\frac{1}{4^q}\left(\binom{2q}{q}+\sum_{j=0}^{q-1}\binom{2q}{j}\cos 2(q-j)\frac{i\pi}{k}\right)-2^{l-2}\\
&= \sum_{q=0}^{\left[\frac{l-1}{2}\right]}\binom{l-1}{2q}\frac{1}{4^q}\left((k+1)\binom{2q}{q}+2\sum_{j=0}^{q-1}\binom{2q}{j}\right)-2^{l-2}\\
&= \sum_{q=0}^{\left[\frac{l-1}{2}\right]}\binom{l-1}{2q}\frac{1}{4^q}\left(k\binom{2q}{q}+4^q\right)-2^{l-2}\\
&= k\sum_{q=0}^{\left[\frac{l-1}{2}\right]}\binom{l-1}{2q}\binom{2q}{q}.
\end{aligned}$$

We end up with

$$|P_k(u)(x)-u(x)|\leq\frac{\left\|u^{(l)}\right\|_\infty}{l!}\sum_{q=0}^{\left[\frac{l-1}{2}\right]}\binom{l-1}{2q}\binom{2q}{q},$$

and Proposition 8.3.2 is proven (the lacking $\frac{1}{2^l}$ factor coming from the time rescaling). □

Chapter 9

Existence and continuation of traveling waves for the suspended bridge equation

Abstract

This chapter is taken from [58]. We prove existence of symmetric homoclinic orbits for the suspension bridge equation $u'''' + \beta u'' + e^u - 1 = 0$ for all parameter values $\beta \in [0.5, 1.9]$. For each β, a parameterization of the stable manifold is computed and the symmetric homoclinic orbits are obtained by solving a projected boundary value problem using Chebyshev series. The proof is computer-assisted and combines the uniform contraction theorem and the radii polynomial approach, which provides an efficient means of determining a set, centered at a numerical approximation of a solution, on which a Newton-like operator is a contraction.

9.1 Introduction

One of the simplest models [162, 124] for a suspension bridge is the partial differential equation (PDE)

$$\frac{\partial^2 U}{\partial T^2} = -\frac{\partial^4 U}{\partial X^4} - e^U + 1. \tag{9.1}$$

Here $U(T, X)$ describes the deflection of the roadway from the rest state $U = 0$ as a function of time T and the spatial variable X (in the direction of traffic). This paper is concerned with traveling wave solutions of (9.1), i.e., solutions $U(T, X) = u(X - cT)$ describing a disturbance with profile u propagating at velocity c along the surface of the bridge. In particular, we apply a computer-assisted proof method to show that there is a large range of velocities for which such a solitary wave exists.

Looking for traveling waves of (9.1) with wave speed c leads to the ordinary differential equation

$$u'''' + c^2 u'' + e^u - 1 = 0. \tag{9.2}$$

For large positive and negative values of the independent variable $t = X - cT$ we assume the solution to converge to the equilibrium $u = 0$. Due to the reversibility symmetry of the PDE in both time and space, we may restrict our attention to symmetric solutions. Hence, setting

$\beta = c^2$, we are looking for symmetric homoclinic orbits satisfying

$$\begin{cases} u'''' + \beta u'' + e^u - 1 = 0 \\ u(-t) = u(t) \\ \lim_{t\to\infty} u(t) = 0. \end{cases} \tag{9.3}$$

Fourth order differential equations of the form $u'''' + \beta u'' + f(u) = 0$ for various nonlinearities f have been studied extensively. For the bistable nonlinearity $f(u) = u^3 - u$ the equation is a standard model in pattern formation, called the Swift-Hohenberg equation (see [180] and references therein), whereas the quadratic nonlinearity $f(u) = u^2 - u$ appears, for example, in the study of water waves [81]. For the piecewise linear case $f(u) = \max\{u, 0\}$ homoclinic solutions were obtained in [162, 98]. For the problem with the exponential nonlinearity $f(u) = e^u - 1$ a family of periodic solutions was established in [179].

In [98] the question about existence of a symmetric homoclinic orbit of (9.3) is raised. This question was addressed by variational methods in [68], where the authors proved the result for *almost all* parameter values $\beta \in (0, 2)$. In [193] the existence of homoclinic orbits was demonstrated for all $\beta \in (0, c_*^2) \approx (0, 0.5516)$, again using variational methods as well as intricate estimates on the second variation. In a different direction, using a computer-assisted proof, it was proven in [80] that (9.3) has at least 36 homoclinic solutions for the single parameter value $\beta = 1.69$.

In the present paper we complement the above results by proving the following.

Theorem 9.1.1. *For all parameter values $\beta \in [0.5, 1.9]$ there exists a symmetric homoclinic orbit of* (9.3)*.*

We remark that for $|\beta| < 2$ the origin is a saddle-focus, while for $\beta > 2$ it is a saddle-center. Furthermore, we note the integral identity $\int_{\mathbb{R}} |u''|^2 - \beta |u'|^2 = -\int_{\mathbb{R}} (e^u - 1)u$. Since the right hand side is non-positive, homoclinic orbits are excluded for $\beta \leq 0$. It is thus expected that the parameter range for which homoclinics exist is $\beta \in (0, 2)$, or, equivalently, wave speeds $c \in (0, \sqrt{2})$. Our method for proving the result in Theorem 9.1.1 is computer-assisted. While it can certainly be extended somewhat beyond the interval $[0.5, 1.9]$, it is not possible to cover the entire range $(0, 2)$ in this way. Indeed, as β decreases towards 0 the amplitude of the solution diverges (u becomes very negative), whereas when β tends to 2 the homoclinic orbit collapses onto the trivial solution. In both limit regimes computer-assisted proofs become harder and harder. Since the result in [193] already covers the range $\beta \in (0, 0.55]$, we thus focus on the parameter range $[0.5, 1.9]$. We note that at $\beta = 2$ a Hamiltonian-Hopf bifurcation occurs. In future work we intend to unfold this bifurcation and subsequently connect the homoclinic orbit that bifurcates to the branch covered by Theorem 9.1.1 (at that point we will know how far we have to push the current continuation technique beyond $\beta = 1.9$ to connect all the way to the bifurcation point).

The rest of the paper is dedicated to the proof of Theorem 9.1.1. Our approach begins by rewriting (9.3) as a first order system for $(u_1, u_2, u_3, u_4) = (u, u', u'', u''')$ and then making the change of variables $(v_1, v_2, v_3, v_4) = (e^{u_1} - 1, u_2, u_3, u_4)$ to obtain

$$\begin{cases} v_1' = v_2 + v_1 v_2 \\ v_2' = v_3 \\ v_3' = v_4 \\ v_4' = -\beta v_3 - v_1. \end{cases} \tag{9.4}$$

There are two reasons for performing this change of variables. First, it turns the system into a polynomial vector field, which has technical advantages when performing the analysis to derive the necessary bounds. Second, while u_1 may become very negative for small values of β, the

variable v_1 is always bounded from below by -1. Our goal is now to prove the existence of symmetric homoclinic solutions to (9.4) for all $\beta \in [0.5, 1.9]$.

We split the problem into two parts. On the one hand a rigorous computational description of the local (un)stable manifold is required. On the other hand we need to solve, via a rigorous computational technique, a boundary value problem for the part of the orbit between the local invariant manifolds. We attack both parts by a continuation technique in the context of the radii polynomial approach. This parametrized Newton-Kantorovich method, adapted to a computational setting, is introduced in Section 9.2. In Section 9.3 we combine this with the parameterization method to obtain descriptions of the local (un)stable manifold of the equilibrium $0 \in \mathbb{R}^4$. Essentially the same technique is then applied in Section 9.4 in a Chebyshev series setting to solve the boundary value problem. These two aspects are then combined into a rigorous computational continuation of the homoclinic solution to (9.3). We note that for smaller values of β the boundary value problem is the more difficult part of the problem, as the orbit makes a bigger and bigger excursion away from the origin. On the other hand, for values of β close to 2 it is more difficult to obtain the local (un)stable manifold of the origin, as the real part of the eigenvalues tends to 0. The algorithmic issues encountered when implementing the proof of Theorem 9.1.1 are discussed in Section 9.5.

Finally, let us mention that there is a growing literature on the subject of computer-assisted methods for proving existence of connecting orbits, see [198, 60, 65, 203, 204, 207, 211]. The main novel contribution of the current paper is to do rigorous continuation of a homoclinic orbit over a large range of parameter values. The method is generally applicable for connecting orbits problems in parameter dependent problems. In that sense Theorem 9.1.1, while providing a new result for traveling waves in the suspension bridge problem which complements earlier work, is an illustration.

9.2 The radii polynomial approach

In this section we present the functional analytic setup of our continuation method, which is formulated in terms of the *radii polynomials*, see Definition 9.2.4. It will be used both to find the stable manifold and to solve the boundary value problem. For more details and proofs we refer to [79, 109, 64].

Consider a sequence of Banach spaces $(X_1, \|\cdot\|_{X_1}), \dots, (X_d, \|\cdot\|_{X_d})$ and the (product) Banach space

$$X = X_1 \times X_2 \times \cdots \times X_d,$$

with the induced norm defined by

$$\|x\|_X = \max\left(\|x^{(1)}\|_{X_1}, \dots, \|x^{(d)}\|_{X_d}\right),$$

where $x = (x^{(1)}, \dots, x^{(d)}) \in X$ with $x^{(j)} \in X_j$ for $j = 1, \dots, d$. Denote by

$$B_r(y) = \{x \in X \mid \|x - y\|_X \leq r\}$$

the closed ball of radius $r > 0$ centered at $y \in X$.

Consider an interval of parameters $[\beta_0, \beta_1] \subset \mathbb{R}$ and $T : [\beta_0, \beta_1] \times X \to X$ a Fréchet differentiable operator. For each $j = 1, \dots, d$, denote by $T^{(j)} : [\beta_0, \beta_1] \times X \to X_j$ the projection of T onto X_j. Let $\bar{x}_{\beta_0}, \bar{x}_{\beta_1} \in X$ be approximate fixed points of $T(\beta_0, \cdot)$ and $T(\beta_1, \cdot)$, respectively, and define the linear interpolation

$$\bar{x}_\beta := \frac{\beta_1 - \beta}{\beta_1 - \beta_0}\bar{x}_{\beta_0} + \frac{\beta - \beta_0}{\beta_1 - \beta_0}\bar{x}_{\beta_1}. \tag{9.5}$$

Define the line of *centers* by $\{\bar{x}_\beta \mid \beta \in [\beta_0, \beta_1]\} \subset X$. For each $j = 1, \ldots, d$, define the bounds

$$\sup_{\beta\in[\beta_0,\beta_1]} \left\| T^{(j)}(\beta, \bar{x}_\beta) - \bar{x}_\beta^{(j)} \right\|_{X_j} \leq Y^{(j)}, \tag{9.6}$$

$$\sup_{\substack{b,c\in B_r(0)\\ \beta\in[\beta_0,\beta_1]}} \left\| D_x T^{(j)}(\beta, \bar{x}_\beta + b)c \right\|_{X_j} \leq Z^{(j)}(r), \tag{9.7}$$

for some $Y^{(j)} > 0$ and $Z^{(j)} : \mathbb{R}^+ \mapsto \mathbb{R}^+ : r \to Z^{(j)}(r)$. The goal of the radii polynomial approach is to provide an efficient way to prove that an operator is a uniform contraction over a subset of X. This subset consists of small balls around the line of centers, provided by the linear interpolation between two numerical approximations of solutions at different parameter values.

Definition 9.2.1. *Let X be a Banach space and $B \subset X$. Let $[\beta_0, \beta_1] \subset \mathbb{R}$ be a set of parameters. A function $\tilde{T} : [\beta_0, \beta_1] \times B \to B$ is a* uniform contraction *if there exists a constant κ such that $0 < \kappa < 1$ and such that $\|\tilde{T}(\beta, x) - \tilde{T}(\beta, y)\|_X \leq \kappa \|x - y\|_X$ for all $x, y \in B$ and all $\beta \in [\beta_0, \beta_1]$.*

The following result is a restatement of the uniform contraction principle (e.g. see [101] for a proof).

Theorem 9.2.2 (Uniform Contraction Principle). *Suppose there exists some $r > 0$ such that*

$$\tilde{T} : \begin{cases} [\beta_0, \beta_1] \times B_r(0) \longrightarrow B_r(0) \\ \qquad (\beta, x) \longmapsto \tilde{T}(\beta, x) := T(\beta, x + \bar{x}_\beta) - \bar{x}_\beta \end{cases} \tag{9.8}$$

is a uniform contraction, then for every $\beta \in [\beta_0, \beta_1]$, there exists a unique $\tilde{x}(\beta) \in B_r(\bar{x}_\beta)$ such that $T(\beta, \tilde{x}(\beta)) = \tilde{x}(\beta)$. Moreover, the function $\beta \mapsto \tilde{x}(\beta)$ is of class C^k if $(\beta, x) \mapsto T(\beta, x)$ is of class C^k.

With the bounds Y and Z on the residue and the derivative of T, see Equations (9.6) and (9.7), contractivity can be checked explicitly. This is expressed by the next theorem (we refer to [79, 109, 64] for a proof).

Theorem 9.2.3. *Given a set of parameters $[\beta_0, \beta_1] \subset \mathbb{R}$, consider the set of centers $\{\bar{x}_\beta \mid \beta \in [\beta_0, \beta_1]\}$ with $\bar{x}_\beta$ given by* (9.5)*. Assume that $T : [\beta_0, \beta_1] \times X \to X$ is an operator satisfying the bounds* (9.6) *and* (9.7)*. If there exists $r > 0$ such that $Y^{(j)} + Z^{(j)}(r) < r$, for each $j = 1, \ldots, d$, then $\tilde{T}$, defined by* (9.8)*, is a uniform contraction (on $B_r(0)$).*

Assuming we have determined explicit bounds $Y^{(j)}$ and $Z^{(j)}(r)$, where in practice the latter is a polynomial with positive coefficients, satisfying (9.6) and (9.7). It is convenient to introduce the radii polynomials, which provide an efficient way in verifying the hypotheses of Theorem 9.2.3.

Definition 9.2.4. *Let $Y = (Y^{(1)}, \ldots, Y^{(d)})$ and $Z = (Z^{(1)}, \ldots, Z^{(d)})$ be the bounds on the operator T_β as given by* (9.6) *and* (9.7) *respectively. We define the radii polynomials as*

$$p_j(r) := Y^{(j)} + Z^{(j)}(r) - r, \quad j = 1, \ldots, d. \tag{9.9}$$

One can see that the radii polynomials depend on the upper bounds Y and Z, and therefore they are not uniquely defined. But the smaller these bounds are, the higher the chances are to prove that the operator T_β is a contraction over a ball around the approximation. The following result shows how the radii polynomials are used in practice to give us the value of r for which we can apply Theorem 9.2.3.

Proposition 9.2.5. *Let*

$$\mathcal{I} := \bigcap_{j=1}^{d} \{r > 0 \mid p_j(r) < 0\},$$

and assume that $\mathcal{I} \neq \emptyset$. Then $\mathcal{I}$ is an open interval of $\mathbb{R}^+$, i.e., $\mathcal{I} = (r_{\min}, r_{\max})$. For any $r_0 \in (r_{\min}, r_{\max})$, $\tilde{T} : B_{r_0}(0) \times [\beta_0, \beta_1] \longrightarrow B_{r_0}(0)$ is a uniform contraction.

9.3 Parameterization of the stable manifold

In this section we compute an approximate parameterization of the (local) stable manifold at 0, and provide explicit error bounds on this parameterization. This is done by combining the ideas of the *parameterization method* (first introduced in [82, 83, 84], see also [130]) and of *rigorous computation* (following the approach of [66, 77]). Having computed the parameterization, we will be able to obtain the homoclinic connection in the next section by taking advantage of the fact that it is now enough to compute an orbit on a finite time interval, i.e., an orbit that ends up in the local stable and unstable manifolds (or rather, we compute and verify an orbit that starts from the symmetric section and ends up, after some finite time, in the local stable manifold, see (9.3)).

9.3.1 Looking for the stable manifold as a zero finding problem $F(\beta, a) = 0$

The first step is to recast the problem of finding a parameterization as looking for a zero of a map F, which is the aim of this section. Setting

$$\Psi_\beta(v) := \begin{pmatrix} v_2 + v_1 v_2 \\ v_3 \\ v_4 \\ -\beta v_3 - v_1 \end{pmatrix},$$

Equation (9.4) is rewritten as $v' = \Psi_\beta(v)$. The Jacobian at the origin is

$$D\Psi_\beta(0) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & -\beta & 0 \end{pmatrix},$$

and one finds that for $\beta \in [0, 2)$ it has two complex conjugated eigenvalues with negative real part, which we denote by $\lambda(\beta)$ and $\lambda^*(\beta)$:

$$\lambda(\beta) = -\frac{1}{2}\sqrt{2-\beta} + \mathrm{i}\frac{1}{2}\sqrt{2+\beta}. \tag{9.10}$$

The associated eigenvectors are given by $V(\beta)$ and $V^*(\beta)$, where

$$V(\beta) = \begin{pmatrix} 1 \\ \lambda(\beta) \\ \lambda(\beta)^2 \\ \lambda(\beta)^3 \end{pmatrix}. \tag{9.11}$$

The stable manifold at 0 is thus two dimensional. Since Ψ_β is analytic we may look for an analytic local parameterization of this manifold. We will look for this parameterization as a power series

$$Q_\beta(\theta) = \sum_{|\alpha| \geq 0} a_\alpha(\beta)\theta^\alpha, \quad \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \in \mathbb{C}^2, \; a_\alpha(\beta) = \begin{pmatrix} a_\alpha^{(1)}(\beta) \\ a_\alpha^{(2)}(\beta) \\ a_\alpha^{(3)}(\beta) \\ a_\alpha^{(4)}(\beta) \end{pmatrix} \in \mathbb{C}^4, \tag{9.12}$$

with standard multi-index notation: $\alpha \in \mathbb{N}^2$, $|\alpha| = \alpha_1 + \alpha_2$, $\theta^\alpha = \theta_1^{\alpha_1}\theta_2^{\alpha_2}$, and satisfying

$$Q_\beta(0) = 0, \quad DQ_\beta(0) = \begin{pmatrix} V(\beta) & V^*(\beta) \end{pmatrix}, \tag{9.13}$$

together with the invariance equation

$$DQ_\beta(\theta)\begin{pmatrix}\lambda(\beta) & 0\\ 0 & \lambda^*(\beta)\end{pmatrix}\theta = \Psi_\beta(Q_\beta(\theta)). \tag{9.14}$$

Remark 9.3.1. *Even though the vector field Ψ_β is real, the fact that we have two complex eigenvalues makes it easier to first look for a parameterization Q_β of the complex manifold and then recover the real parameterization (the one which will be of interest in the next section for computing the homoclinic orbit) by considering*

$$P_\beta(\theta) := Q_\beta(\theta_1 + \mathrm{i}\theta_2, \theta_1 - \mathrm{i}\theta_2), \quad \textit{for } \theta \in \mathbb{R}^2,$$

see [172, 66] for more details. This is due to the underlying symmetry $a_{(\alpha_2,\alpha_1)} = a^*_{(\alpha_1,\alpha_2)}$, which is respected by the function F introduced below.

Plugging the power series (9.12) into the invariance equation (9.14) we get

$$\sum_{|\alpha|\geq 0}(\alpha_1\lambda(\beta)+\alpha_2\lambda^*(\beta))a_\alpha(\beta)\theta^\alpha = \sum_{|\alpha|\geq 0}\begin{pmatrix} a^{(2)}_\alpha(\beta) + (a^{(1)}(\beta)\star a^{(2)}(\beta))_\alpha\\ a^{(3)}_\alpha(\beta)\\ a^{(4)}_\alpha(\beta)\\ -a^{(1)}_\alpha(\beta) - \beta a^{(3)}_\alpha(\beta)\end{pmatrix}\theta^\alpha, \tag{9.15}$$

where $\star$ stands for the Cauchy product. We recall that, given two sequences u and v of complex numbers (indexed over $\mathbb{N}^2$), their Cauchy product is the sequence defined by

$$(u\star v)_\alpha = \sum_{0\leq\sigma\leq\alpha} u_\sigma v_{\alpha-\sigma}, \qquad \text{for all } \alpha\in\mathbb{N}^2,$$

where $\sigma\leq\alpha$ means $\sigma_1\leq\alpha_1$ and $\sigma_2\leq\alpha_2$ (and similarly $\alpha-\sigma = (\alpha_1-\sigma_1, \alpha_2-\sigma_2)$).

Notice that the additional conditions (9.13) imply that the coefficients of total degree 0 and 1 are equal on both sides of Equation (9.15).

Finding an analytic parameterization of the local manifold is now equivalent to find a zero of $F(\beta,\cdot)$, defined component-wise by

$$F(\beta,a) = \begin{cases} a_{0,0} & \text{if } \alpha=(0,0),\\ a_{1,0} - V(\beta) & \text{if } \alpha=(1,0),\\ a_{0,1} - V^*(\beta) & \text{if } \alpha=(0,1),\\ (\alpha_1\lambda(\beta)+\alpha_2\lambda^*(\beta))a_\alpha - \begin{pmatrix} a^{(2)}_\alpha + (a^{(1)}\star a^{(2)})_\alpha\\ a^{(3)}_\alpha\\ a^{(4)}_\alpha\\ -a^{(1)}_\alpha - \beta a^{(3)}_\alpha\end{pmatrix} & \text{for } |\alpha|\geq 2.\end{cases}$$

9.3.2 Getting to the fixed point formulation

Let $\nu\geq 1$ and denote by ℓ^1_ν the Banach space of complex valued sequences $u=(u_\alpha)_{|\alpha|\geq 0}$ such that

$$\|u\|_{1,\nu} := \sum_{|\alpha|=0}^{\infty}|u_\alpha|\nu^{|\alpha|} < \infty.$$

This space is a Banach algebra under the Cauchy product, which gives us control on the quadratic terms.

Lemma 9.3.2. *For $u,v\in\ell^1_\nu$, $\|u\star v\|_{1,\nu}\leq\|u\|_{1,\nu}\|v\|_{1,\nu}$.*

Definition 9.3.3. *In this section we consider*

$$X := (\ell^1_\nu)^4, \quad \textit{with the norm } \|a\|_X := \max_{j=1,\dots,4} \|a^{(j)}\|_{1,\nu}.$$

We are going to look for zeros a of $F(\beta,\cdot)$ in the space X. Notice that $a \in X$ means that

$$\sum_{|\alpha|\geq 0} |a^{(j)}_\alpha| \nu^{|\alpha|} < \infty, \quad \text{for } j = 1,\dots,4,$$

which ensures that the associated parameterization Q_β is well defined at least for

$$|\theta|_\infty := \max(|\theta_1|, |\theta_2|) \leq \nu.$$

We now explain how to rigorously determine a parameterization of the manifold for all values of β in a given interval $[\beta_0, \beta_1]$. It will be more convenient to work with a rescaled parameter s ranging between 0 and 1. Therefore we define

$$\beta_s = \beta_0 + s(\beta_1 - \beta_0) = \beta_0 + s\Delta\beta, \quad \text{for } s \in [0,1].$$

We want to get a parameterization $a(s)$ such that

$$F(\beta_s, a(s)) = 0, \quad \text{for } s \in [0,1].$$

Since we are working on the interval $[\beta_0, \beta_1]$, parameterized by $s \in [0,1]$, we have altered the notation of the parametrization of the coefficients a slightly compared to Section 9.3.1, namely $a(s)$ instead of $a(\beta)$. We will use $a(s)$ throughout the remainder of the paper, except in Section 9.4.1, where the notation $a(\beta)$ is more appropriate.

We first compute approximate zeros $\bar{a}(0)$ and $\bar{a}(1)$ of $F(\beta_0,\cdot)$ and $F(\beta_1,\cdot)$ respectively, by solving numerically the truncated problem (for $s = 0$ and $s = 1$)

$$F^{[N]}(\beta_s,\cdot) := (F_\alpha(\beta_s,\cdot))_{0\leq|\alpha|<N} = 0,$$

for some $N \geq 1$, and by padding the obtained solutions with 0 to get elements of $X = (\ell^1_\nu)^4$. We then define for $s \in [0,1]$

$$\bar{a}(s) := \bar{a}(0) + s(\bar{a}(1) - \bar{a}(0)) = \bar{a}(0) + s\Delta\bar{a}.$$

If $\bar{a}(0)$ and $\bar{a}(1)$ are two good approximate zeros (of $F(\beta_0,\cdot)$ and $F(\beta_1,\cdot)$ respectively) and if $|\beta_1 - \beta_0|$ is not too large, $\bar{a}(s)$ should be a good approximate zero of $F(\beta_s,\cdot)$ for each $s \in [0,1]$. We are going to reformulate this claim into a mathematical statement and prove that in a given neighbourhood of $\bar{a}(s)$ there exists a unique zero of $F(\beta_s,\cdot)$ for all $s \in [0,1]$. To put this in the framework described in Section 9.2, we consider the operator

$$T(\beta, a) = a - AF(\beta, a),$$

where A, defined below, is an approximate inverse of $D_aF(\beta_0, \bar{a}(0))$. Namely, for N large enough,

$$A^\dagger := \begin{pmatrix} D_aF^{[N]}(\beta_0, \bar{a}(0)) & & 0 & \\ & \tilde{M}_N & & \\ 0 & & \tilde{M}_{N+1} & \\ & & & \ddots \end{pmatrix}$$

should be a reasonably good approximation of $D_aF(\beta_0, \bar{a}(0))$, where, for any $k \geq N$, $\tilde{M}_k$ is the $4(k+1)$ by $4(k+1)$ block diagonal matrix

$$\tilde{M}_k := \begin{pmatrix} k\lambda(\beta_0)I_4 & & & 0 \\ & ((k-1)\lambda(\beta_0) + \lambda^*(\beta_0))I_4 & & \\ 0 & & \ddots & \\ & & & k\lambda^*(\beta_0)I_4 \end{pmatrix},$$

with I_4 the 4 by 4 identity matrix. Finally, we define A as

$$A := \begin{pmatrix} J & & 0 & \\ & M_N & & \\ 0 & & M_{N+1} & \\ & & & \ddots \end{pmatrix}, \tag{9.16}$$

where J is a numerical approximation of $\left(D_a F^{[N]}(\beta_0, \bar{a}(0))\right)^{-1}$, while the $M_k = \tilde{M}_k^{-1}$ are exact inverses. The operators $A^\dagger$ and A are then approximate inverses of each other: approximate in the finite part and exact in the infinite tail.

Remark 9.3.4. *To make sense of this matrix representation of $A^\dagger$ and A, as well as $\tilde{M}_k$ and M_k, one should think of a_α as an infinite vector where the elements are ordered according to increasing degree $|\alpha| = \alpha_1 + \alpha_2$ and within fixed degree by increasing α_2, while also taking into account that each a_α is a vector in $\mathbb{C}^4$. This means that a is represented as*

$$a = \begin{pmatrix} a_{0,0} \\ a_{1,0} \\ a_{0,1} \\ a_{2,0} \\ a_{1,1} \\ a_{0,2} \\ \vdots \end{pmatrix}, \text{ where } a_\alpha = \begin{pmatrix} a_\alpha^{(1)} \\ a_\alpha^{(2)} \\ a_\alpha^{(3)} \\ a_\alpha^{(4)} \end{pmatrix} \text{ for each } \alpha \in \mathbb{N}^2, \text{ and that } a^{(j)} = \begin{pmatrix} a_{0,0}^{(j)} \\ a_{1,0}^{(j)} \\ a_{0,1}^{(j)} \\ a_{2,0}^{(j)} \\ a_{1,1}^{(j)} \\ a_{0,2}^{(j)} \\ \vdots \end{pmatrix} \text{ for } j = 1, \dots, 4.$$

The above representation describes the operators as infinite matrices where each element $A_{\alpha',\alpha}$ is a linear operator on $\mathbb{C}^4$, i.e. a 4×4 matrix that we will occasionally denote by $A_{\alpha',\alpha} = \{A_{\alpha',\alpha}^{(i,j)}\}_{1 \le i,j \le 4}$.

We now follow the ideas described in Section 9.2, using the Banach space $X = \left(\ell^1_\nu\right)^4$ endowed with the norm $\|a\|_X = \max_{j=1,\dots,4} \|a^{(j)}\|_{1,\nu}$. In the next subsections we are going to compute the bounds $Y^{(j)}$ and $Z^{(j)}(r)$ and the associated radii polynomials, and then prove that for some positive r each radii polynomial $p_j(r)$ is negative, which will yield (for each $s \in [0,1]$) the existence of a unique zero $a(\beta_s)$ of $F(\beta_s, \cdot)$ in the ball of radius r around $\bar{a}(s)$. At this point we will know that $\bar{a}(s)$ defines an approximate parameterization of the stable manifold, with an error bound controlled by r. We will use this in Section 9.4 to prove the existence of a homoclinic orbit for all $\beta \in [\beta_0, \beta_1]$. Moreover, derivatives of the manifold with respect to θ, which will be needed in Section 9.4.1, can also be approximately computed with rigorous control on the error bound, see Lemma 9.3.8.

9.3.3 The bound Y

In this section we focus on the bound Y defined in (9.6). Let $|A|$ denote the component-wise absolute value of A. In order to define the bound we are looking for, we try to bound every term $(T(\beta_s, \bar{a}(s)) - \bar{a}(s))_\alpha^{(j)}$ with $|\alpha| \ge 0$ and $j = 1, 2, 3, 4$:

$$\begin{aligned} &\left|(T(\beta_s, \bar{a}(s)) - \bar{a}(s))_\alpha^{(j)}\right| = \left|(AF(\beta_s, \bar{a}(s)))_\alpha^{(j)}\right| \\ &\quad \le \Bigg(|A| \Bigg(|F(\beta_0, \bar{a}(0))| + |D_a F(\beta_0, \bar{a}(0))\Delta\bar{a} + D_\beta F(\beta_0, \bar{a}(0))\Delta\beta| \\ &\qquad + \frac{1}{2} \max_{s \in [0,1]} \left| D_{aa}^2 F(\beta_s, \bar{a}(s))(\Delta\bar{a})^2 + 2D_{a\beta}^2 F(\beta_s, \bar{a}(s))\Delta\bar{a}\Delta\beta + D_{\beta\beta}^2 F(\beta_s, \bar{a}(s))(\Delta\beta)^2 \right| \Bigg) \Bigg)_\alpha^{(j)}. \end{aligned}$$

A straightforward calculation (using that $|\lambda(\beta)| = 1$ and computing the derivatives of λ and V with respect to β) yields that, for all $|\alpha| \geq 0$,

$$\frac{1}{2}\max_{s\in[0,1]}\left|D^2_{aa}F_\alpha(\beta_s,\bar a(s))(\Delta\bar a)^2 + 2D^2_{a\beta}F_\alpha(\beta_s,\bar a(s))\Delta\bar a\Delta\beta + D^2_{\beta\beta}F_\alpha(\beta_s,\bar a(s))(\Delta\beta)^2\right| \leq G_\alpha,$$

where

$$G_\alpha := \begin{cases} 0 & \alpha = (0,0), \\ \dfrac{1}{2}\left(\dfrac{1}{4}\sqrt{\dfrac{4+3\beta_1^2}{(2-\beta_1)^3(2+\beta_1)^3}}\begin{pmatrix}0\\1\\2\\3\end{pmatrix} + \dfrac{1}{4}\dfrac{1}{(2-\beta_1)(2+\beta_1)}\begin{pmatrix}0\\0\\2\\6\end{pmatrix}\right)(\Delta\beta)^2 & |\alpha| = 1, \\ \dfrac{1}{4}\sqrt{\dfrac{(\alpha_1+\alpha_2)^2}{2-\beta_1} + \dfrac{(\alpha_1-\alpha_2)^2}{2+\beta_0}}\Delta\beta\Delta\bar a_\alpha + \begin{pmatrix}\left|\Delta\bar a^{(1)} \star \Delta\bar a^{(2)}\right|_\alpha \\ 0 \\ 0 \\ \left|\Delta\beta\Delta\bar a^{(3)}_\alpha\right|\end{pmatrix} & |\alpha| \geq 2. \end{cases}$$

Since $(\bar a(s))_\alpha = 0$ for all $|\alpha| \geq N$ and $F(\beta,\cdot)$ is quadratic in a, we have that $F_\alpha(\beta_s,\bar a(s))$ vanishes as soon as $|\alpha| \geq 2N-1$. Therefore, we define $\tilde F$ component-wise by

$$\tilde F_\alpha = \begin{cases} |F(\beta_0,\bar a(0))|_\alpha + |D_aF(\beta_0,\bar a(0))\Delta\bar a + D_\beta F(\beta_0,\bar a(0))\Delta\beta|_\alpha + G_\alpha & |\alpha| < 2N-1, \\ 0 & |\alpha| \geq 2N-1, \end{cases}$$

and then set

$$Y^{(j)} = \left\|\left(|A|\tilde F\right)^{(j)}\right\|_{1,\nu},$$

so that

$$\left\|(T(\beta_s,\bar a(s)) - \bar a(s))^{(j)}\right\|_{1,\nu} \leq Y^{(j)} \qquad \text{for } j = 1,\dots,4,\ s\in[0,1].$$

9.3.4 The bound Z

In this section we derive the bound Z defined in (9.7). Let $b, c \in B_r(0)$. We split $D_aT(\beta_s,\bar a(s)+b)c$ in three terms which will be easier to bound separately. For each $j = 1,\dots,4$,

$$\begin{aligned}\left\|(D_aT(\beta_s,\bar a(s)+b)c)^{(j)}\right\|_{1,\nu} &= \left\|\left(\left(I - AD_aF(\beta_s,\bar a(s)+b)\right)c\right)^{(j)}\right\|_{1,\nu} \\ &\leq \left\|\left(\left(I-AA^\dagger\right)c\right)^{(j)}\right\|_{1,\nu} + \left\|\left(A\left(D_aF(\beta_s,\bar a(s)+b) - A^\dagger\right)c\right)^{(j)}\right\|_{1,\nu} \\ &\leq \left\|\left(\left(I-AA^\dagger\right)c\right)^{(j)}\right\|_{1,\nu} + \left\|\left(A\left(D_aF(\beta_s,\bar a(s)) - A^\dagger\right)c\right)^{(j)}\right\|_{1,\nu} \\ &\qquad + \left\|\left(AD^2_{aa}F(\beta_s,\bar a(s))(b,c)\right)^{(j)}\right\|_{1,\nu} \\ &\leq Z_0^{(j)}r + Z_1^{(j)}r + Z_2^{(j)}r^2.\end{aligned}$$

The bounds $Z_i := \left(Z_i^{(1)}, Z_i^{(2)}, Z_i^{(3)}, Z_i^{(4)}\right) \in \mathbb{R}^4$ $(i = 0,1,2)$ are given in the following subsections.

The bound Z_0

From the definitions of A and $A^\dagger$ we get

$$I - AA^\dagger = \begin{pmatrix} I_{\frac{4N(N+1)}{2}} - JD_aF^{[N]}(\beta_0,\bar a(0)) & 0 & \\ 0 & 0 & \\ & & \ddots \end{pmatrix}.$$

The finite matrix $B = I_{2N(N+1)} - JD_aF^{[N]}(\beta_0, \bar{a}(0))$ can be computed using interval arithmetic. To obtain the bound Z_0 we only need to compute the operator norm of B (as acting on $(\ell^1_\nu)^4$). This is the content of the following lemma.

Lemma 9.3.5. *Let $h = (h_\alpha)_{\alpha\in\mathbb{N}^2} \in \ell^1_\nu$ (with $h_\alpha \in \mathbb{C}$ for all α) and Γ a linear operator acting on ℓ^1_ν. Then*

$$\sup_{\|h\|_{1,\nu}=1} \|\Gamma h\|_{1,\nu} = \sup_{\alpha\in\mathbb{N}^2} \frac{1}{\nu^{|\alpha|}} \sum_{\alpha'\in\mathbb{N}^2} |\Gamma_{\alpha',\alpha}| \nu^{|\alpha'|}.$$

In particular, if Γ consists in a finite block $\Gamma^{[N]}$ of size $N(N+1)/2 \times N(N+1)/2$ and a diagonal tail $(\gamma_\alpha)_{|\alpha|\geq N}$

$$\Gamma = \begin{pmatrix} \Gamma^{[N]} & & 0 & \\ & \gamma_{N,0} & & \\ 0 & & \gamma_{N-1,1} & \\ & & & \ddots \end{pmatrix},$$

then

$$\sup_{\|h\|_{1,\nu}=1} \|\Gamma h\|_{1,\nu} = \max\left(\max_{|\alpha|<N} \frac{1}{\nu^{|\alpha|}} \sum_{|\alpha'|<N} |\Gamma_{\alpha',\alpha}| \nu^{|\alpha'|}, \sup_{|\alpha|\geq N} |\gamma_\alpha| \right).$$

Hence, we define

$$K^{(i,j)}(B) := \max_{0\leq|\alpha|<N} \frac{1}{\nu^{|\alpha|}} \sum_{|\alpha'|<N} |B^{(i,j)}_{\alpha',\alpha}| \nu^{|\alpha'|}, \tag{9.17}$$

with the notation $B^{(i,j)}_{\alpha',\alpha}$ introduced in Remark 9.3.4, and set

$$Z_0^{(i)} = \sum_{j=1}^{4} K^{(i,j)}(B),$$

to obtain

$$\left\| \left(\left(I - AA^\dagger \right) c \right)^{(j)} \right\|_{1,\nu} \leq Z_0^{(j)} r, \quad \text{for } j = 1, \dots, 4. \tag{9.18}$$

The bound Z_1

This term is the most involved one to bound tightly, so again we split it into several parts that we bound separately. For each $j = 1, \dots, 4$,

$$\begin{aligned}
\left\| \left(A \left(D_aF(\beta_s, \bar{a}(s)) - A^\dagger \right) c \right)^{(j)} \right\|_{1,\nu} \leq & \left\| \left(|A| \left| \left(D_aF(\beta_0, \bar{a}(s)) - A^\dagger \right) c \right| \right)^{(j)} \right\|_{1,\nu} \\
& + \left\| \left(|A| \max_{\eta\in[0,1]} |\Delta\beta| \left| D^2_{\beta a}F(\beta_\eta, \bar{a}(s))c \right| \right)^{(j)} \right\|_{1,\nu} \\
\leq & \left\| \left(|A| \left| \left(D_aF(\beta_0, \bar{a}(0)) - A^\dagger \right) c \right| \right)^{(j)} \right\|_{1,\nu} \\
& + \left\| \left(|A| \left| D^2_{aa}F(\beta_0, \bar{a}(0))(\Delta\bar{a}, c) \right| \right)^{(j)} \right\|_{1,\nu} \\
& + \left\| \left(|A| \max_{\eta\in[0,1]} |\Delta\beta| \left| D^2_{\beta a}F(\beta_\eta, \bar{a}(s))c \right| \right)^{(j)} \right\|_{1,\nu}.
\end{aligned}$$

Let us focus first on the first term. Since

$$D_aF^{[N]}(\beta_0, \bar{a}(0))c^{[N]} = \left(D_aF(\beta_0, \bar{a}(0))c \right)^{[N]},$$

we get that

$$\left(\left(D_aF(\beta_0,\bar{a}(0))-A^\dagger\right)c\right)^{[N]}=0.$$

For the tail $|\alpha|\geq N$ we find

$$d_\alpha:=\left(\left(D_aF(\beta_0,\bar{a}(0))-A^\dagger\right)c\right)_\alpha=\begin{pmatrix}c_\alpha^{(2)}+(\bar{a}(0)^{(1)}\star c^{(2)})_\alpha+(\bar{a}(0)^{(2)}\star c^{(1)})_\alpha\\ c_\alpha^{(3)}\\ c_\alpha^{(4)}\\ -c_\alpha^{(1)}-\beta_0 c_\alpha^{(3)}\end{pmatrix},$$

which we estimate by

$$\begin{aligned}\left\|d_\alpha^{(1)}\right\|_{1,\nu}&\leq\left(1+\left\|\bar{a}(0)^{(1)}\right\|_{1,\nu}+\left\|\bar{a}(0)^{(2)}\right\|_{1,\nu}\right)r\\ \left\|d_\alpha^{(2)}\right\|_{1,\nu}&\leq r\\ \left\|d_\alpha^{(3)}\right\|_{1,\nu}&\leq r\\ \left\|d_\alpha^{(4)}\right\|_{1,\nu}&\leq(1+\beta_0)\,r.\end{aligned}$$

Now we use Lemma 9.3.5 again and from the fact that $|(n-k)\lambda(\beta_0)+k\lambda^*(\beta_0)|\geq n|\Re(\lambda(\beta_0))|=n\frac{\sqrt{2-\beta_0}}{2}$ we infer that

$$\left\|(Ad)^{(j)}\right\|_{1,\nu}\leq\frac{2}{N\sqrt{2-\beta_0}}\left\|d^{(j)}\right\|_{1,\nu},$$

and we are done with the first term. For the second term

$$D^2_{aa}F_\alpha(\beta_0,\bar{a}(0))(\Delta\bar{a},c)=\begin{pmatrix}(\Delta\bar{a}^{(1)}\star c^{(2)})_\alpha+(\Delta\bar{a}^{(2)}\star c^{(1)})_\alpha\\ 0\\ 0\\ 0\end{pmatrix}.$$

Again we use Lemma 9.3.5 to obtain

$$\left\|\left(|A|\left|D^2_{aa}F_0(\bar{a}(0))(\Delta\bar{a},c)\right|\right)^{(j)}\right\|_{1,\nu}\leq$$
$$\begin{cases}\max\left(K^{(1,1)}(J),\frac{2}{N\sqrt{2-\beta_0}}\right)\left(\|\Delta\bar{a}^{(1)}\|_{1,\nu}+\|\Delta\bar{a}^{(2)}\|_{1,\nu}\right)r & j=1,\\ K^{(j,1)}(J)\left(\|\Delta\bar{a}^{(1)}\|_{1,\nu}+\|\Delta\bar{a}^{(2)}\|_{1,\nu}\right)r & j=2,3,4,\end{cases}$$

where we recall that J is the block of A corresponding to the floating point data, see (9.16), and $K^{(i,j)}$ is defined by (9.17). Finally, computing the derivative of λ with respect to β, we get that

$$\max_{\eta\in[0,1]}|\Delta\beta|\left|D^2_{\beta a}F(\beta_\eta,\bar{a}(s))c\right|_\alpha\leq\begin{cases}0 & |\alpha|<2,\\ \Delta\beta\left(\frac{|\alpha|}{2\sqrt{(2-\beta_1)(2+\beta_1)}}|c_\alpha|+\begin{pmatrix}0\\0\\0\\\left|c_\alpha^{(3)}\right|\end{pmatrix}\right) & |\alpha|\geq 2.\end{cases}$$

Now we need the following lemma, which is a slightly modified version of Lemma 9.3.5.

Lemma 9.3.6. *Let $c = (c_\alpha)_{\alpha\in\mathbb{N}^2} \in \left(\ell^1_\nu\right)^4$. We denote by e the vector such that for all α, $e_\alpha = |\alpha| c_\alpha$. Then for all $1 \le j \le 4$*

$$\left\|(Ae)^{(i)}\right\|_{1,\nu} \le \max\left(\sum_{j=1}^{4} \tilde{K}^{(i,j)}(J), \frac{2}{\sqrt{2-\beta_0}}\right) r$$

where

$$\tilde{K}^{(i,j)}(J) := \max_{|\alpha|<N}\left(\frac{|\alpha|}{\nu^{|\alpha|}} \sum_{|\alpha'|<N} \left|J^{(i,j)}_{\alpha',\alpha}\right| \nu^{|\alpha'|}\right).$$

Using Lemmas 9.3.5 and 9.3.6 we infer that

$$\left\|\left(|A| \max_{\eta\in[0,1]} |\Delta\beta| \left|D^2_{\beta a} F(\beta_\eta, \bar{a}(s))c\right|\right)^{(j)}\right\|_{1,\nu} \le \begin{cases} \Delta\beta\left(\overline{K}_j + K^{(j,4)}(J)\right) r & j = 1,2,3 \\ \Delta\beta\left(\overline{K}_4 + \max\left(K^{(4,4)}(J), \frac{2}{N\sqrt{2-\beta_0}}\right)\right) r & j = 4, \end{cases}$$

where

$$\overline{K}_i := \frac{\max\left(\sum_{j=1}^4 \tilde{K}^{(i,j)}(J), \frac{2}{\sqrt{2-\beta_0}}\right)}{2\sqrt{(2-\beta_1)(2+\beta_1)}}.$$

Finally, putting everything together, we define

$$\begin{aligned}
Z_1^{(1)} &= \frac{2\left(1 + \left\|\bar{a}(0)^{(1)}\right\|_{1,\nu} + \left\|\bar{a}(0)^{(2)}\right\|_{1,\nu}\right)}{N\sqrt{2-\beta_0}} \\
&\quad + \max\left(K^{(1,1)}(J), \frac{2}{N\sqrt{2-\beta_0}}\right)\left(\left\|\Delta\bar{a}^{(1)}\right\|_{1,\nu} + \left\|\Delta\bar{a}^{(2)}\right\|_{1,\nu}\right) + \Delta\beta\left(\overline{K}_1 + K^{(1,4)}(J)\right), \\
Z_1^{(2)} &= \frac{2}{N\sqrt{2-\beta_0}} + K^{(2,1)}(J)\left(\left\|\Delta\bar{a}^{(1)}\right\|_{1,\nu} + \left\|\Delta\bar{a}^{(2)}\right\|_{1,\nu}\right) + \Delta\beta\left(\overline{K}_2 + K^{(2,4)}(J)\right), \\
Z_1^{(3)} &= \frac{2}{N\sqrt{2-\beta_0}} + K^{(3,1)}(J)\left(\left\|\Delta\bar{a}^{(1)}\right\|_{1,\nu} + \left\|\Delta\bar{a}^{(2)}\right\|_{1,\nu}\right) + \Delta\beta\left(\overline{K}_3 + K^{(3,4)}(J)\right), \\
Z_1^{(4)} &= \frac{2(1+\beta_0)}{N\sqrt{2-\beta_0}} + K^{(4,1)}(J)\left(\left\|\Delta\bar{a}^{(1)}\right\|_{1,\nu} + \left\|\Delta\bar{a}^{(2)}\right\|_{1,\nu}\right) \\
&\quad + \Delta\beta\left(\overline{K}_4 + \max\left(K^{(4,4)}(J), \frac{2}{N\sqrt{2-\beta_0}}\right)\right),
\end{aligned}$$

so that

$$\left\|\left(A\left(D_a F(\beta_s, \bar{a}(s)) - A^\dagger\right)c\right)^{(j)}\right\|_{1,\nu} \le Z_1^{(j)} r \qquad \text{for } j = 1,\dots,4,\ s \in [0,1].$$

The bound Z_2

Since

$$D^2_{aa} F_\alpha(\beta_s, \bar{a}(s))(b,c) = \begin{cases} 0 & |\alpha| \le 1, \\ \begin{pmatrix} (b^{(1)} \star c^{(2)})_\alpha + (b^{(2)} \star c^{(1)})_\alpha \\ 0 \\ 0 \\ 0 \end{pmatrix} & |\alpha| \ge 2, \end{cases}$$

we directly use one more time Lemma 9.3.5 and set

$$Z_2^{(1)} = 2\max\left(K^{(1,1)}(J), \frac{2}{N\sqrt{2-\beta_0}}\right), \quad Z_2^{(2)} = 2K^{(2,1)}(J), \quad Z_2^{(3)} = 2K^{(3,1)}(J),$$

$$Z_2^{(4)} = 2K^{(4,1)}(J),$$

so that

$$\left\| \left(AD^2_{aa} F(\beta_s, \bar{a}(s))(b,c) \right)^{(j)} \right\|_{1,\nu} \leq Z_2^{(j)} r^2 \qquad \text{for } j = 1,2,3,4,\ s \in [0,1].$$

9.3.5 Use of the uniform contraction principle and error bounds

Following (9.9), we set

$$p^{(j)}(r) := Y^{(j)} + \left(Z_0^{(j)} + Z_1^{(j)} - 1 \right) r + Z_2^{(j)} r^2, \quad \text{for } j = 1, \ldots, 4. \tag{9.19}$$

If we find an $r > 0$ such that $p^{(j)}(r) < 0$ for all $j = 1, \ldots, 4$, then according to Proposition 9.2.5 we have validated the numerical approximation $\bar{a}(s)$ of the local stable manifold for $\beta = \beta_s$, for every $s \in [0,1]$.

Proposition 9.3.7. *For every $s \in [0,1]$, let*

$$\overline{Q}_{\beta_s}(\theta) = \sum_{|\alpha|=0}^{N-1} \bar{a}_\alpha(s) \theta^\alpha$$

be the approximate parameterization of the complex local stable manifold that we have computed (for $\beta = \beta_s$). Assume that there exists an $r > 0$ such that $p^{(j)}(r) < 0$ for all $j = 1, \ldots, 4$. Then, for each $s \in [0,1]$, there exists a parameterization Q_{β_s} of the complex local stable manifold (for $\beta = \beta_s$) of the form

$$Q_{\beta_s}(\theta) = \sum_{|\alpha|=0}^{\infty} a_\alpha(s) \theta^\alpha,$$

which is well defined for all $\theta \in \mathbb{C}^2$ satisfying $|\theta|_\infty \leq \nu$. Let

$$\hat{h}_{\beta_s}(\theta) := Q_{\beta_s}(\theta) - \overline{Q}_{\beta_s}(\theta), \tag{9.20}$$

then we have the error bound $|\hat{h}_{\beta_s}(\theta)|_\infty \leq r$ for all $|\theta|_\infty \leq \nu$. These statements still hold true for the real (approximate and exact) local stable manifold, defined by

$$P_{\beta_s}(\theta) := Q_{\beta_s}(\theta_1 + \mathrm{i}\theta_2, \theta_1 - \mathrm{i}\theta_2) \tag{9.21}$$

$$\overline{P}_{\beta_s}(\theta) := \overline{Q}_{\beta_s}(\theta_1 + \mathrm{i}\theta_2, \theta_1 - \mathrm{i}\theta_2) \tag{9.22}$$

for all $\theta \in \mathbb{R}^2$ satisfying $|\theta|_2 := \sqrt{\theta_1^2 + \theta_2^2} \leq \nu$.

Proof. Proposition 9.2.5 yields that, for each $s \in [0,1]$, there exists a unique fixed point $a(s)$ of $T(\beta_s, \cdot)$ in the ball of radius r around $\bar{a}(s)$. The operator A is injective since its non-diagonal part J is invertible. The latter follows from the fact that, see (9.18),

$$\| I_{2N(N+1)} - J D_a F^{[N]}(\beta_0, \bar{a}(0)) \|_{B(X^{[N]}, X^{[N]})} \leq \max_{1 \leq j \leq 4} Z_0^{(j)} < 1,$$

where the final inequality is implied by $p_j(r) < 0$. Here the operator norm on $X^{[N]} \cong \mathbb{R}^{2N(N+1)}$ is induced by the one on $X = (\ell^1_\nu)^4$. Hence the fixed point $a(s)$ of T solves $F(\beta_s, a(s)) = 0$. By

construction Q_{β_s} is a parameterization of the local stable manifold defined for $|\theta|_\infty \leq \nu$, and for such θ,

$$\begin{aligned}
\left|Q_{\beta_s}(\theta) - \overline{Q}_{\beta_s}(\theta)\right|_\infty &= \left|\sum_{|\alpha|=0}^{\infty} \left(a_\alpha(s) - \bar{a}_\alpha(s)\right)\theta^\alpha\right|_\infty \\
&= \max_{j=1,\ldots,4}\left|\sum_{|\alpha|=0}^{\infty}\left(a_\alpha^{(j)}(s) - \bar{a}_\alpha^{(j)}(s)\right)\theta^\alpha\right| \\
&\leq \max_{j=1,\ldots,4}\sum_{|\alpha|=0}^{\infty}\left|a_\alpha^{(j)}(s) - \bar{a}_\alpha^{(j)}(s)\right|\nu^{|\alpha|} \\
&= \max_{j=1,\ldots,4}\left\|a^{(j)}(s) - \bar{a}^{(j)}(s)\right\|_{1,\nu} \\
&\leq r. \qquad \square
\end{aligned}$$

In the following section we use these approximations to rigorously prove the existence of homoclinic orbits for every parameter β in $[0.5, 1.9]$. To do so, we will also need control on the derivative of the parameterization P_{β_s}, which is provided by the theory of analytic functions. Define

$$h_{\beta_s}(\theta) := P_{\beta_s}(\theta) - \overline{P}_{\beta_s}(\theta), \quad \theta \in \mathbb{R}^2,\ |\theta|_2 \leq \nu. \tag{9.23}$$

For all $s \in [0,1]$, the function $\hat{h}_{\beta_s}$, defined by (9.20), is analytic. Since $h_{\beta_s}(\theta) = \hat{h}_{\beta_s}(\theta_1 + \mathrm{i}\theta_2, \theta_1 - \mathrm{i}\theta_2)$, we can control the derivative of h_{β_s} (on a smaller domain) by a bound on $\hat{h}_{\beta_s}$. This is the content of the following lemma, of which the proof can be found in [172].

Lemma 9.3.8. *Assume that $\hat{h} : D_{\infty,\nu}(\mathbb{C}^2) \subset \mathbb{C}^2 \to \mathbb{C}^4$ is analytic, where*

$$D_{\infty,\nu}(\mathbb{C}^2) := \left\{\theta \in \mathbb{C}^2,\ |\theta|_\infty \leq \nu\right\},$$

and $\delta > 0$ is such that

$$\max_{\theta \in D_{\infty,\nu}(\mathbb{C}^2)}\left|\hat{h}(\theta)\right|_\infty \leq \delta. \tag{9.24}$$

Consider $h : D_{2,\nu}(\mathbb{R}^2) \subset \mathbb{R}^2 \to \mathbb{R}^4$ defined by $h(\theta) = \hat{h}(\theta_1 + \mathrm{i}\theta_2, \theta_1 - \mathrm{i}\theta_2)$, where

$$D_{2,\nu}(\mathbb{R}^2) := \left\{\theta \in \mathbb{R}^2,\ |\theta|_2 \leq \nu\right\}.$$

Then for any $\rho < \nu$ we have

$$\max_{\theta \in D_{2,\rho}(\mathbb{R}^2)}\left|\frac{\partial h^{(j)}}{\partial \theta_i}(\theta)\right|_\infty \leq \frac{4\pi}{\nu \ln(\frac{\nu}{\rho})}\delta \qquad \text{for } j = 1,\ldots,4,\ i = 1,2. \tag{9.25}$$

9.4 Parameterized families of symmetric homoclinic orbits

In this section, we apply the technique of Section 9.2 in a Chebyshev series setting to rigorously prove existence of parameterized families of symmetric homoclinic orbits. More precisely, we present all necessary estimates and bounds in order to demonstrate that solutions of (9.3) exist for all $\beta \in [0.5, 1.9]$.

9.4.1 A projected boundary value problem formulation

We begin by transforming the symmetric homoclinic orbit problem (9.3) into a projected boundary value problem (BVP). In order to set up the projected BVP, we first use the symmetry of the orbit to simplify the problem and therefore solve only for "half of the orbit". The following lemma provides a strategy to do this.

Lemma 9.4.1. *Let u_0, u_2 and t_0 be arbitrary numbers, and let $u(t)$ be the solution of the initial value problem*

$$\begin{cases} u''''(t) + \beta u''(t) + e^{u(t)} - 1 = 0, \\ (u(t_0), u'(t_0), u''(t_0), u'''(t_0)) = (u_0, 0, u_2, 0)\,. \end{cases}$$

Then $u(-t + 2t_0) = u(t)$ for all t for which the solution u is defined.

Proof. It is straightforward to verify that $u(-t + 2t_0)$ is also a solution of the initial value problem. By the theorem of existence and uniqueness for ODEs, it follows that $u(-t+2t_0) = u(t)$ for all t in the domain definition of u. □

Using the previous result, we fix a number $t_0 = L > 0$, and it follows that to solve (9.3), it is enough to solve

$$\begin{cases} u''''(t) + \beta u''(t) + e^{u(t)} - 1 = 0, \\ u'(-L) = 0, \quad u'''(-L) = 0, \\ \lim_{t\to\infty} (u(t), u'(t), u''(t), u'''(t)) = 0. \end{cases} \tag{9.26}$$

The idea now is to modify the boundary value problem (9.26) in a way that the boundary value at $t = \infty$ is removed by a projected boundary value at $t = L$ where we impose at that time that $(u(L), u'(L), u''(L), u'''(L)) \in W^s_{\text{loc}}(0)$, a local stable manifold at 0. In order to achieve this step, we use the theory of Section 9.3 to obtain a real-valued parameterization P_β of $W^s_{\text{loc}}(0)$ at the parameter value $\beta \in [0.5, 1.9]$:

$$P_\beta(\theta) = Q_\beta(\theta_1 + \mathrm{i}\theta_2, \theta_1 - \mathrm{i}\theta_2) = \sum_{|\alpha|=0}^{\infty} a_\alpha(\beta)(\theta_1 + \mathrm{i}\theta_2)^{\alpha_1}(\theta_1 - \mathrm{i}\theta_2)^{\alpha_2},$$

which is well-defined for all $\theta \in D_{2,\tilde\nu}(\mathbb{R}^2) = \left\{\theta \in \mathbb{R}^2 : |\theta|_2 = \sqrt{\theta_1^2 + \theta_2^2} \le \tilde\nu\right\}$, where the size $\tilde\nu = \tilde\nu(\beta)$ of the domain of P_β changes as the parameter $\beta \in [0.5, 1.9]$ varies. Using the parameterization, we impose that

$$(u(L), u'(L), u''(L), u'''(L))^T = P_\beta(\theta) \tag{9.27}$$

for some $\theta \in D_{2,\tilde\nu}(\mathbb{R}^2)$, which implies that the orbit lies in the stable manifold. This introduces an indeterminacy that needs to be resolved. Namely, there is a one parameter family of pairs (L, θ) solving (9.27) while describing the same orbit. To overcome this, we impose that $\theta \in \partial D_{2,\rho}(\mathbb{R}^2) = \left\{\theta \in \mathbb{R}^2 : |\theta|_2 = \sqrt{\theta_1^2 + \theta_2^2} = \rho\right\}$, for some fixed $\rho < \tilde\nu$, and we solve for the angle ψ. More precisely, we consider θ such that $\sqrt{\theta_1^2 + \theta_2^2} = \rho$ by setting $\theta_1 + \mathrm{i}\theta_2 = \rho e^{\mathrm{i}\psi}$ for some $\psi \in [0, 2\pi)$. In this case, the evaluation of the parameterization of the local stable manifold along $\partial D_{2,\rho}(\mathbb{R}^2)$ reduces to

$$\begin{aligned} P_\beta(\psi) &= \sum_{|\alpha|=0}^{\infty} a_\alpha(\beta)(\theta_1 + \mathrm{i}\theta_2)^{\alpha_1}(\theta_1 - \mathrm{i}\theta_2)^{\alpha_2} \\ &= \sum_{|\alpha|=0}^{\infty} a_\alpha(\beta)\rho^{\alpha_1} e^{\mathrm{i}\alpha_1\psi}\rho^{\alpha_2} e^{-\mathrm{i}\alpha_2\psi} \\ &= \sum_{|\alpha|=0}^{\infty} a_\alpha(\beta)\rho^{|\alpha|} e^{\mathrm{i}(\alpha_1-\alpha_2)\psi}. \end{aligned}$$

We slightly abuse notation by using the same notation P_β to denote both $P_\beta(\theta)$ and $P_\beta(\psi)$. We can therefore define the projected BVP

$$\begin{cases} u''''(t) + \beta u''(t) + e^{u(t)} - 1 = 0, \quad t \in [-L, L], \\ u'(-L) = 0, \quad u'''(-L) = 0, \\ (u(L), u'(L), u''(L), u'''(L))^T = P_\beta(\psi), \end{cases} \tag{9.28}$$

where $L > 0$ and $\psi \in [0, 2\pi)$ are variables. As in Section 9.1, we make the change of variables

$$(v^{(1)}, v^{(2)}, v^{(3)}, v^{(4)}) := (e^{u_1} - 1, u_2, u_3, u_4)$$

and set $v = (v^{(1)}, v^{(2)}, v^{(3)}, v^{(4)})$ to obtain that $v' = \Psi_\beta(v)$, where $\Psi_\beta : \mathbb{R}^4 \to \mathbb{R}^4$ is the vector field given by the right-hand side of (9.4). We rescale time via $t \mapsto t/L$ so that (9.28) becomes

$$\begin{cases} \dot{v} = L\Psi_\beta(v), \quad t \in [-1, 1], \\ v^{(2)}(-1) = 0, \quad v^{(4)}(-1) = 0, \\ v(1) = P_\beta(\psi). \end{cases} \tag{9.29}$$

A triplet (L, ψ, v) satisfying (9.29) thus corresponds to a symmetric homoclinic solution of the suspension bridge equation. The rest of this section is dedicated to applying the technique of Section 9.2 in a Chebyshev series setting to rigorously prove existence of parameterized families of solutions of the projected BVP (9.29) for all $\beta \in [0.5, 1.9]$. This begins by defining a zero finding problem $F = 0$ whose solutions correspond to symmetric homoclinic solutions of the suspension bridge equation.

9.4.2 Setting up the zero finding problem using Chebyshev series

Now that $v^{(i)}(t)$ is defined on $[-1, 1]$ and needs to solve a boundary value problem, describing $v^{(i)}(t)$ in terms of a Chebyshev series is a natural choice, see [105, 156, 60, 67]. Denote by $T_k : [-1, 1] \to \mathbb{R}$ the k-th Chebyshev polynomial with $k \geq 0$, where $T_0(t) = 1$, $T_1(t) = t$ and $T_{k+1}(t) = 2tT_k(t) - T_{k-1}(t)$ for $k \geq 1$. One way to characterize the Chebyshev polynomials is through the identity $T_k(t) = \cos(k \arccos t)$, from which it follows that $\|T_k\|_\infty = 1$, $T_k(1) = 1$, and $T_k(-1) = (-1)^k$.

For each $i = 1, 2, 3, 4$, we expand $v^{(i)}$ using a Chebyshev series expansion, that is

$$v^{(i)}(t) = x_0^{(i)} + 2\sum_{k=1}^{\infty} x_k^{(i)} T_k(t). \tag{9.30}$$

For each $i = 1, 2, 3, 4$, denote by $x^{(i)} := \{x_k^{(i)}\}_{k \geq 0}$ the infinite dimensional vector of Chebyshev coefficients of $v^{(i)}$. The vector field is analytic (polynomial) and therefore the solutions (if they exist) of the projected BVP (9.29) are analytic. By the Paley-Wiener theorem, this implies that the Chebyshev coefficients of each component of v decay geometrically to zero. Hence, there exists a number $\nu > 1$ such that $x^{(i)} \in \ell^1_\nu$ for each $i = 1, 2, 3, 4$, where

$$\ell^1_\nu = \left\{ a = (a_k)_{k \geq 0} : \|a\|_{1,\nu} := |a_0| + 2\sum_{k=1}^{\infty} |a_k| \nu^k < \infty \right\}.$$

We remark that throughout this section $\nu \geq 1$.

Remark 9.4.2 (Notation). *The decay rate ν in the definition of the Banach space ℓ^1_ν appears both in the current section and in Section 9.3. Both values need not to be the same. Therefore, to avoid confusion, we denote by $\tilde{\nu}$ the value from Section 9.3. Moreover, although the sequence space ℓ^1_ν as considered above is slightly different from the one used in Section 9.3, we nevertheless use the same notation, since the spaces and norms are completely analogous to those used in Section 9.3.2.*

The dual space can be characterized as follows.

Lemma 9.4.3. *The dual space $(\ell^1_\nu)^*$ is isomorphic to*

$$\ell^\infty_{\nu^{-1}} = \left\{ c = (c_k)_{k\geq 0} : \|c\|_{\infty,\nu^{-1}} := \max\left(|c_0|, \tfrac{1}{2}\sup_{k\geq 1} |c_k|\nu^{-k} \right) < \infty \right\}.$$

For all $a \in \ell^1_\nu$ and $c \in \ell^\infty_{\nu^{-1}}$ we have

$$\left| \sum_{k\geq 0} c_k a_k \right| \leq \|c\|_{\infty,\nu^{-1}} \|a\|_{1,\nu}. \tag{9.31}$$

The following lemma is analogous to Lemma 9.3.5.

Lemma 9.4.4. *Let $\Gamma \in B(\ell^1_\nu)$, the space of bounded linear operators from ℓ^1_ν to itself, acting as $(\Gamma a)_i = \sum_{j\geq 0} \Gamma_{i,j} a_j$. Define the weights $\omega = (\omega_k)_{k\geq 0}$ by $\omega_0 = 1$ and $\omega_k = 2\nu^k$ for $k \geq 1$. Then*

$$\|\Gamma\|_{B(\ell^1_\nu)} = \sup_{j\geq 0} \frac{1}{\omega_j} \sum_{i\geq 0} |\Gamma_{i,j}| \omega_i.$$

The Banach space of unknowns $x := (L, \psi, x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)})$ is

$$X := \mathbb{R}^2 \times (\ell^1_\nu)^4, \tag{9.32}$$

endowed with the norm

$$\|x\|_X := \max\left\{ |L|, |\psi|, \|x^{(1)}\|_{1,\nu}, \|x^{(2)}\|_{1,\nu}, \|x^{(3)}\|_{1,\nu}, \|x^{(4)}\|_{1,\nu} \right\}.$$

In terms of Chebyshev coefficients the differential equation $\dot v = L\Psi_\beta(v)$ becomes (see e.g. [156])

$$\begin{cases} f^{(1)}_k(\beta, x) := 2kx^{(1)}_k + L[x^{(2)}_{k\pm 1} + (x^{(1)} * x^{(2)})_{k\pm 1}] = 0, \\ f^{(2)}_k(\beta, x) := 2kx^{(2)}_k + Lx^{(3)}_{k\pm 1} = 0, \\ f^{(3)}_k(\beta, x) := 2kx^{(3)}_k + Lx^{(4)}_{k\pm 1} = 0, \\ f^{(4)}_k(\beta, x) := 2kx^{(4)}_k + L[-x^{(1)}_{k\pm 1} - \beta x^{(3)}_{k\pm 1}] = 0, \end{cases} \tag{9.33}$$

for all $k \geq 1$. Here $x^{(i)}_{k\pm 1} := x^{(i)}_{k+1} - x^{(i)}_{k-1}$, and $*$ denotes the discrete convolution product $* : \ell^1_\nu \times \ell^1_\nu \to \ell^1_\nu$ defined as follows. Let $a, b \in \ell^1_\nu$, then for all $k \geq 0$ the k-th entry of the convolution product $a * b$ is given by

$$(a * b)_k = \sum_{\substack{k_1+k_2=k \\ k_1,k_2\in\mathbb{Z}}} a_{|k_1|} b_{|k_2|}.$$

The choice of norm and convolution product is justified by the fact ℓ^1_ν is a Banach algebra, that is $\|a * b\|_{1,\nu} \leq \|a\|_{1,\nu} \|b\|_{1,\nu}$, for all $a, b \in \ell^1_\nu$.

The symmetry conditions $v^{(2)}(-1) = v^{(4)}(-1) = 0$ reduce to

$$\eta^{(1)}(\beta, x) := x^{(2)}_0 + 2\sum_{k=1}^\infty x^{(2)}_k (-1)^k = 0, \tag{9.34}$$

$$\eta^{(2)}(\beta, x) := x^{(4)}_0 + 2\sum_{k=1}^\infty x^{(4)}_k (-1)^k = 0, \tag{9.35}$$

and the boundary conditions $v(1) = P_\beta(\psi)$ become

$$f^{(i)}_0(\beta, x) := x^{(i)}_0 + 2\sum_{k=1}^\infty x^{(i)}_k - P^{(i)}_\beta(\psi) = 0 \qquad \text{for } i = 1, 2, 3, 4. \tag{9.36}$$

The full set of equations that we want to solve is thus $F(\beta, x) = 0$, where

$$F := \left(\eta^{(1)}, \eta^{(2)}, F^{(1)}, F^{(2)}, F^{(3)}, F^{(4)}\right), \qquad \text{with } F^{(i)} := \left\{f_k^{(i)}\right\}_{k\geq 0}. \tag{9.37}$$

In order to solve rigorously the problem $F(\beta, x) = 0$ in the Banach space X, for all $\beta \in [0.5, 1.9]$, we apply the radii polynomial approach of Section 9.2.

9.4.3 The finite dimensional reduction of the zero finding problem

Having identified the operator F given in (9.37) whose zeros correspond to symmetric homoclinic orbits of (9.2), the next step is to compute numerical approximations, which requires considering a finite dimensional projection of the Banach space X given in (9.32). Given a sequence $a = (a_k)_{k\geq 0} \in \ell^1_\nu$, denote by $a^{[m]} = (a_0, \dots, a_{m-1}) \in \mathbb{R}^m$ the Galerkin projection of a onto the first m Chebyshev coefficients. Given an infinite dimensional vector $x = (L, \psi, x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) \in X$, denote

$$x^{[m]} := \left(L, \psi, (x^{(1)})^{[m]}, (x^{(2)})^{[m]}, (x^{(3)})^{[m]}, (x^{(4)})^{[m]}\right) \in \mathbb{R}^2 \times (\mathbb{R}^m)^4 \cong \mathbb{R}^{4m+2}. \tag{9.38}$$

In this context, the finite dimensional Banach space $\mathbb{R}^{4m+2}$ is the finite dimensional projection of $X = \mathbb{R}^2 \times (\ell^1_\nu)^4$, and $x^{[m]}$ is the finite dimensional projection of x. We slightly abuse the notation by denoting $x^{[m]} \in X$ as the vector built from $x^{[m]} \in \mathbb{R}^{4m+2}$ by padding each entry $(x^{(i)})^{[m]}$ $(i = 1, 2, 3, 4)$ with infinitely many zeros. The finite dimensional projection of F given in (9.37) is defined as

$$\begin{aligned} F^{[m]} : \mathbb{R} \times \mathbb{R}^{4m+2} &\to \mathbb{R}^{4m+2} \\ (\beta, x^{[m]}) &\mapsto F^{[m]}(\beta, x^{[m]}) := \left(F(\beta, x^{[m]})\right)^{[m]}. \end{aligned}$$

We want to compute on $F^{[m]}$, but it depends on the parameterization P_{β_s}, which itself depends on *infinitely* many Taylor coefficients. To remedy this, we consider a finite dimensional reduction of P_{β_s}. Recalling (9.22), for every s, denote by $\overline{P}_{\beta_s}$ the computable approximation of the stable manifold given by

$$\begin{aligned} \overline{P}_{\beta_s}(\psi) &= \sum_{|\alpha|<N} (\bar{a}_{0,\alpha} + s(\bar{a}_{1,\alpha} - \bar{a}_{0,\alpha}))\rho^{|\alpha|} e^{\mathrm{i}\psi(\alpha_1-\alpha_2)} \\ &= \sum_{|\alpha|<N} (\bar{a}_{0,\alpha} + s\Delta\bar{a}_\alpha)\rho^{|\alpha|} e^{\mathrm{i}\psi(\alpha_1-\alpha_2)} \\ &= \sum_{|\alpha|<N} \bar{a}_{0,\alpha}\rho^{|\alpha|} e^{\mathrm{i}\psi(\alpha_1-\alpha_2)} + s\sum_{|\alpha|<N} \Delta\bar{a}_\alpha\rho^{|\alpha|} e^{\mathrm{i}\psi(\alpha_1-\alpha_2)} \\ &= \overline{P}_{\beta_0}(\psi) + s\Delta\overline{P}(\psi), \end{aligned} \tag{9.39}$$

where $\bar{a}_{0,\alpha}$ and $\bar{a}_{1,\alpha}$ are the numerical approximations of the coefficients of the stable manifold for β_0 and β_1 respectively.

Finally, let $\overline{F}(\beta, x^{[m]})$ denote the finite dimensional projection of the operator using a Galerkin projection on the last four components and using the finite dimensional approximation $\overline{P}_\beta$ for the parameterization of the stable manifold. More explicitly,

$$\overline{F}(\beta, x^{[m]}) := \left(\overline{\eta}^{(1)}(\beta, x^{[m]}), \overline{\eta}^{(2)}(\beta, x^{[m]}), \overline{F}^{(1)}(\beta, x^{[m]}), \dots, \overline{F}^{(4)}(\beta, x^{[m]})\right), \tag{9.40}$$

with $\overline{F}^{(i)}(\beta, x^{[m]}) := \left\{\overline{f}_k^{(i)}(\beta, x^{[m]})\right\}_{k=0}^{m-1}$ for $i = 1, 2, 3, 4$, and

$$\overline{\eta}^{(1)}(\beta, x^{[m]}) := x_0^{(2)} + 2\sum_{k=1}^{m-1} x_k^{(2)}(-1)^k, \qquad \overline{\eta}^{(2)}(\beta, x^{[m]}) := x_0^{(4)} + 2\sum_{k=1}^{m-1} x_k^{(4)}(-1)^k$$

$$\overline{f}_0^{(i)}(\beta, x^{[m]}) := x_0^{(i)} + 2\sum_{k=1}^{m-1} x_k^{(i)} - \overline{P}_\beta^{(i)}(\psi) = 0 \qquad \text{for } i = 1, 2, 3, 4,$$

while $\overline{f}_k^{(i)}(\beta, x^{[m]}) = f_k^{(i)}(\beta, x^{[m]})$ for all $k = 1, \ldots, m-1$, see (9.33). Having identified $\overline{F} : \mathbb{R} \times \mathbb{R}^{4m+2} \to \mathbb{R}^{4m+2} : (\beta, x^{[m]}) \mapsto \overline{F}(\beta, x^{[m]})$ defined in (9.40) as the finite dimensional reduction of F given in (9.37), we can apply the finite dimensional Newton's method to find numerical approximations. The next step is to define an infinite dimensional Newton-like operator $T : \mathbb{R} \times X \to X$ on which we apply the uniform contraction principle (via the radii polynomial approach of Section 9.2).

9.4.4 The Newton-like operator for the homoclinic orbit

Let $\beta_0 < \beta_1$ be two different parameter values, and consider two numerical approximations $\bar{x}_0$ and $\bar{x}_1$ such that $F(\beta_0, \bar{x}_0) \approx 0$ and $F(\beta_1, \bar{x}_1) \approx 0$. In practice we find $\bar{x}_i$ by solving $\overline{F}(\beta_i, \cdot) = 0$ numerically. For every $s \in [0, 1]$, set

$$\bar{x}_s = \bar{x}_0 + s\Delta\bar{x}, \qquad \Delta\bar{x} := \bar{x}_1 - \bar{x}_0$$

and

$$\beta_s = \beta_0 + s\Delta\beta, \qquad \Delta\beta := \beta_1 - \beta_0.$$

We denote

$$\bar{x}_s = (\bar{L}_s, \bar{\psi}_s, \bar{x}_s^{(1)}, \bar{x}_s^{(2)}, \bar{x}_s^{(3)}, \bar{x}_s^{(4)}) \in X$$

for $s \in [0, 1]$, and we recall that each $\bar{x}_s^{(j)}$ is obtained from $(\bar{x}_s^{(j)})^{[m]} \in \mathbb{R}^m$ by padding with zeros. Similarly, we denote

$$\Delta\bar{x} = (\Delta\bar{L}, \Delta\bar{\psi}, \Delta\bar{x}^{(1)}, \Delta\bar{x}^{(2)}, \Delta\bar{x}^{(3)}, \Delta\bar{x}^{(4)}).$$

We now construct a fixed point operator $T(\beta, x) = x - AF(\beta, x)$ so that it is a uniform contraction over the interval of parameters $[\beta_0, \beta_1]$, whose fixed points $x = x(\beta)$ correspond to zeros of $F(\beta, \cdot)$ at a given parameter value $\beta \in [\beta_0, \beta_1]$. The operator A is constructed as an approximate inverse of $DF(\beta_0, \bar{x}_0)$. Let $\bar{x}_0$ be such that $\overline{F}(\beta_0, \bar{x}_0) \approx 0$ and let $A^{[m]} \approx (D\overline{F}(\beta_0, \bar{x}_0))^{-1}$ be a numerical approximation of the inverse of the Jacobian matrix. We decompose the $(4m+2) \times (4m+2)$ matrix $A^{[m]}$, into 36 blocks as

$$A^{[m]} = \begin{pmatrix} A_{1,1}^{[m]} & A_{1,2}^{[m]} & A_{1,3}^{[m]} & \cdots & A_{1,6}^{[m]} \\ A_{2,1}^{[m]} & A_{2,2}^{[m]} & A_{2,3}^{[m]} & \cdots & A_{2,6}^{[m]} \\ A_{3,1}^{[m]} & A_{3,2}^{[m]} & A_{3,3}^{[m]} & \cdots & A_{3,6}^{[m]} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{6,1}^{[m]} & A_{6,2}^{[m]} & A_{6,3}^{[m]} & \cdots & A_{6,6}^{[m]} \end{pmatrix}. \tag{9.41}$$

Here $A_{i,j}^{[m]}$ is scalar for $1 \le i, j \le 2$, $A_{i,j}^{[m]}$ is a row vector of length m for $1 \le i \le 2$, $3 \le j \le 6$, $A_{i,j}^{[m]}$ is a column vector of length m for $3 \le i \le 6$, $1 \le j \le 2$, and $A_{i,j}^{[m]}$ is a $m \times m$ matrix for $3 \le i, j \le 6$.

Definition 9.4.5 (Definition of A). *We extend this finite dimensional operator $A^{[m]} = \{A_{i,j}^{[m]} | 1 \le i, j \le 6\}$ to an operator $A = \{A_{i,j} | 1 \le i, j \le 6\}$ on X defined block-wise as*

- $A_{i,j} \in \mathbb{R}$ *for* $1 \le i, j \le 2$, *where* $A_{i,j} = A_{i,j}^{[m]}$;
- $A_{i,j} \in (\ell^1_\nu)^*$ *for* $1 \le i \le 2$ *and* $3 \le j \le 6$, *where* $A_{i,j}$ *is* $A_{i,j}^{[m]}$ *padded with zeros;*
- $A_{i,j} \in \ell^1_\nu$ *for* $3 \le i \le 6$ *and* $1 \le j \le 2$, *where* $A_{i,j}$ *is* $A_{i,j}^{[m]}$ *padded with zeros;*
- $A_{i,j} \in B(\ell^1_\nu, \ell^1_\nu)$ *for* $3 \le i, j \le 6$, *where*

$$(A_{i,j}x^{(j-2)})_k = \begin{cases} \left(A_{i,j}^{[m]}(x^{(j-2)})^{[m]}\right)_k & \text{if } 0 \le k \le m-1, \\ \frac{\delta_{i,j}}{2k} x_k^{(j-2)} & \text{if } k \ge m, \end{cases} \tag{9.42}$$

with $\delta_{i,j}$ the usual Kronecker delta.

Here $(\ell^1_\nu)^*$ is the dual of ℓ^1_ν. As an example, for $1 \le i \le 2$, $3 \le j \le 6$, we have $A_{i,j}a = A^{[m]}_{i,j}a^{[m]}$. The action of A on $x = (L, \psi, x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) \in X$ is thus

$$(Ax)^{(i)} = A_{i,1}L + A_{i,2}\psi + \sum_{j=3}^{6} A_{i,j}x^{(j-2)}, \qquad \text{for } 1 \le i \le 6,$$

where $(Ax)^{(i)} \in \mathbb{R}$ for $i = 1, 2$ and $(Ax)^{(i)} \in \ell^1_\nu$ for $i = 3, 4, 5, 6$.

We consider the Newton-like operator

$$T(\beta_s, x) = x - AF(\beta_s, x) \tag{9.43}$$

where $s \in [0, 1]$ and A is as in Definition 9.4.5.

Lemma 9.4.6. *Given the operator A as in Definition 9.4.5. Then $T : \mathbb{R} \times X \to X$.*

Proof. Consider $x \in X = \mathbb{R}^2 \times (\ell^1_\nu)^4$ and $\beta \in \mathbb{R}$. By construction of A, in particular the infinite diagonal tail chosen in (9.42), it is straightforward to verify that $(AF(\beta, x))^{(i)} \in \mathbb{R}$ for $i = 1, 2$ and $(AF(\beta, x))^{(i)} \in \ell^1_\nu$ for $3 \le i \le 6$. □

Showing the existence of parameterized fixed points of T defined in (9.43) is done by applying the general technique of Section 9.2. This requires computing the bounds $Y^{(j)}$ satisfying (9.6) and the bounds $Z^{(j)}$ satisfying (9.7) for $j = 1, \dots, 6$. We recall that since $X = \prod_{j=1}^{6} X_j = \mathbb{R}^2 \times (\ell^1_\nu)^4$, we have that $\|\cdot\|_{X_j}$ denotes the absolute value for $j = 1, 2$ and the ℓ^1_ν norm for $j = 3, 4, 5, 6$.

9.4.5 The Y bound for the homoclinic orbit problem

We recall the definition of the bounds $Y^{(j)}$ in (9.6). In our context, $Y^{(j)}$ is a bound satisfying

$$\sup_{s\in[0,1]} \|(AF(\beta_s, \bar{x}_s))^{(j)}\|_{X_j} \le Y^{(j)}.$$

We begin by expanding each component of

$$F(\beta_s, \bar{x}_s) = \left(\eta^{(1)}(\beta_s, \bar{x}_s), \eta^{(2)}(\beta_s, \bar{x}_s), F^{(1)}(\beta_s, \bar{x}_s), \dots, F^{(4)}(\beta_s, \bar{x}_s)\right)$$

as a polynomial in s. Given $s \in [0, 1]$ and $j = 1, 2, 3, 4$, denote $\bar{x}^{(j)}_s = \left(\bar{x}^{(j)}_{s,k}\right)_{k\ge 0}$.

First, $\eta^{(1)}(\beta_s, \bar{x}_s) = S^{(1)}_0 + sS^{(1)}_1$ and $\eta^{(2)}(\beta_s, \bar{x}_s) = S^{(2)}_0 + sS^{(2)}_1$, where

$$S^{(1)}_0 := \bar{x}^{(2)}_{0,0} + 2\sum_{k=1}^{m-1}(-1)^k\bar{x}^{(2)}_{0,k}, \qquad S^{(1)}_1 := \Delta\bar{x}^{(2)}_0 + 2\sum_{k=1}^{m-1}(-1)^k\Delta\bar{x}^{(2)}_k, \tag{9.44}$$

$$S^{(2)}_0 := \bar{x}^{(4)}_{0,0} + 2\sum_{k=1}^{m-1}(-1)^k\bar{x}^{(4)}_{0,k}, \qquad S^{(2)}_1 := \Delta\bar{x}^{(4)}_0 + 2\sum_{k=1}^{m-1}(-1)^k\Delta\bar{x}^{(4)}_k. \tag{9.45}$$

Let us now expand $F^{(1)}(\beta_s, \bar{x}_s), \dots, F^{(4)}(\beta_s, \bar{x}_s)$ as polynomials in s, and recall that their first component depend on the exact parameterization of the stable manifold P_{β_s} which involves infinitely many Taylor coefficients. The work from Section 9.3 provides the existence of a function $h_s : D_{2,\tilde{\nu}}(\mathbb{R}^2) \to \mathbb{R}^4$, see (9.23), such that

$$P_{\beta_s}(\theta) = \overline{P}_{\beta_s}(\theta) + h_s(\theta).$$

As before, we slightly abuse notation by denoting $P_{\beta_s}(\psi) = \overline{P}_{\beta_s}(\psi)+h_s(\psi)$, where $\theta_1+\mathrm{i}\theta_2 = \rho e^{\mathrm{i}\psi}$ for a fixed $\rho < \tilde{\nu}$. We then split the operator as

$$F(\beta_s, \bar{x}_s) = F^{(N)}(\beta_s, \bar{x}_s) + H_s(\bar{\psi}_s),$$

where $F^{(N)}$ denotes the full infinite dimensional F operator but evaluated using the (finitely computable) approximation of the parameterization $\overline{P}_{\beta_s}$ of order N, and where

$$H_s(\psi) := \begin{pmatrix} 0 \\ 0 \\ (h_s^{(1)}(\psi), 0, 0, \ldots) \\ (h_s^{(2)}(\psi), 0, 0, \ldots) \\ (h_s^{(3)}(\psi), 0, 0, \ldots) \\ (h_s^{(4)}(\psi), 0, 0, \ldots) \end{pmatrix}.$$

The size of $h_s^{(i)}(\psi)$ can be estimated by r_m using Proposition 9.3.7, where r_m is the validation radius for the manifold for $\beta_0 \le \beta \le \beta_1$. In addition, since we know that the zeroth order term in $h_s^{(i)}$ vanishes, i.e. $a_0(s) = \bar{a}_0(s) = 0$, we obtain a slightly sharper bound for any $\rho < \tilde{\nu}$:

$$\begin{aligned} |h_s^{(i)}(\psi)| &= \left| \sum_{|\alpha|=0}^{\infty} (a_\alpha^{(i)}(s) - \bar{a}_\alpha^{(i)}(s)) \rho^{|\alpha|} e^{\mathrm{i}\psi(\alpha_1-\alpha_2)} \right| \le \sum_{|\alpha|=1}^{\infty} |a_\alpha^{(i)}(s) - \bar{a}_\alpha^{(i)}(s)| \left(\frac{\rho}{\tilde{\nu}}\right)^{|\alpha|} \tilde{\nu}^{|\alpha|} \\ &\le \frac{\rho}{\tilde{\nu}} \sum_{|\alpha|=1}^{\infty} |a_\alpha^{(i)}(s) - \bar{a}_\alpha^{(i)}(s)| \tilde{\nu}^{|\alpha|} = \frac{\rho}{\tilde{\nu}} \|a^{(i)}(s) - \bar{a}^{(i)}(s)\|_{1,\tilde{\nu}} \le \frac{\rho}{\tilde{\nu}} r_m. \end{aligned}$$

Hence, we can estimate $H_s(\bar{\psi}_s)$ elementwise by

$$|H_s(\bar{\psi}_s)| \le \mu := \begin{pmatrix} 0 \\ 0 \\ (\frac{\rho}{\tilde{\nu}} r_m, 0, 0, \ldots) \\ (\frac{\rho}{\tilde{\nu}} r_m, 0, 0, \ldots) \\ (\frac{\rho}{\tilde{\nu}} r_m, 0, 0, \ldots) \\ (\frac{\rho}{\tilde{\nu}} r_m, 0, 0, \ldots) \end{pmatrix}.$$

Denoting

$$F^{(N)} = \left(\eta^{(1)}, \eta^{(2)}, F^{(1,N)}, F^{(2,N)}, F^{(3,N)}, F^{(4,N)}\right)$$

with $F^{(j,N)} = \left\{ f_0^{(j,N)}, f_1^{(j)}, f_2^{(j)}, f_3^{(j)}, \ldots \right\}$ for $= 1, 2, 3, 4$, we rewrite $f_0^{(j,N)}(\beta_s, \bar{x}_s)$ as a polynomial in s, where we use $\Delta\overline{P}$ as defined in (9.39):

$$\begin{aligned} f_0^{(j,N)}(\beta_s, \bar{x}_s) &= \bar{x}_{0,0}^{(j)} + s\Delta\bar{x}_0^{(j)} + 2\sum_{k=1}^{m-1} [\bar{x}_{0,k}^{(j)} + s\Delta\bar{x}_k^{(j)}] - \overline{P}_{\beta_s}^{(j)}(\bar{\psi}_s) \\ &= \left(\bar{x}_{0,0}^{(j)} + 2\sum_{k=1}^{m-1} \bar{x}_{0,k}^{(j)} - \overline{P}_{\beta_0}^{(j)}(\bar{\psi}_s) \right) + s \left(\Delta\bar{x}_0^{(j)} + 2\sum_{k=1}^{m-1} \Delta\bar{x}_k^{(j)} - \Delta\overline{P}^{(j)}(\bar{\psi}_s) \right) \\ &= \left(\bar{x}_{0,0}^{(j)} + 2\sum_{k=1}^{m-1} \bar{x}_{0,k}^{(j)} - \overline{P}_{\beta_0}^{(j)}(\bar{\psi}_0) \right) \\ &\quad + s \left(\Delta\bar{x}_0^{(j)} + 2\sum_{k=1}^{m-1} \Delta\bar{x}_k^{(j)} - \Delta\overline{P}^{(j)}(\bar{\psi}_0) - \Delta\bar{\psi}\frac{d}{d\psi}\overline{P}_{\beta_0}^{(j)}(\xi) \right) - s^2\Delta\bar{\psi}\frac{d}{d\psi}\Delta\overline{P}^{(j)}(\zeta) \\ &:= S_{0,0}^{(j+2)} + sS_{1,0}^{(j+2)} + s^2 S_{2,0}^{(j+2)}, \end{aligned} \tag{9.46}$$

for some ξ, ζ between $\bar\psi_0$ and $\bar\psi_1$ (using the mean value theorem). To obtain an explicit computable expression for $S^{(j+2)}_{1,0}$ and $S^{(j+2)}_{2,0}$, we determine

$$\begin{aligned}\frac{d}{d\psi}\overline{P}^{(j)}_{\beta_0}(\xi) &= \mathrm{i}\sum_{|\alpha|<N}\bar a_{0,\alpha}(\alpha_1-\alpha_2)\rho^{|\alpha|}e^{\mathrm{i}\xi(\alpha_1-\alpha_2)},\\ \frac{d}{d\psi}\Delta\overline{P}^{(j)}(\zeta) &= \mathrm{i}\sum_{|\alpha|<N}\Delta\bar a_{\alpha}(\alpha_1-\alpha_2)\rho^{|\alpha|}e^{\mathrm{i}\zeta(\alpha_1-\alpha_2)},\end{aligned} \tag{9.47}$$

by an interval arithmetic calculation, i.e., replacing ξ and ζ by the interval $[\bar\psi_0, \bar\psi_1]$.

For $k \geq 1$, we set $(j = 1, 2, 3, 4)$

$$f^{(j)}_k(\beta_s, \bar x_s) = S^{(j+2)}_{0,k} + S^{(j+2)}_{1,k}s + S^{(j+2)}_{2,k}s^2 + S^{(j+2)}_{3,k}s^3, \tag{9.48}$$

where the third order term is nonzero for $j = 1$ and $j = 4$ only. All terms are collected in Table 9.4.5. Then, it is possible to write the whole operator as

$$F(\beta_s, \bar x_s) = S_0 + sS_1 + s^2S_2 + s^3S_3 + H_s(\overline{\psi}_s), \tag{9.49}$$

where $S_i = (S^{(1)}_i, S^{(2)}_i, \{S^{(3)}_{i,k}\}_{k\geq0}, \{S^{(4)}_{i,k}\}_{k\geq0}, \{S^{(5)}_{i,k}\}_{k\geq0}, \{S^{(6)}_{i,k}\}_{k\geq0})$ for $i = 0, 1, 2, 3$.

i	Coefficients $S^{(3)}_{i,k}$ for $k \geq 1$
0	$2k\bar x^{(1)}_{0,k} + \bar L_0\left[\bar x^{(2)}_{0,k\pm1} + (\bar x^{(1)}_0 * \bar x^{(2)}_0)_{k\pm1}\right]$
1	$2k\Delta\bar x^{(1)}_k + \bar L_0\left[\Delta\bar x^{(2)}_{k\pm1} + (\Delta\bar x^{(1)} * \bar x^{(2)}_0)_{k\pm1} + (\bar x^{(1)}_0 * \Delta\bar x^{(2)})_{k\pm1}\right]$ $+\Delta\bar L\left[\bar x^{(2)}_{0,k\pm1} + (\bar x^{(1)}_0 * \bar x^{(2)}_0)_{k\pm1}\right]$
2	$\bar L_0(\Delta\bar x^{(1)} * \Delta\bar x^{(2)})_{k\pm1} + \Delta\bar L\left[\Delta\bar x^{(2)}_{k\pm1} + (\Delta\bar x^{(1)} * \bar x^{(2)}_0)_{k\pm1} + (\bar x^{(1)}_0 * \Delta\bar x^{(2)})_{k\pm1}\right]$
3	$\Delta\bar L(\Delta\bar x^{(1)} * \Delta\bar x^{(2)})_{k\pm1}$
i	Coefficients $S^{(4)}_{i,k}$ for $k \geq 1$
0	$2k\bar x^{(2)}_{0,k} + \bar L_0\bar x^{(3)}_{0,k\pm1}$
1	$\bar L_0\Delta\bar x^{(3)}_{k\pm1} + 2k\Delta\bar x^{(2)}_k + \Delta\bar L\bar x^{(3)}_{0,k\pm1}$
2	$\Delta\bar L\Delta\bar x^{(3)}_{k\pm1}$
3	0
i	Coefficients $S^{(5)}_{i,k}$ for $k \geq 1$
0	$2k\bar x^{(3)}_{0,k} + \bar L_0\bar x^{(4)}_{0,k\pm1}$
1	$\bar L_0\Delta\bar x^{(4)}_{k\pm1} + 2k\Delta\bar x^{(3)}_k + \Delta\bar L\bar x^{(4)}_{0,k\pm1}$
2	$\Delta\bar L\Delta\bar x^{(4)}_{k\pm1}$
3	0
i	Coefficients $S^{(6)}_{i,k}$ for $k \geq 1$
0	$2k\bar x^{(4)}_{0,k} - \bar L_0\left[\bar x^{(1)}_{0,k\pm1} + \beta_0\bar x^{(3)}_{0,k\pm1}\right]$
1	$2k\Delta\bar x^{(4)}_k - \Delta\bar L\left[\bar x^{(1)}_{0,k\pm1} + \beta_0\bar x^{(3)}_{0,k\pm1}\right] - \bar L_0\left[\Delta\bar x^{(1)}_{k\pm1} + \beta_0\Delta\bar x^{(3)}_{k\pm1} + \Delta\beta\bar x^{(3)}_{0,k\pm1}\right]$
2	$-\Delta\bar L\left[\Delta\bar x^{(1)}_{k\pm1} + \beta_0\Delta\bar x^{(3)}_{k\pm1} + \Delta\beta\bar x^{(3)}_{0,k\pm1}\right] - \bar L_0\Delta\beta\Delta\bar x^{(3)}_{k\pm1}$
3	$-\Delta\bar L\Delta\beta\Delta\bar x^{(3)}_{k\pm1}$

Table 9.1 – Coefficients $S^{(j)}_{i,k}$ for the splitting of $F(\beta_s, \bar x_s)$ as a polynomial in s, as given in (9.49). The coefficients $S^{(j)}_{i,k}$ for $k = 0$ and $3 \leq j \leq 6$ are provided in (9.46). The coefficients $S^{(j)}_{i,k}$ for $j = 1, 2$ are provided in (9.44) and (9.45).

Since we evaluate F using a finite dimensional approximation, $F(\beta_s, \bar{x}_s)$ will contain only a finite number of nonzero elements. First, we consider the entries not exceeding the dimension of the finite dimension approximation. This part gets 'hit' by $A^{[m]}$, and is bounded component by component:

$$|A^{[m]}F(\beta_s, \bar{x}_s)^{[m]}| \leq V := |A^{[m]}S_0^{[m]}| + |A^{[m]}S_1^{[m]}| + |A^{[m]}S_2^{[m]}| + |A^{[m]}S_3^{[m]}| + |A^{[m]}|\mu^{[m]},$$

where $V = (V^{(1)}, V^{(2)}, \{V_k^{(3)}\}_{k=0}^{m-1}, \{V_k^{(4)}\}_{k=0}^{m-1}, \{V_k^{(5)}\}_{k=0}^{m-1}, \{V_k^{(6)}\}_{k=0}^{m-1})^T$. Concerning terms that exceed the dimension of the finite dimensional projection, by using the definition of A, one gets for $j = 3, 4, 5, 6$

$$\left|\left(A_{j,j}F^{(j-2)}(\beta_s, \bar{x}_s)\right)_k\right| = \left|\frac{1}{2k} f_k^{(j-2)}(\beta_s, \bar{x}_s)\right| \qquad \text{for } k \geq m.$$

The expansion of $f_k^{(j-2)}(\beta_s, \bar{x}_s)$ in powers of s is given by (9.48) with the coefficients in Table 9.4.5, We note that all $S_{i,k}^{(j)}$ vanish for $k \geq 2m$. To be precise, $S_{i,k}^{(3)}$ vanishes for $k \geq 2m$, whereas when $j = 4, 5, 6$ then $S_{i,k}^{(j)}$ vanishes already for $k \geq m+1$. Hence we define the estimates ($3 \leq j \leq 6$, $m \leq k \leq 2m-1$)

$$\frac{1}{2k}\left|f_k^{(j-2)}(\beta_s, \bar{x}_s)\right| \leq W_k^{(j)} := \frac{1}{2k}\left(|S_{0,k}^{(j)}| + |S_{1,k}^{(j)}| + |S_{2,k}^{(j)}| + |S_{3,k}^{(j)}|\right).$$

Having estimated all the terms appearing in the expression $\|(T(\beta_s, \bar{x}_s) - \bar{x}_s)^{(j)}\|_{X_j}$, we set

$$Y^{(j)} := \begin{cases} V^{(j)}, & j = 1, 2, \\ V_0^{(j)} + 2\displaystyle\sum_{k=1}^{m-1} V_k^{(j)}\nu^k + 2\sum_{k=m}^{2m-1} W_k^{(j)}\nu^k, & j = 3, 4, 5, 6. \end{cases}$$

By construction, we have

$$\left\|(T(\beta_s, \bar{x}_s) - \bar{x}_s)^{(j)}\right\|_{X_j} \leq Y^{(j)} \qquad \text{for all } s \in [0, 1] \text{ and } j = 1, \ldots, 6.$$

9.4.6 The Z bound for the homoclinic orbit problem

We recall the definition of the bounds $Z^{(j)}$ in (9.7). In our context, $Z^{(j)}$ is a bound satisfying

$$\sup_{\substack{b,c \in B_r(0) \\ s \in [0,1]}} \left\|D_x T^{(j)}(\beta, \bar{x}_s + b)c\right\|_{X_j} \leq Z^{(j)}(r).$$

To simplify the manipulations of the expressions appearing in the bounds, we introduce an operator $A^\dagger = \{A_{i,j}^\dagger | 1 \leq i, j \leq 6\}$, where the splitting is explained in Definition 9.4.5. This operator $A^\dagger$ is on the one hand an 'almost inverse' of the operator A, and on the other hand it approximates $D_x F(\beta_0, \bar{x}_0)$. We define $A^\dagger$ piecewise, where we use the decomposition of the Jacobian $(D_x\overline{F}(\beta_0, \bar{x}_0)) = (D_x\overline{F}(\beta_0, \bar{x}_0))_{i,j}$ into 36 blocks as in (9.41):

- $A_{i,j}^\dagger \in \mathbb{R}$ for $1 \leq i, j \leq 2$, where $A_{i,j}^\dagger = (D_x\overline{F}(\beta_0, \bar{x}_0))_{i,j}$;
- $A_{i,j}^\dagger \in (\ell_\nu^1)^*$ for $1 \leq i \leq 2$ and $3 \leq j \leq 6$, where $A_{i,j}^\dagger$ is $(D_x\overline{F}(\beta_0, \bar{x}_0))_{i,j}$ padded with zeros;
- $A_{i,j}^\dagger \in \ell_\nu^1$ for $3 \leq i \leq 6$ and $1 \leq j \leq 2$, where $A_{i,j}^\dagger$ is $(D_x\overline{F}(\beta_0, \bar{x}_0))_{i,j}$ padded with zeros;
- $A_{i,j}^\dagger \in B(\ell_\nu^1, \ell_{\nu'}^1)$ for $3 \leq i, j \leq 6$, with $\nu' < \nu$, where

$$(A_{i,j}^\dagger x^{(j-2)})_k = \begin{cases} \left((D_x\overline{F}(\beta_0, \bar{x}_0))_{i,j}(x^{(j-2)})^{[m]}\right)_k & \text{if } 0 \leq k \leq m-1, \\ \delta_{i,j} 2k x_k^{(j-2)} & \text{if } k \geq m. \end{cases}$$

Now, we use $A^\dagger$ to perform the splitting

$$\begin{aligned} DT(\beta_s, \bar{x}_s + b)c &= [I - ADF(\beta_s, \bar{x}_s + b)]c \\ &= [I - AA^\dagger]c - A[DF(\beta_s, \bar{x}_s + b)c - A^\dagger c]. \end{aligned} \tag{9.50}$$

As in Section 9.3, the bound on the first term in (9.50) can be directly computed. We set $B = I - AA^\dagger$, whose nonzero elements are represented by the finite matrix $I_{4m+2} - A^{[m]} D_x \overline{F}(\beta_0, \bar{x}_0)$, and we use Lemmas 9.4.3 and 9.4.4 to derive the bounds

$$Z_0^{(i)} := \begin{cases} \displaystyle\sum_{j=1}^{2} |B_{i,j}| + \sum_{j=3}^{6} \|B_{i,j}\|_{\infty,\nu^{-1}} & \text{for } i = 1, 2, \\ \displaystyle\sum_{j=1}^{2} \|B_{i,j}\|_{1,\nu} + \sum_{j=3}^{6} \|B_{i,j}\|_{B(\ell^1_\nu)} & \text{for } i = 3, 4, 5, 6, \end{cases} \tag{9.51}$$

with the norms introduced in Section 9.4.2. This provides the desired bound on the first term of (9.50). For the second term, we set $u, v \in B_1(0)$ such that $b = ru$ and $c = rv$. We denote $v = (v_L, v_\psi, v^{(1)}, v^{(2)}, v^{(3)}, v^{(4)})$, and similarly for u, b and c. First, for $i = 1, 2$, we have

$$\begin{aligned} [DF(\beta_s, \bar{x}_s + b)c - A^\dagger c]^{(i)} &= \left| c_0^{(2i)} + 2\sum_{k=1}^{\infty} (-1)^k c_k^{(2i)} - c_0^{(2i)} - 2\sum_{k=1}^{m-1} (-1)^k c_k^{(2i)} \right| \\ &= \left| 2\sum_{k=m}^{\infty} (-1)^k c_k^{(2i)} \right| \le 2\sum_{k=m}^{\infty} |v_k^{(2i)}| r \le \frac{1}{\nu^m} r, \end{aligned} \tag{9.52}$$

where the final inequality follows from Lemma 9.4.3. Next we consider the $k = 0$ term of the other four components. For $i = 1, 2, 3, 4$ one finds

$$\begin{aligned} &\left| [DF(\beta_s, \bar{x}_s + b)c - A^\dagger c]_0^{(i+2)} \right| \\ &\quad = \left| \left[-\frac{dP_{\beta_s}^{(i)}}{d\psi}(\overline{\psi}_s + b_\psi)c_\psi + c_0^{(i)} + 2\sum_{k=1}^{\infty} c_k^{(i)} \right] - \left[-\frac{d\overline{P}_{\beta_0}^{(i)}}{d\psi}(\overline{\psi}_0)c_\psi + c_0^{(i)} + 2\sum_{k=1}^{m-1} c_k^{(i)} \right] \right| \\ &\quad \le \left| \frac{d\overline{P}_{\beta_0}^{(i)}}{d\psi}(\overline{\psi}_s + ru_\psi) - \frac{d\overline{P}_{\beta_0}^{(i)}}{d\psi}(\overline{\psi}_s) + \frac{d\overline{P}_{\beta_0}^{(i)}}{d\psi}(\overline{\psi}_s) - \frac{d\overline{P}_{\beta_0}^{(i)}}{d\psi}(\overline{\psi}_0) + s\frac{d\Delta\overline{P}^{(i)}}{d\psi}(\overline{\psi}_s + b_\psi) \right| r \\ &\qquad + \left(\left| \frac{dh_s}{d\psi}(\overline{\psi}_s + b_\psi) \right| + \left| 2\sum_{k=m}^{\infty} v_k^{(i)} \right| \right) r \\ &\quad \le \left| \frac{d^2\overline{P}_{\beta_0}^{(i)}}{d\psi^2}(\zeta_s) \right| r^2 + \left(\left| \frac{d^2\overline{P}_{\beta_0}^{(i)}}{d\psi^2}(\xi_s)\Delta\bar{\psi} \right| + \left| \frac{d\Delta\overline{P}^{(i)}}{d\psi}(\overline{\psi}_s + b_\psi) \right| \right) sr \qquad (9.53) \\ &\qquad + \left(\rho \left| \frac{\partial h_s}{\partial\theta_1}(\overline{\psi}_s + b_\psi) \right| + \rho \left| \frac{\partial h_s}{\partial\theta_2}(\overline{\psi}_s + b_\psi) \right| + \frac{1}{\nu^m} \right) r, \qquad (9.54) \end{aligned}$$

where ζ_s is in $[\bar{\psi}_s - r, \bar{\psi}_s + r]$, and ξ_s is in $[\bar{\psi}_0, \bar{\psi}_s]$. A direct computation shows that

$$\left| \frac{d^2\overline{P}_{\beta_0}^{(i)}}{d\psi^2}(\psi) \right| = \left| \sum_{|\alpha|<N} -\bar{a}_{0,\alpha}^{(i)} \rho^{|\alpha|} (\alpha_1 - \alpha_2)^2 e^{\mathrm{i}\psi(\alpha_1 - \alpha_2)} \right| \le \sum_{|\alpha|<N} \left| \bar{a}_{0,\alpha}^{(i)} \right| \rho^{|\alpha|} (\alpha_1 - \alpha_2)^2.$$

Combining this with (9.47) gives us a bound on the terms in (9.53):

$$\begin{aligned} \Lambda^{(i)} &:= |\Delta\bar{\psi}| \sum_{|\alpha|<N} \left| \bar{a}_{0,\alpha}^{(i)} \right| \rho^{|\alpha|} (\alpha_1 - \alpha_2)^2 + \sum_{|\alpha|<N} \left| \Delta\bar{a}_{0,\alpha}^{(i)} \right| \rho^{|\alpha|} |\alpha_1 - \alpha_2|, \\ \tilde{\Lambda}^{(i)} &:= \sum_{|\alpha|<N} \left| \bar{a}_{0,\alpha}^{(i)} \right| \rho^{|\alpha|} (\alpha_1 - \alpha_2)^2. \end{aligned}$$

The remaining terms in (9.54) are estimated using Lemma 9.3.8. We obtain, for $i = 1, 2, 3, 4$,

$$\left|[DF(\beta_s, \bar{x}_s + b)c - A^\dagger c]_0^{(i+2)}\right| \leq \mathcal{W}_1^{(i+2)} r + \tilde{\Lambda}^{(i)} r^2, \tag{9.55}$$

with, for $j = 3, 4, 5, 6$,

$$\mathcal{W}_1^{(j)} := \left(\Lambda^{(j-2)} + \frac{8\pi\rho r_m}{\tilde{\nu} \ln \frac{\tilde{\nu}}{\rho}} + \frac{1}{\nu^m}\right). \tag{9.56}$$

For $k \neq 0$, we consider separately the coefficients of r, r^2 and r^3:

$$\Big(DF(\beta_s, \bar{x}_s + ru)rv - A^\dagger rv\Big)_k^{(i)} = \tilde{z}_{1,k}^{(i)} r + \tilde{z}_{2,k}^{(i)} r^2 + \tilde{z}_{3,k}^{(i)} r^3, \qquad \text{for } i = 3, 4, 5, 6.$$

The term $-A^\dagger v$ contributes to the s-independent part of $\tilde{z}_{1,k}^{(i)}$ only. Since in $\tilde{z}_{1,k}^{(i)}$ some of the terms involving $(v^{(j)})^{[m]}$ will cancel, it is useful to introduce $\widehat{v}^{(j)}$ as follows:

$$\widehat{v}_k^{(j)} := \begin{cases} 0 & \text{if } k < m, \\ v_k^{(j)} & \text{if } k \geq m. \end{cases}$$

Using this notation, for $\tilde{z}_{1,k}^{(3)}$ and $1 \leq k \leq m-1$, one finds

$$\begin{aligned}
\tilde{z}_{1,k}^{(3)} = {} & \bar{L}_0 \left[(\bar{x}_0^{(1)} * \widehat{v}^{(2)})_{k\pm1} + (\widehat{v}^{(1)} * \bar{x}_0^{(2)})_{k\pm1} + \delta_{k,m-1}\widehat{v}_{k\pm1}^{(2)}\right] \\
& + s\Big(v_L \left[\Delta\bar{x}_{k\pm1}^{(2)} + (\bar{x}_0^{(1)} * \Delta\bar{x}^{(2)})_{k\pm1} + (\Delta\bar{x}^{(1)} * \bar{x}_0^{(2)})_{k\pm1}\right] \\
& \quad + \bar{L}_0 \left[(\Delta\bar{x}^{(1)} * v^{(2)})_{k\pm1} + (v^{(1)} * \Delta\bar{x}^{(2)})_{k\pm1}\right] \\
& \quad + \Delta\bar{L} \left[v_{k\pm1}^{(2)} + (\bar{x}_0^{(1)} * v^{(2)})_{k\pm1} + (v^{(1)} * \bar{x}_0^{(2)})_{k\pm1}\right] \Big) \\
& + s^2 \Big(\Delta\bar{L}[(\Delta\bar{x}^{(1)} * v^{(2)})_{k\pm1} + (v^{(1)} * \Delta\bar{x}^{(2)})_{k\pm1}] + v_L(\Delta\bar{x}^{(1)} * \Delta\bar{x}^{(2)})_{k\pm1} \Big).
\end{aligned}$$

Clearly $\delta_{k,m-1}\widehat{v}_{k\pm1}^{(2)} = \delta_{k,m-1}v_{k+1}^{(2)} = \widehat{v}_{k\pm1}^{(2)}$, for $k \leq m-1$, and the Kronecker $\delta_{k,m-1}$ may be viewed as superfluous. For $k \geq m$ we find

$$\begin{aligned}
\tilde{z}_{1,k}^{(3)} = {} & \bar{L}_0 \left[(\bar{x}_0^{(1)} * v^{(2)})_{k\pm1} + (v^{(1)} * \bar{x}_0^{(2)})_{k\pm1} + v_{k\pm1}^{(2)}\right] + v_L \left[(\bar{x}_0^{(1)} * \bar{x}_0^{(2)})_{k\pm1} + \delta_{k,m}\bar{x}_{0,k\pm1}^{(2)}\right] \\
& + s\Big(v_L \left[\delta_{k,m}\Delta\bar{x}_{k\pm1}^{(2)} + (\bar{x}_0^{(1)} * \Delta\bar{x}^{(2)})_{k\pm1} + (\Delta\bar{x}^{(1)} * \bar{x}_0^{(2)})_{k\pm1}\right] \\
& \quad + \bar{L}_0 \left[(\Delta\bar{x}^{(1)} * v^{(2)})_{k\pm1} + (v^{(1)} * \Delta\bar{x}^{(2)})_{k\pm1}\right] \\
& \quad + \Delta\bar{L} \left[v_{k\pm1}^{(2)} + (\bar{x}_0^{(1)} * v^{(2)})_{k\pm1} + (v^{(1)} * \bar{x}_0^{(2)})_{k\pm1}\right] \Big) \\
& + s^2\Big(\Delta\bar{L} \left[(\Delta\bar{x}^{(1)} * v^{(2)})_{k\pm1} + (v^{(1)} * \Delta\bar{x}^{(2)})_{k\pm1}\right] + v_L(\Delta\bar{x}^{(1)} * \Delta\bar{x}^{(2)})_{k\pm1} \Big).
\end{aligned}$$

Once again the Kronecker $\delta_{k,m}$ may be viewed as superfluous. For $\tilde{z}_{4,k}^{(1)}$, $\tilde{z}_{5,k}^{(1)}$ and $\tilde{z}_{6,k}^{(1)}$, one finds

$$\tilde{z}_{1,k}^{(4)} = \begin{cases} \delta_{k,m-1}\bar{L}_0 v_{k+1}^{(3)} + s[\Delta\bar{L}v_{k\pm1}^{(3)} + v_L\Delta\bar{x}_{k\pm1}^{(3)}] & \text{for } 1 \leq k \leq m-1 \\ \bar{L}_0 v_{k\pm1}^{(3)} + v_L\delta_{k,m}\bar{x}_{0,k\pm1}^{(3)} + s[\Delta\bar{L}v_{k\pm1}^{(3)} + \delta_{k,m}v_L\Delta\bar{x}_{k\pm1}^{(3)}] & \text{for } k \geq m, \end{cases}$$

$$\tilde{z}_{1,k}^{(5)} = \begin{cases} \delta_{k,m-1}\bar{L}_0 v_{k+1}^{(4)} + s[\Delta\bar{L}v_{k\pm1}^{(4)} + v_L\Delta\bar{x}_{k\pm1}^{(4)}] & \text{for } 1 \leq k \leq m-1 \\ \bar{L}_0 v_{k\pm1}^{(4)} + v_L\delta_{k,m}\bar{x}_{0,k\pm1}^{(4)} + s[\Delta\bar{L}v_{k\pm1}^{(4)} + \delta_{k,m}v_L\Delta\bar{x}_{k\pm1}^{(4)}] & \text{for } k \geq m, \end{cases}$$

and

$$\tilde{z}^{(6)}_{1,k} = \begin{cases} -\delta_{k,m-1}\bar{L}_0[\beta_0 v^{(3)}_{k+1} + v^{(1)}_{k+1}] - s\Big(v_L[\Delta\beta\bar{x}^{(3)}_{0,k\pm1} + \Delta\bar{x}^{(1)}_{k\pm1} + \beta_0\Delta\bar{x}^{(3)}_{k\pm1}] \\ \qquad +\Delta\bar{L}v^{(1)}_{k\pm1} + [\bar{L}_0\Delta\beta + \Delta\bar{L}\beta_0]v^{(3)}_{k\pm1}\Big) - s^2[\Delta\bar{L}\Delta\beta v^{(3)}_{k\pm1} + v_L\Delta\beta\Delta\bar{x}^{(3)}_{k\pm1}] \qquad \text{for } 1 \le k \le m-1 \\ -\bar{L}_0[\beta_0 v^{(3)}_{k\pm1} + v^{(1)}_{k\pm1}] - \delta_{k,m}v_L[\beta_0\bar{x}^{(3)}_{0,k\pm1} + \bar{x}^{(1)}_{0,k\pm1}] \\ \qquad -s\Big(\delta_{k,m}v_L[\Delta\beta\bar{x}^{(3)}_{0,k\pm1} + \Delta\bar{x}^{(1)}_{k\pm1} + \beta_0\Delta\bar{x}^{(3)}_{k\pm1}] + \Delta\bar{L}v^{(1)}_{k\pm1} \\ \qquad\quad +[\bar{L}_0\Delta\beta + \Delta\bar{L}\beta_0]v^{(3)}_{k\pm1}\Big) - s^2[\Delta\bar{L}\Delta\beta v^{(3)}_{k\pm1} + \delta_{k,m}v_L\Delta\beta\Delta\bar{x}^{(3)}_{k\pm1}] \qquad \text{for } k \ge m. \end{cases}$$

The $\tilde{z}^{(2)}_{i,k}$ and $\tilde{z}^{(3)}_{i,k}$ coefficients are still to be determined. For $k \neq 0$, they are given in Table 9.2. Thus, we set $\tilde{z}^{(i)}_1 = \{\tilde{z}^{(i)}_{1,k}\}_{k\ge0}$, $\tilde{z}^{(i)}_2 = \{\tilde{z}^{(i)}_{2,k}\}_{k\ge0}$ and $\tilde{z}^{(i)}_3 = \{\tilde{z}^{(i)}_{3,k}\}_{k\ge0}$. We note that values of $\tilde{z}^{(i)}_{1,0}$

Coefficients in front of r^2, for $k \ge 1$	
$\tilde{z}^{(3)}_{2,k}$	$v_L\big((\bar{x}^{(1)}_0 * u^{(2)})_{k\pm1} + (u^{(1)} * \bar{x}^{(2)}_0)_{k\pm1} + s[(\Delta\bar{x}^{(1)} * u^{(2)})_{k\pm1} + (u^{(1)} * \Delta\bar{x}^{(2)})_{k\pm1}] + u^{(2)}_{k\pm1}\big)$ $+u_L\big((\bar{x}^{(1)}_0 * v^{(2)})_{k\pm1} + (v^{(1)} * \bar{x}^{(2)}_0)_{k\pm1} + s[(\Delta\bar{x}^{(1)} * v^{(2)})_{k\pm1} + (v^{(1)} * \Delta\bar{x}^{(2)})_{k\pm1}] + v^{(2)}_{k\pm1}\big)$ $+s\Delta\bar{L}\big((u^{(1)} * v^{(2)})_{k\pm1} + (v^{(1)} * u^{(2)})_{k\pm1}\big) + \bar{L}_0\big((u^{(1)} * v^{(2)})_{k\pm1} + (v^{(1)} * u^{(2)})_{k\pm1}\big)$
$\tilde{z}^{(4)}_{2,k}$	$u_L v^{(3)}_{k\pm1} + v_L u^{(3)}_{k\pm1}$
$\tilde{z}^{(5)}_{2,k}$	$u_L v^{(4)}_{k\pm1} + v_L u^{(4)}_{k\pm1}$
$\tilde{z}^{(6)}_{2,k}$	$-u_L(\beta_s v^{(3)}_{k\pm1} + v^{(1)}_{k\pm1}) - v_L(\beta_s u^{(3)}_{k\pm1} + u^{(1)}_{k\pm1})$
Coefficients in front of r^3, for $k \ge 1$	
$\tilde{z}^{(3)}_{3,k}$	$u_L(u^{(1)} * v^{(2)})_{k\pm1} + u_L(u^{(2)} * v^{(1)})_{k\pm1} + v_L(u^{(1)} * u^{(2)})_{k\pm1}$
$\tilde{z}^{(4)}_{3,k}$ $\tilde{z}^{(5)}_{3,k}$ $\tilde{z}^{(6)}_{3,k}$	0

Table 9.2 – Coefficients $\tilde{z}^{(i)}_{2,k}$ and $\tilde{z}^{(i)}_{3,k}$ for $k \neq 0$.

and $\tilde{z}^{(i)}_{2,0}$ are not explicitly given, but (9.55) provides bounds on these terms. We are going to abuse notation by referring to these bounds as $\tilde{z}^{(i)}_{1,0}$ and $\tilde{z}^{(i)}_{2,0}$, where we will correct for this abuse below whenever these terms get involved. We set $\tilde{z}^{(i)}_{3,0} = 0$.

For $l = 1, 2$, one can estimate, using Equation (9.52) and the definition of $\tilde{z}^{(i)}_j$,

$$\Big|\Big(A[DF(\beta_s, \bar{x}_s + b)c - A^\dagger c]\Big)^{(l)}\Big| \le \left(\sum_{i=1}^{2}\frac{|A_{l,i}|}{\nu^m} + \sum_{i=3}^{6}\Big|A_{l,i}\tilde{z}^{(i)}_1\Big|\right) r + \sum_{i=3}^{6}\|A_{l,i}\|_{\infty,\nu^{-1}}\left(\|\tilde{z}^{(i)}_2\|_{1,\nu}r^2 + \|\tilde{z}^{(i)}_3\|_{1,\nu}r^3\right), \tag{9.57}$$

and for $l = 3, 4, 5, 6$

$$\Big\|\Big(A[DF(\beta_s, \bar{x}_s + b)c - A^\dagger c]\Big)^{(l)}\Big\|_{1,\nu} \le \left(\sum_{i=1}^{2}\frac{\|A_{l,i}\|_{1,\nu}}{\nu^m} + \sum_{i=3}^{6}\Big\|A_{l,i}\tilde{z}^{(i)}_1\Big\|_{1,\nu}\right) r + \sum_{i=3}^{6}\|A_{l,i}\|_{B(\ell^1_\nu)}\left(\|\tilde{z}^{(i)}_2\|_{1,\nu}r^2 + \|\tilde{z}^{(i)}_3\|_{1,\nu}r^3\right). \tag{9.58}$$

Apart from $|A_{l,i}|$ for $l, i = 1, 2$, which are scalars, it is not immediately obvious how to compute or estimate the terms in (9.57) and (9.58) explicitly. The norms $\|A_{l,i}\|_{\infty,\nu^{-1}}$ for $l = 1, 2$, $i = 3, 4, 5, 6$ and $\|A_{l,i}\|_{1,\nu}$ for $i = 1, 2$, $l = 3, 4, 5, 6$ can be computed directly, since they are represented by row and column vectors of length m. The operator norms $\|A_{l,i}\|_{B(\ell^1_\nu)}$ can be computed using Lemma 9.4.4, since for $l \neq i$ they are represented by finite matrices, whereas for $l = i$ they have a decaying diagonal tail (see the analogous Lemma 9.3.5).

The norms $\|\tilde{z}_2^{(i)}\|_{1,\nu}$ and $\|\tilde{z}_3^{(i)}\|_{1,\nu}$ in the quadratic and cubic terms in r can be estimated using the Banach algebra structure. Taking into account the bound on $\tilde{z}_{2,0}^{(i)}$ in (9.55), this leads to bounds

$$\|\tilde{z}_2^{(i)}\|_{1,\nu} \leq \mathcal{W}_2^{(i)} \qquad \text{for } i = 3, 4, 5, 6,$$

with

$$\mathcal{W}_2^{(3)} := \tilde{\Lambda}^{(1)} + 2\left(\nu + \frac{1}{\nu}\right)\left(\|\bar{x}_0^{(1)}\|_{1,\nu} + \|\bar{x}_0^{(2)}\|_{1,\nu} + \|\Delta\bar{x}^{(1)}\|_{1,\nu} + \|\Delta\bar{x}^{(2)}\|_{1,\nu} + 1 + \bar{L}_0 + |\Delta\bar{L}|\right), \tag{9.59}$$

$$\mathcal{W}_2^{(4)} := \tilde{\Lambda}^{(2)} + 2\left(\nu + \frac{1}{\nu}\right), \tag{9.60}$$

$$\mathcal{W}_2^{(5)} := \tilde{\Lambda}^{(3)} + 2\left(\nu + \frac{1}{\nu}\right), \tag{9.61}$$

$$\mathcal{W}_2^{(6)} := \tilde{\Lambda}^{(4)} + 2\left(\nu + \frac{1}{\nu}\right)(\beta_1 + 1), \tag{9.62}$$

and

$$\|\tilde{z}_3^{(3)}\|_{1,\nu} \leq \mathcal{W}_3^{(3)} := 3\left(\nu + \frac{1}{\nu}\right). \tag{9.63}$$

The factor $\nu + \nu^{-1}$ in the expressions above is due to the shift in index (to the right and to the left) in $u_{k\pm1}^{(i)}$, $v_{k\pm1}^{(i)}$, etc.

This leaves us with estimating $|A_{l,i}\tilde{z}_1^{(i)}|$ and $\|A_{l,i}\tilde{z}_1^{(i)}\|_{1,\nu}$. Since these appear in the terms that are linear in r, a direct triangle inequality bound would be too rough for the method to succeed. Hence we estimate these terms more carefully below.

For the term in front of r in equation (9.57), for $l = 1, 2$, we have

$$\sum_{i=3}^{6}\left|A_{l,i}\tilde{z}_i^{(1)}\right| \leq \sum_{i=3}^{6}|(A_{l,i})_0|\,\mathcal{W}_1^{(i)} + \sum_{i=3}^{6}\left|\sum_{k=1}^{m-1}(A_{l,i})_k\tilde{z}_{1,k}^{(i)}\right|.$$

Here we have corrected for our abuse of notation regarding $\tilde{z}_{1,0}^{(i)}$ by splitting it off using the triangle inequality.

Remark 9.4.7. *We use the bound* (9.31) *to estimate the convolution*

$$\sup_{\|v\|_{1,\nu}\leq1}|(a*v)_k| = \sup_{\|v\|_{1,\nu}\leq1}\left|\sum_{k'\in\mathbb{Z}}v_{|k'|}a_{|k-k'|}\right| \leq \max\left\{|a_k|, \sup_{k'\geq1}\frac{|a_{|k-k'|}| + |a_{|k+k'|}|}{2\nu^{k'}}\right\} := \mathcal{Q}_k(a).$$

A similar estimate leads to

$$\sup_{\|v\|_{1,\nu}\leq1}|(a*\widehat{v})_k| \leq \sup_{k'\geq m}\frac{|a_{|k-k'|}| + |a_{|k+k'|}|}{2\nu^{k'}} := \hat{\mathcal{Q}}_k(a).$$

Some of the terms in $\tilde{z}_{1,k}^{(i)}$ are computable directly, while others need to be estimated. To present these estimates in a structured way we introduce several computable constants. For the

convolution terms involving either v or $\widehat{v}$ in $\tilde{z}^{(3)}_{1,k}$ we introduce (for $k \geq 1$)

$$\begin{aligned}
\omega^{(i)}_k &:= \mathcal{Q}_{k-1}(\bar{x}^{(i)}) + \mathcal{Q}_{k+1}(\bar{x}^{(i)}),\\
\hat{\omega}^{(i)}_k &:= \hat{\mathcal{Q}}_{k-1}(\bar{x}^{(i)}) + \hat{Q}_{k+1}(\bar{x}^{(i)}),\\
\Delta\omega^{(i)}_k &:= \mathcal{Q}_{k-1}(\Delta\bar{x}^{(i)}) + \mathcal{Q}_{k+1}(\Delta\bar{x}^{(i)}).
\end{aligned}$$

Here $\mathcal{Q}_k(\cdot)$ and $\hat{\mathcal{Q}}_k(\cdot)$, defined in Remark 9.4.7, can be computed (at least finitely many of them) since $\bar{x}^{(i)}$ and $\Delta\bar{x}^{(i)}$ have only finitely many nonzero components. We now set, for $k = 1, \dots, m-1$,

$$\begin{aligned}
z^{(3)}_k &:= |\Delta\bar{L}|\Big[\frac{2}{\nu^{k-1}} + \omega^{(1)}_k + \omega^{(2)}_k\Big] + (\bar{L}_0 + |\Delta\bar{L}|)\big[\Delta\omega^{(1)}_k + \Delta\omega^{(2)}_k\big] + \bar{L}_0\big[\hat{\omega}^{(1)}_k + \hat{\omega}^{(2)}_k\big],\\
z^{(4)}_k &:= 2\frac{|\Delta\bar{L}|}{\nu^{k-1}},\\
z^{(5)}_k &:= 2\frac{|\Delta\bar{L}|}{\nu^{k-1}},\\
z^{(6)}_k &:= 2\frac{|\bar{L}_0\Delta\beta + \Delta\bar{L}\beta_0| + |\Delta\bar{L}|}{\nu^{k-1}} + 2\frac{\big|\Delta\bar{L}\Delta\beta\big|}{\nu^{k-1}},
\end{aligned}$$

as well as

$$\begin{aligned}
\hat{z}^{(3)}_k &:= \Delta\bar{x}^{(2)}_{k\pm1} + (\bar{x}^{(1)}_0 * \Delta\bar{x}^{(2)})_{k\pm1} + (\Delta\bar{x}^{(1)} * \bar{x}^{(2)}_0)_{k\pm1},\\
\hat{z}^{(4)}_k &:= \Delta\bar{x}^{(3)}_{k\pm1},\\
\hat{z}^{(5)}_k &:= \Delta\bar{x}^{(4)}_{k\pm1},\\
\hat{z}^{(6)}_k &:= \Delta\beta\bar{x}^{(3)}_{0,k\pm1} + \Delta\bar{x}^{(1)}_{k\pm1} + \beta_0\Delta\bar{x}^{(3)}_{k\pm1},
\end{aligned}$$

and

$$\begin{aligned}
\hat{\hat{z}}^{(3)}_k &:= (\Delta\bar{x}^{(1)} * \Delta\bar{x}^{(2)})_{k\pm1},\\
\hat{\hat{z}}^{(6)}_k &:= \Delta\beta\Delta\bar{x}^{(3)}_{k\pm1}.
\end{aligned}$$

Recall that $|A|$ denotes the component-wise absolute value. Then we have the computable estimates ($l = 1, 2$)

$$\begin{aligned}
\Big|\sum_{k=1}^{m-1}(A_{l,3})_k\tilde{z}^{(3)}_{1,k}\Big| &\leq \mathcal{Z}_{l,3} := \frac{(|A|_{l,3})_{m-1}\bar{L}_0}{\nu^m} + \sum_{k=1}^{m-1}(|A|_{l,3})_k z^{(3)}_k + \Big|\sum_{k=1}^{m-1}(A_{l,3})_k\hat{z}^{(3)}_k\Big| + \Big|\sum_{k=1}^{m-1}(A_{l,3})_k\hat{\hat{z}}^{(3)}_k\Big|,\\
\Big|\sum_{k=1}^{m-1}(A_{l,4})_k\tilde{z}^{(4)}_{1,k}\Big| &\leq \mathcal{Z}_{l,4} := \frac{(|A|_{l,4})_{m-1}\bar{L}_0}{\nu^m} + \sum_{k=1}^{m-1}(|A|_{l,4})_k z^{(4)}_k + \Big|\sum_{k=1}^{m-1}(A_{l,4})_k\hat{z}^{(4)}_k\Big|,\\
\Big|\sum_{k=1}^{m-1}(A_{l,5})_k\tilde{z}^{(5)}_{1,k}\Big| &\leq \mathcal{Z}_{l,5} := \frac{(|A|_{l,5})_{m-1}\bar{L}_0}{\nu^m} + \sum_{k=1}^{m-1}(|A|_{l,5})_k z^{(5)}_k + \Big|\sum_{k=1}^{m-1}(A_{l,5})_k\hat{z}^{(5)}_k\Big|,\\
\Big|\sum_{k=1}^{m-1}(A_{l,6})_k\tilde{z}^{(6)}_{1,k}\Big| &\leq \mathcal{Z}_{l,6} := \frac{(|A|_{l,6})_{m-1}\bar{L}_0(\beta_0+1)}{\nu^m} + \sum_{k=1}^{m-1}(|A|_{l,6})_k z^{(6)}_k + \Big|\sum_{k=1}^{m-1}(A_{l,6})_k\hat{z}^{(6)}_k\Big| + \Big|\sum_{k=1}^{m-1}(A_{l,6})_k\hat{\hat{z}}^{(6)}_k\Big|.
\end{aligned}$$

For $l = 3, 4, 5, 6$, we split the estimate in three terms because of the way the $\tilde{z}^{(1)}_{i,0}$ bounds and A are defined. Using (9.55), we get ($i, l = 3, 4, 5, 6$)

$$\Big\|A_{l,i}\tilde{z}^{(i)}_1\Big\|_{1,\nu} \leq \mathcal{W}^{(i)}_1\sum_{j=0}^{m-1}|(A_{l,i})_{j0}| + 2\sum_{j=1}^{m-1}\Big|\sum_{k=1}^{m-1}(A_{l,i})_{j,k}\tilde{z}^{(i)}_{1,k}\Big|\nu^j + 2\delta_{l,i}\sum_{j\geq m}\frac{1}{2j}|\tilde{z}^{(i)}_{1,j}|\nu^j. \tag{9.64}$$

Again we have dealt with the $\tilde{z}^{(i)}_{1,0}$ terms separately to take into account our abuse of notation.

The final two terms in (9.64) still need to be estimated. The first of these can be estimated in the same way as above, which we write (for $3 \le l, i \le 6$) compactly as

$$\sum_{j=1}^{m-1} \left| \sum_{k=1}^{m-1} (A_{l,i})_{j,k} \tilde{z}^{(3)}_{1,k} \right| \nu^j \le \mathcal{Z}_{l,i} := \sum_{j=1}^{m-1} (\mathcal{Z}_{l,i})_j,$$

with

$$(\mathcal{Z}_{l,i})_j := \frac{(|A|_{l,i})_{j,m-1} \bar{L}_0 (\delta_{i,6}\beta_0 + 1)}{\nu^m} + \sum_{k=1}^{m-1} (|A|_{l,i})_{j,k} z^{(i)}_k + \left| \sum_{k=1}^{m-1} (A_{l,i})_{j,k} \hat{z}^{(i)}_k \right| + \left| \sum_{k=1}^{m-1} (A_{l,i})_{j,k} \hat{\hat{z}}^{(i)}_k \right|,$$

where one should read $\hat{\hat{z}}^{(4)}_k = \hat{\hat{z}}^{(5)}_k = 0$.

For the final 'tail' terms in (9.64), we bound these as we did for $z^{(i)}_2$ and $z^{(i)}_3$ coefficients. We obtain

$$\begin{aligned}
\sum_{j \ge m} \frac{|\tilde{z}^{(3)}_{1,j}|}{j} \nu^j \le \mathcal{Z}^\infty_3 := {} & \frac{1}{2m} \left(\nu + \frac{1}{\nu} \right) \left(\bar{L}_0 + |\Delta \bar{L}| \right) \left(\|\bar{x}^{(1)}_0\|_{1,\nu} + \|\bar{x}^{(2)}_0\|_{1,\nu} + \|\Delta \bar{x}^{(1)}\|_{1,\nu} + \|\Delta \bar{x}^{(2)}\|_{1,\nu} + 1 \right) \\
& + \sum_{k=m}^{2m-1} \frac{\nu^k}{k} \left(\left| (\bar{x}^{(1)}_0 * \bar{x}^{(2)}_0)_{k\pm 1} \right| + \left| (\bar{x}^{(1)}_0 * \Delta \bar{x}^{(2)})_{k\pm 1} \right| + \left| (\Delta \bar{x}^{(1)} * \bar{x}^{(2)}_0)_{k\pm 1} \right| \right) \\
& + \sum_{k=m}^{2m-1} \frac{\nu^k}{k} \left| (\Delta \bar{x}^{(1)} * \Delta \bar{x}^{(2)})_{k\pm 1} \right| + \frac{\nu^m}{m} \left(|\Delta \bar{x}^{(2)}_{m-1}| + |\bar{x}^{(2)}_{0,m-1}| \right) \\
\sum_{j \ge m} \frac{|\tilde{z}^{(4)}_{1,j}|}{j} \nu^j \le \mathcal{Z}^\infty_4 := {} & \frac{1}{2m} \left(\nu + \frac{1}{\nu} \right) \left(\bar{L}_0 + |\Delta \bar{L}| \right) + \frac{\nu^m}{m} \left(|\bar{x}^{(3)}_{0,m-1}| + |\Delta \bar{x}^{(3)}_{m-1}| \right), \\
\sum_{j \ge m} \frac{|\tilde{z}^{(5)}_{1,j}|}{j} \nu^j \le \mathcal{Z}^\infty_5 := {} & \frac{1}{2m} \left(\nu + \frac{1}{\nu} \right) \left(\bar{L}_0 + |\Delta \bar{L}| \right) + \frac{\nu^m}{m} \left(|\bar{x}^{(4)}_{0,m-1}| + |\Delta \bar{x}^{(4)}_{m-1}| \right), \\
\sum_{j \ge m} \frac{|\tilde{z}^{(6)}_{1,j}|}{j} \nu^j \le \mathcal{Z}^\infty_6 := {} & \frac{1}{2m} \left(\nu + \frac{1}{\nu} \right) \left(\bar{L}_0 + |\Delta \bar{L}| \right) (1 + \beta_1) \\
& + \frac{\nu^m}{m} \left(\beta_1 \left(|\bar{x}^{(3)}_{0,m-1}| + |\Delta \bar{x}^{(3)}_{m-1}| \right) + |\bar{x}^{(1)}_{0,m-1}| + |\Delta \bar{x}^{(1)}_{m-1}| \right).
\end{aligned}$$

Therefore, recalling (9.56) and (9.59)–(9.63), for $l = 1, 2$, we set

$$\begin{aligned}
Z^{(l)}_1 &:= \sum_{i=1}^{2} \frac{|A_{l,i}|}{\nu^m} + \sum_{i=3}^{6} |(A_{l,i})_0| \, \mathcal{W}^{(i)}_1 + \sum_{i=3}^{6} \mathcal{Z}_{l,i}, \\
Z^{(l)}_2 &:= \sum_{i=3}^{6} \|A_{l,i}\|_{\infty,\nu^{-1}} \mathcal{W}^{(i)}_2, \\
Z^{(l)}_3 &:= \|A_{l,3}\|_{\infty,\nu^{-1}} \mathcal{W}^{(3)}_3,
\end{aligned}$$

and for $l = 3, 4, 5, 6$, we set

$$\begin{aligned}
Z^{(l)}_1 &:= \sum_{i=1}^{2} \frac{\|A_{l,i}\|_{1,\nu}}{\nu^m} + \sum_{i=3}^{6} \mathcal{W}^{(i)}_1 \sum_{j=1}^{m-1} |(A_{l,i})_{j,0}| + \sum_{i=3}^{6} \mathcal{Z}_{l,i} + \mathcal{Z}^\infty_l, \\
Z^{(l)}_2 &:= \sum_{i=3}^{6} \|A_{l,i}\|_{B(\ell^1_\nu)} \mathcal{W}^{(i)}_2, \\
Z^{(l)}_3 &:= \|A_{l,3}\|_{B(\ell^1_\nu)} \, \mathcal{W}^{(3)}_3.
\end{aligned}$$

Finally, by construction, for all $s \in [0,1]$ and $l = 1, \dots, 6$, we have

$$\sup_{b,c \in B(r)} \left\| (DT(\beta_s, \bar{x}_s + b)c)^{(l)} \right\|_{X^{(l)}} \leq \left(Z_0^{(l)} + Z_1^{(l)} \right) r + Z_2^{(l)} r^2 + Z_3^{(l)} r^3.$$

9.4.7 Use of the uniform contraction principle

Using the computable bounds $Y^{(l)}$ and $Z^{(l)}$ constructed in the previous two sections, we set

$$p^{(l)}(r) := Y^{(l)} + \left(Z_0^{(l)} + Z_1^{(l)} - 1 \right) r + Z_2^{(l)} r^2 + Z_3^{(l)} r^3, \qquad l = 1, \dots, 6. \tag{9.65}$$

If we find an $r > 0$ such that $p^{(l)}(r) < 0$ for all $l = 1, \dots, 6$, then according to Proposition 9.2.5 we have validated the numerical approximation $\bar{x}_s$ of solutions to the BVP (9.29), for every $s \in [0,1]$, and hence we have proven the existence of symmetric homoclinic orbits for all $\beta \in [\beta_0, \beta_1]$.

Proposition 9.4.8. *For every $s \in [0,1]$, let*

$$\overline{v}_s^{(i)}(t) = \bar{x}_{s,0}^{(i)} + 2 \sum_{k=0}^{m-1} \bar{x}_{s,k}^{(i)} T_k(t), \qquad \text{for } i = 1, 2, 3, 4,$$

be the approximate solution of (9.29) that we have computed for $\beta = \beta_s$, $L = \bar{L}_s$ and $\psi = \bar{\psi}_s$. Assume that there exists an $r > 0$ such that $p^{(l)}(r) < 0$ for all $l = 1, \dots, 6$. Then, for each $s \in [0,1]$, there exists a solution of (9.29) for $\beta = \beta_s$ of the form

$$v_s^{(i)}(t) = x_{s,0}^{(i)} + 2 \sum_{k=0}^{\infty} x_{s,k}^{(i)} T_k(t), \qquad \text{for } i = 1, 2, 3, 4,$$

and some $L = L_s$ and $\psi = \psi_s$ satisfying $|L_s - \bar{L}_s| \leq r$ and $|\psi_s - \bar{\psi}_s| \leq r$. This solution corresponds to a (symmetric) homoclinic orbit of (9.3). Furthermore, let

$$g_s^{(i)}(t) = v_s^{(i)}(t) - \overline{v}_s^{(i)}(t) \qquad \text{for } i = 1, 2, 3, 4,$$

then we have the following uniform error bound on the (central part of) the homoclinic orbit in phase space: $|g_s^{(i)}(t)| \leq r$ for all $t \in [-1,1]$, $s \in [0,1]$ and $i = 1, 2, 3, 4$.

Proof. Proposition 9.2.5 yields that, for each $s \in [0,1]$, there exists a unique fixed point x_s of $T(\beta_s, \cdot)$ in the ball of radius r around $\bar{x}_s$. The operator A is injective since its non-diagonal part $A^{[m]}$ is invertible. The latter follows from the fact that, see (9.51),

$$\left\| I_{4m+2} - A^{[m]} D_x \overline{F}(\beta_0, \bar{x}_0) \right\|_{B(X^{[m]})} \leq \max_{1 \leq l \leq 6} Z_0^{(l)} < 1,$$

where the final inequality is implied by $p^{(l)}(r) < 0$. Here the operator norm on $X^{[m]} \cong \mathbb{R}^{4m+2}$ is induced by the one on X. Hence the fixed point x_s of T solves $F(\beta_s, x_s) = 0$, and by construction v_s is a solution of (9.29), which through the change of variables from Section 9.4.1 corresponds to a homoclinic solution of (9.3). The error bound follows from

$$\begin{aligned}
\|v_s^{(i)}(t) - \overline{v}_s^{(i)}(t)\|_\infty &= \left\| x_{s,0}^{(i)} - \bar{x}_{s,0}^{(i)} + 2 \sum_{k \geq 1} (x_{s,k}^{(i)} - \bar{x}_{s,k}^{(i)}) T_k(t) \right\|_\infty \\
&\leq |x_{s,0}^{(i)} - \bar{x}_{s,0}^{(i)}| + 2 \sum_{k \geq 1} |x_{s,k}^{(i)} - \bar{x}_{s,k}^{(i)}| \\
&\leq |x_{s,0}^{(i)} - \bar{x}_{s,0}^{(i)}| + 2 \sum_{k \geq 1} |x_{s,k}^{(i)} - \bar{x}_{s,k}^{(i)}| \nu^k \\
&= \|x_s^{(i)} - \bar{x}_s^{(i)}\|_{1,\nu} \leq \|x_s - \bar{x}_s\|_X \leq r.
\end{aligned}$$

□

9.5 Algorithm and results

In this section we discuss some algorithmic issues. In particular, we explain how certain computational constants are chosen and how the two parts of the problem (the manifold computation and the boundary value problem) are joined together to produce the homoclinic orbit. To get the continuation started, the first thing to do is to compute the approximation of the manifold for a fixed value of β. Since the first coefficients of the parameterization depend on the steady state and the eigenvectors, which are known, one can start Newton's method with these values in combination with zeros for all higher order coefficients. If Newton's method does not converge, replacing the starting point with a good approximation for a slightly higher number of Taylor coefficients (which can be computed recursively) will work. Once one good approximation has been found for a particular value of the parameter β, one can use it as the starting point to find another approximation for sufficiently close values of the parameter.

Another important point for the computations is the size of the manifold that we get. Since the stable eigenvalues of the Jacobian at 0 are complex conjugates, we know that asymptotically the orbit spirals toward the origin. If the manifold we compute is large enough to contain most of the spiraling part, then we do not have to compute that part of the orbit using Chebyshev series, which is advantageous. Generally speaking, the larger the manifold is, the easier the remaining part with Chebyshev will be. Therefore we use the method developed in [77] to maximize the image of the parameterization we compute.

A natural approach to obtain a larger manifold is to try and maximize the $\tilde{\nu}$ for which we can validate the parameterization (we recall that its domain of definition is $D_{2,\tilde{\nu}}(\mathbb{R}^2) = \{\theta \in \mathbb{R}^2,\ |\theta|_2 \leq \tilde{\nu}\}$). However, taking $\tilde{\nu} \gg 1$ or $\tilde{\nu} \ll 1$ leads to numerical instabilities (see for instance the quantities $K^{(i,j)}$ defined in Section 9.3.4). The key observation from [77] to avoid this phenomenon is the following. Given a parameterization

$$P(\theta) = \sum_{|\alpha| \geq 0} a_\alpha \theta^\alpha,$$

and, for some $\gamma > 0$, a *rescaled* parameterization (also with rescaled eigenvectors)

$$\tilde{P}(\theta) = \sum_{|\alpha| \geq 0} \tilde{a}_\alpha \theta^\alpha, \quad \text{with} \quad \tilde{a}_\alpha = \gamma^{|\alpha|} a_\alpha,$$

the parameterization P on the domain $D_{2,\gamma}(\mathbb{R}^2)$ defines the same manifold as the rescaled parameterization $\tilde{P}$ on the domain $D_{2,1}(\mathbb{R}^2)$. Therefore we can fix $\tilde{\nu}$ to be 1 and instead look for the largest γ for which we can validate the rescaled parameterization.

Another useful feature of the results of [77] is that they provide the explicit dependence of the bounds Y and Z with respect to the rescaling γ, enabling us to recompute bounds for any rescaling cheaply. In practice, we use the following process:

- Compute an approximate parameterization P (that is, the coefficient a_α).
- Compute the bounds Y and Z for β_0, without the continuation (i.e. take $\Delta a = 0$ and $\Delta\beta = 0$ in every estimate).
- Find the largest γ for which the proof succeeds (i.e. we find an $r > 0$ such that $p^{(i)}(r) < 0$ for all $i = 1, 2, 3, 4$, where the four radii polynomials $p^{(i)}$ are defined in (9.19)) for the rescaled coefficients $\tilde{a}_\alpha = \gamma^{|\alpha|} a_\alpha$, while requiring the coefficients of the linear term (the one front of r) in each radii polynomial $p^{(i)}$ to be less than some threshold $\eta \in (0, 1)$, which will be discussed below. This step yields a parameterization $\tilde{P}$ with rigorous error bounds on the domain $D_{2,1}(\mathbb{R}^2)$.
- Use the parameterization $\tilde{P}$ with this γ for the Chebyshev part and for the continuation.

Before describing in more detail the process of continuation, let us explain the role of the threshold η. Finding a positive root of a radii polynomial is impossible if its linear term is not

negative, because all its other coefficients are always non-negative by construction. If the linear term is just negative enough for the proof to work at the single parameter value β_0, then $\Delta\beta$ has to be taken extremely small for it to remain negative for the uniform proof, since all bounds become worse monotonically in $|\Delta\beta|$. However, we want to take $\Delta\beta$ as large as possible to reduce the number of steps we have to perform to prove the existence of a symmetric homoclinic orbit for all $\beta \in [0.5, 1.9]$. Hence, the addition of this threshold η is a trade off: we get a manifold that is a bit smaller than what we could have had optimally, which makes the proof for the Chebyshev part a bit harder, but we can take larger steps in β, making the total process faster overall. In practice, we use an η close to 0.5 (the value we use varies slightly with β).

Once the approximation for the manifold is maximized and proven for a particular value of β_0, one can use it as the starting point to find the approximation for $\beta_1 > \beta_0$ in order to compute an approximation for the whole interval $[\beta_0, \beta_1]$. We use the same rescaling γ for the entire interval $[\beta_0, \beta_1]$. On the other hand, it is possible to use different scalings for consecutive intervals.

The value of $\Delta\beta = \beta_1 - \beta_0$ that we use is not constant, and varies between 2.5×10^{-4} and 3.9×10^{-6}. The smaller values are needed when $\beta_0 \geq 1.8$. This is due to the fact that proof of the stable manifold becomes harder and harder when β approaches 2. Indeed, when β goes to 2 the real part of the stable eigenvalues (see (9.10)) goes to zero, and the problem of finding the stable manifold becomes singular (this can also be seen in the bounds derived in Section 9.3). Note that when the proof fails for a given interval, a smaller $\Delta\beta$ needs to be used. Thus, the algorithm needs to recompute both the manifold and the orbit for $\beta = \beta_1$. However, $A^\dagger$ and A need not to be computed again for the new proofs since they both only depend on the approximation at $\beta = \beta_0$.

For the manifold all proofs were done using $N = 30$ for the dimension of the truncated power series. For the orbit, the proof succeeds with $m = 350$ for $[\beta_0, \beta_1] \subset [0.5, 1.8]$, and with $m = 400$ otherwise. In Figure 9.1 one can see the profile of the solution for $\beta = 0.5$, $\beta = 1.2$ and $\beta = 1.9$. The left part of the figure shows the decay rate of the solution using the logarithm of the absolute value of the first 50 Chebyshev coefficients. Recall that the first component of the system is given by $v_1 = e^{u_1} - 1$, where u_1 is the first component of the original system, obtained after transforming the fourth order equation to a first order system. One can see that the solution for $\beta = 0.5$ is really close to -1 for a much longer period of time than the other solutions depicted. This behaviour has an impact on the decay of the corresponding Chebyshev series. Moreover, another value affecting the decay rate of the solution is the time rescaling factor L of the orbit. For $\beta = 0.5$ (respectively $\beta = 1.2$ and $\beta = 1.9$) we have $L \approx 3.1312$ (respectively $L \approx 1.7671$ and $L \approx 2.6170$). The first three components of the solution and the local manifold can be seen in Figure 9.2, Figure 9.3 and Figure 9.4 for $\beta = 0.5$, $\beta = 1.2$ and $\beta = 1.9$, respectively. The profiles of the first component $v^{(1)}$ of these three solutions can be compared in Figure 9.5, where half the symmetric homoclinic orbits is depicted. Furthermore, the three corresponding homoclinic solutions of the suspension bridge equation (9.2) in the original u-variable are presented in Figure 9.6.

Finally, to perform the proof successfully for the entire interval range $\beta \in [0.5, 1.9]$ we had to execute the algorithm 7960 times. Each proof took between 7 and 10 seconds on a laptop with an Intel Core i7 4500U processor on MATLAB R2016a. The code which was used to perform the proofs is available at [59] and uses the interval arithmetic package INTLAB [190].

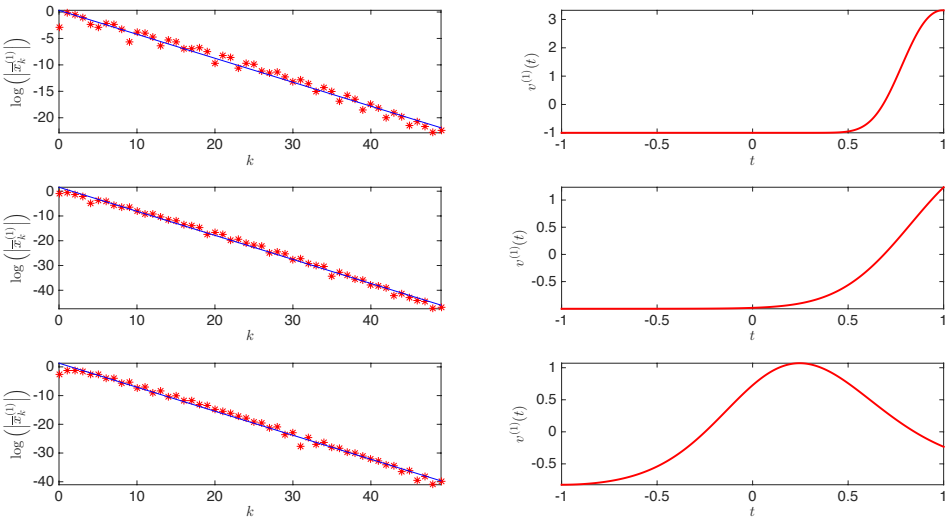


Figure 9.1 – The logarithm of the absolute value of the 50 first coefficients of the first component on the left, and the profile of the first component of the solution on the right. At the top $\beta = 0.5$, in the middle $\beta = 1.2$ and at the bottom $\beta = 1.9$

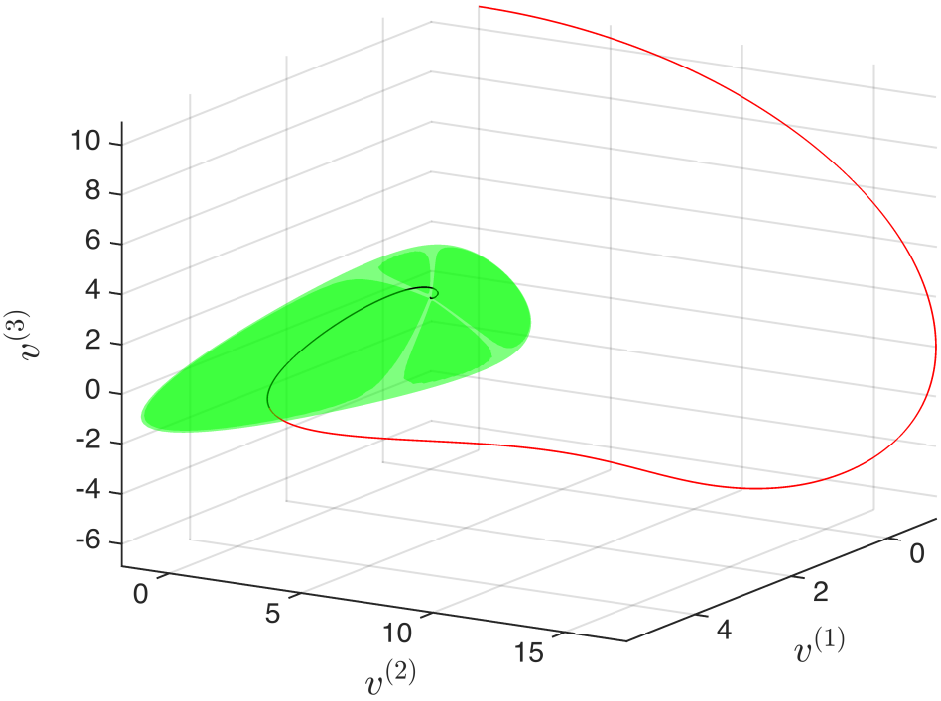


Figure 9.2 – First three components of the solution (red) and the manifold (green) in the case $\beta = 0.5$. The segment in black corresponds to the forward orbit of the solution on the local manifold, where the dynamics is obtained via the conjugacy relation satisfied by the parameterization (e.g. see [84]).

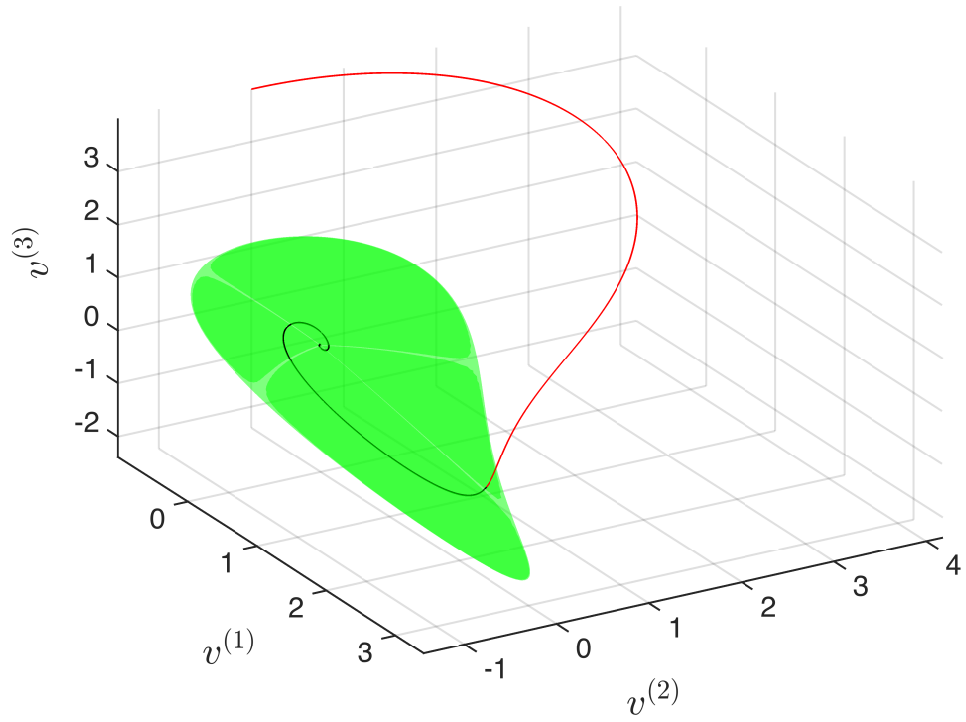


Figure 9.3 – First three components of the solution (red) and the manifold (green) in the case $\beta = 1.2$. The segment in black corresponds to the forward orbit of the solution on the local manifold, where the dynamics is obtained via the conjugacy relation satisfied by the parameterization.

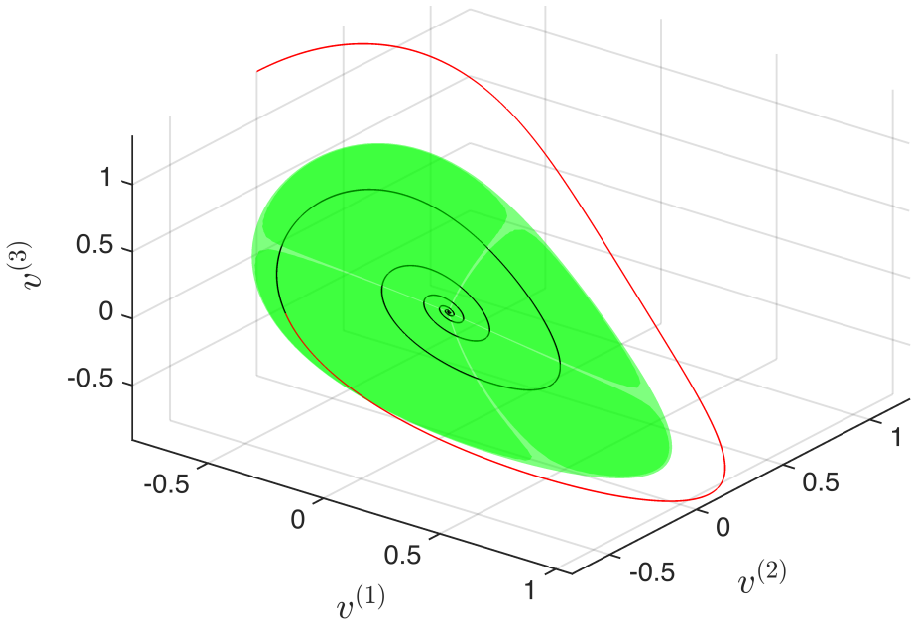


Figure 9.4 – First three components of the solution (red) and the manifold (green) in the case $\beta = 1.9$. The segment in black corresponds to the forward orbit of the solution on the local manifold, where the dynamics is obtained via the conjugacy relation satisfied by the parameterization.

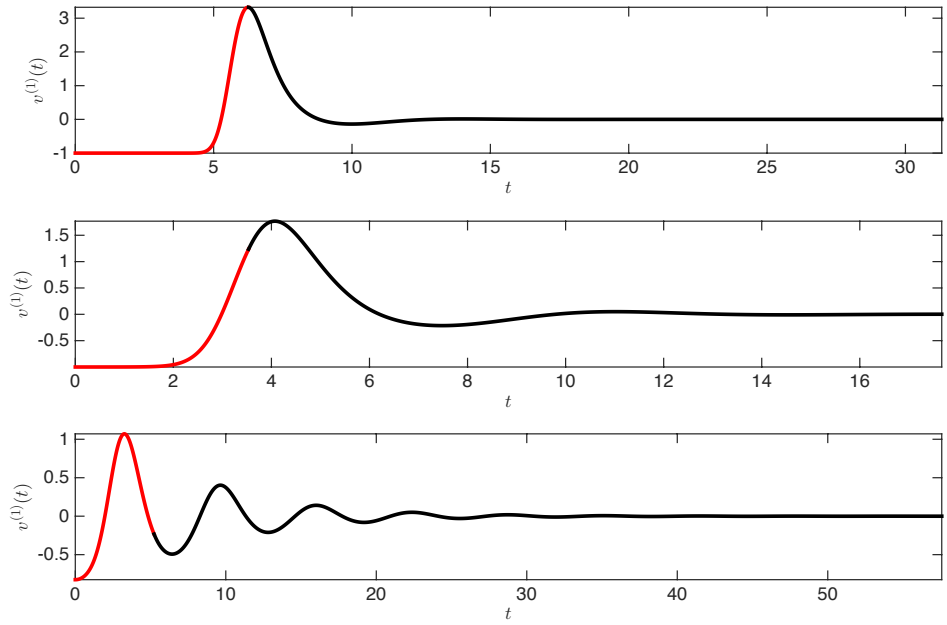


Figure 9.5 – The solution profiles of $v^{(1)}(t)$ for $\beta = 0.5$ (top), $\beta = 1.2$ (middle) and $\beta = 1.9$ (bottom). The parts in red correspond to the part of the solution which was obtained using Chebyshev series, while the parts in black correspond to the part of the solution lying in the local stable manifold computed using Taylor series.

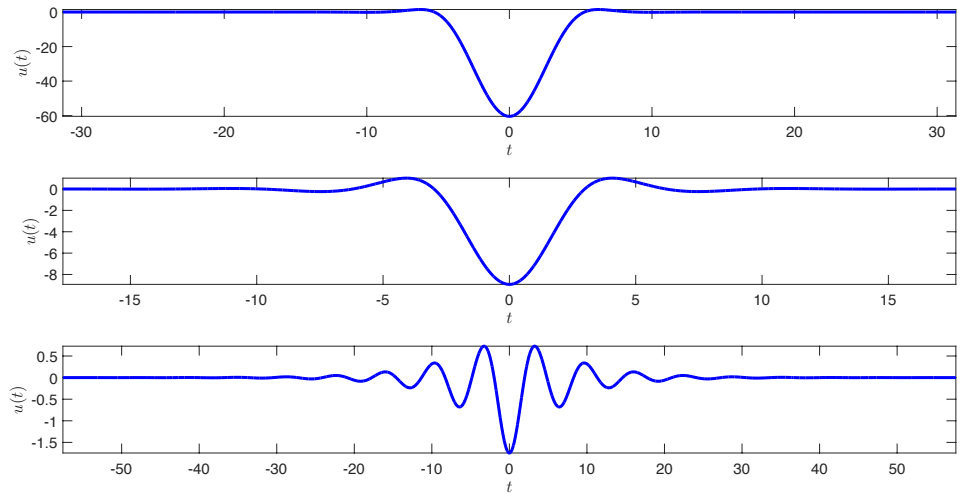


Figure 9.6 – The solution profiles in the variable u of the suspension bridge equation (9.2) for $\beta = 0.5$ (top), $\beta = 1.2$ (middle) and $\beta = 1.9$ (bottom). Notice the different scales of the y-axis for the three solutions.

Chapter 10

Existence and instability of several steady states for a cross-diffusion system

Abstract

This chapter is taken from [71]. We present and apply a computer-assisted method to study steady states of a triangular cross-diffusion system. Our approach consist in an *a posteriori* validation procedure, that is based on using a fixed point argument around a numerically computed solution, in the spirit of the Newton-Kantorovich theorem. It allows us to prove the existence of various non homogeneous steady states for different parameter values. In some situations, we get as many as 13 coexisting steady states. We also apply the *a posteriori* validation procedure to study the linear stability of the obtained steady states, proving that many of them are in fact unstable.

10.1 Introduction

The primary goal of describing physical systems with mathematical models is to be able to explain and predict natural phenomena, within some range of approximation. In some circumstances the mathematical prediction and the experimental evidence don't agree, and a more trustful model is then required. Typically, one can add nonlinear or non homogeneous terms to get a more refined model, but this often seriously complicates the mathematical analysis of the system, which can become very hard, if not impossible, to study analytically. In this situation, numerical simulations allow insight of the phenomena and provide approximate, often very accurate, solutions. Aiming at formulating theorems, a powerful tool to validate approximate solutions into rigorous mathematical statements is provided by the rigorous computational techniques.

The diffusive Lotka-Volterra system, a well known model for population dynamics to study the competition between two species, is paradigmatic of the situation discussed above. It consists in the system

$$\begin{cases} \dfrac{\partial u}{\partial t} = d_1 \Delta u + (r_1 - a_1 u - b_1 v)u, & \text{on } \mathbb{R}_+ \times \Omega, \\ \dfrac{\partial v}{\partial t} = d_2 \Delta v + (r_2 - b_2 u - a_2 v)v, & \text{on } \mathbb{R}_+ \times \Omega, \\ \dfrac{\partial u}{\partial n} = 0 = \dfrac{\partial v}{\partial n}, & \text{on } \mathbb{R}_+ \times \partial\Omega, \end{cases} \tag{10.1}$$

where Ω is a bounded domain of $\mathbb{R}^N$, and $u(t,x), v(t,x) \geq 0$ represent the population densities of two species at time t and position x. The non negative coefficients d_i, r_i, a_i and b_i $(i = 1, 2)$ describe the diffusion, the unhindered growth of the species, the intra-specific competition and the inter-specific competition respectively.

One of the fundamental problems is to determine if and under which assumptions the two species coexist, that converts into proving the existence or non-existence of stable positive equilibrium solutions. Several works have been produced to classify and analyse the stability of the equilibria for (10.1) and of related systems. We refer for instance to [165] for a short review. Of particular interest for our discussion is the result presented in [144]. If the domain Ω is convex, in that paper it is proved that any spatially non-constant equilibrium solution of (10.1) is unstable, if it exists. This implies that if the two species coexist, their densities must be homogeneous in the whole domain.

However, biological observations suggest that two competing species could coexist by forming pattern to avoid each other (a phenomenom called *spatial segregation*). Therefore we would like the model to exhibit stable non homogeneous steady states, but the quoted result shows that this is excluded (at least for convex domains). We point out that stable non homogeneous equilibria have been shown to exist for non convex domains [160], or for systems involving more than two species [142].

In the case of two competing species, to account for the expected stable inhomogeneous steady states, a generalization of (10.1) was proposed in [194]:

$$\begin{cases} \dfrac{\partial u}{\partial t} = \Delta((d_1 + d_{12}v)u) + (r_1 - a_1u - b_1v)u, & \text{on } \mathbb{R}_+ \times \Omega, \\ \dfrac{\partial v}{\partial t} = \Delta((d_2 + d_{21}u)v) + (r_2 - b_2u - a_2v)v, & \text{on } \mathbb{R}_+ \times \Omega, \\ \dfrac{\partial u}{\partial n} = 0 = \dfrac{\partial v}{\partial n}, & \text{on } \mathbb{R}_+ \times \partial\Omega, \end{cases} \tag{10.2}$$

where the added *cross-diffusion* terms $\Delta(uv)$ model that the two species try to avoid each other, by diffusing more when more individuals of the other species are present.

Since its introduction the system (10.2) has been studied extensively, one of the main reason being that it seems to exhibit a much wider variety of steady states than (10.1), especially non homogeneous ones, in accordance to laboratory experiments. Several numerical studies have been presented, displaying intricate bifurcation diagrams of steady states (see for instance [137, 138] and also Figure 10.1). Consider for instance the homogenous equilibria

$$(u_{eq}, v_{eq}) := \left(\frac{r_1a_2 - r_2b_1}{a_1a_2 - b_1b_2}, \frac{r_2a_1 - r_1b_2}{a_1a_2 - b_1b_2} \right) \tag{10.3}$$

in the *strong intra-specific* case, i.e. when $\frac{b_1}{a_2} < \frac{r_1}{r_2} < \frac{a_1}{b_2}$ (the other case $\frac{a_1}{b_2} < \frac{r_1}{r_2} < \frac{b_1}{a_2}$ is known as the *strong inter-specific competition* case). While (u_{eq}, v_{eq}) is stable for (10.1) for any diffusion coefficient $d_1, d_2 \geq 0$, adding strong enough cross-diffusion can destabilize this equilibria, from which new non homogeneous steady states can bifurcate. A link between this *cross-diffusion induced instability* and the standard *Turing instability* for reaction-diffusion systems is made in [137].

While from one side the addition of these nonlinear cross-diffusion terms yields a more reliable model, on the other it seriously complicates the analytical treatment of the system. Even the existence of global classical solutions of (10.2) (completed with non negative initial data) is a challenging question that is still fairly open. Local in time existence can be obtained by the theory of quasilinear parabolic systems [48], but to then get bounds that prevent blowups requires restrictions on the coefficients, as for instance $d_{21} = 0$ [100]. In this particular case, sometimes called *triangular cross-diffusion system*, some recent progresses were also made

in [111] by combining entropy methods with a 3 component reaction-diffusion system without cross-diffusion (first introduced in [137]), that is used to approach (10.2). Entropy methods were also used to improve upon the existing results for the *full* cross-diffusion system, see [140, 110] and the references therein.

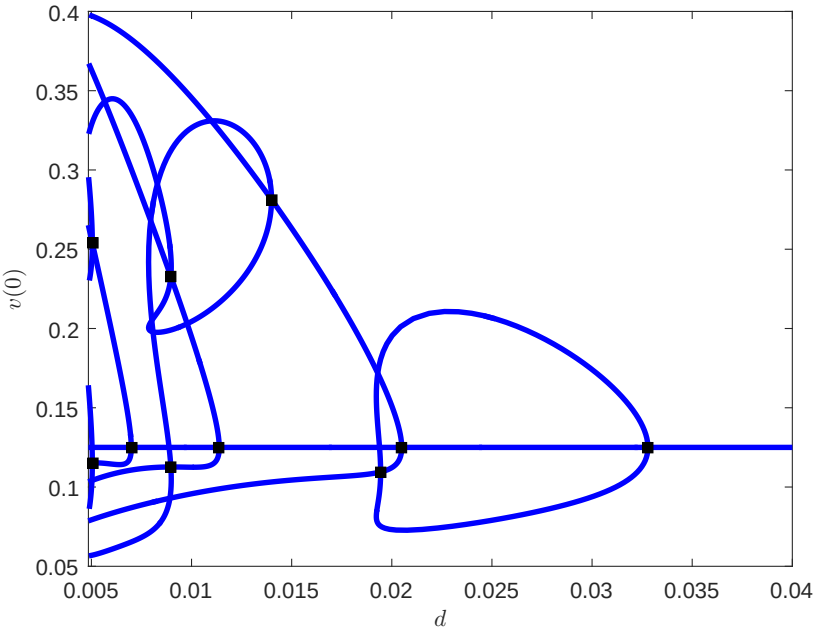


Figure 10.1 – A numerical bifurcation diagram of steady states of (10.2), in the strong intra-specific case. The space domain Ω is $(0,1)$, $r_1 = 5$, $r_2 = 2$, $a_1 = 3$, $a_2 = 3$, $b_1 = 1$, $b_2 = 1$, $d_{12} = 3$, $d_{21} = 0$ and $d_1 = d_2 = d$ is left as the bifurcation parameter.

Existence and stability results for non homogeneous steady states of (10.2) have also been established following different approaches including bifurcations techniques [166], singular perturbation techniques [167, 158] and fixed point index theory [192]. Because of the presence of the cross-diffusion term, the analytical studies are limited to those solutions that are either close the homogeneous steady states, or in very specific parameter ranges. However, the numerically computed bifurcation diagram reveals a very rich structure that includes coexistence of many different steady states for a given set of parameter values, as well as secondary bifurcations. To the best of the author's knowledge, even the existence of these solutions is not yet proved and it seems out of reach of purely analytical techniques.

The aim of this paper is to prove existence and study the linear stability of several non-homogenous steady states of (10.2), significantly far from being perturbations of the homogeneous equilibria, also showing multiplicity of solutions for the same set of parameters. The kind of technique adopted here is often referred to as *validated numerics*, because the goal is to prove the existence of a genuine solution of the problem in a sharp and explicit neighborhood of a numerical one, hence, in this sense, to validate the approximate solution (more details in Section 10.2). More precisely, we follow the so called *radii polynomial* approach, a quite general technique based on the contraction mapping argument that has been adapted to solve several differential problems in areas ranging from dynamical systems to ordinary and partial differential equations through delay differential equation and chaotic dynamics. This is the first time that rigorous computational techniques are applied to PDE system with cross interactions in the leading differential operator. The cross-diffusion terms are indeed a major technical hurdles, since they enfeeble the smoothing effect of the higher order differential operator.

In this work we restrict ourself to the triangular case and assume that the space dimension is 1 (i.e. we fix $d_{21} = 0$ and $\Omega = (0,1)$). The method can easily be extended to higher space dimension (say 2 or 3), but of course the computational cost would increase. The generalization

to a full cross-diffusion system is less straightforward. Indeed, as mentioned above, it will become apparent in the next sections that the cross-diffusion structure hinders the use of our validation method, and that we take advantage of the triangular configuration to overcome this difficulty.

The system we are dealing with is the following:

$$\begin{cases} ((d_1 + d_{12}v)u)'' + (r_1 - a_1u - b_1v)u = 0, & \text{on } (0,1), \\ d_2v'' + (r_2 - b_2u - a_2v)v = 0, & \text{on } (0,1), \\ u'(0) = u'(1) = 0, \\ v'(0) = v'(1) = 0. \end{cases} \tag{10.4}$$

Figure 10.1 depicts a bifurcation diagram of solutions of (10.4), with given values for the parameters r_i, a_i, b_i and d_{12}. This diagram was first obtained numerically in [137], using a 3-component system without cross-diffusion that approaches (10.4). We point out that even in the somewhat restricted framework with $\Omega = (0,1)$ and $d_{12} = 0$, the steady states of (10.2) already manifest very complex and interesting behavior when the parameters vary.

The first result concerns the existence of steady states.

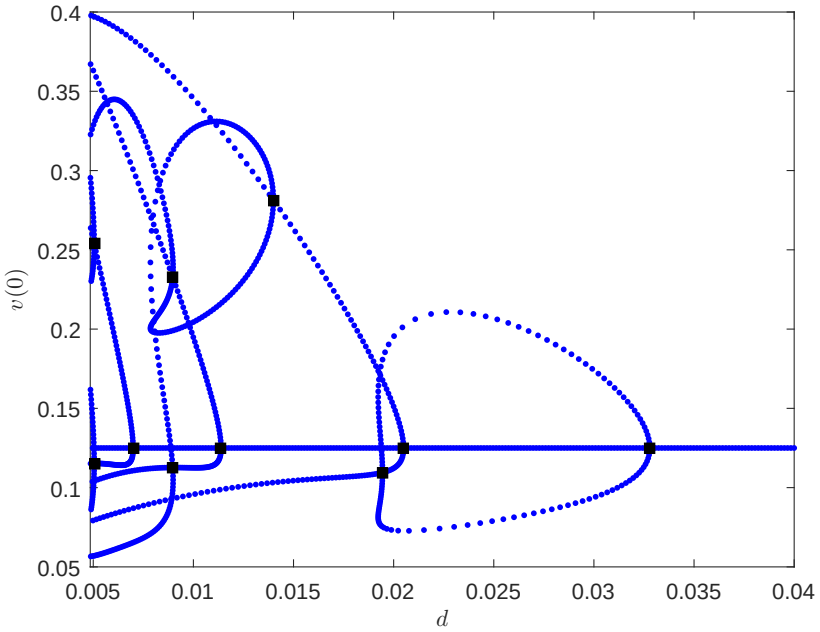


Figure 10.2 – Validated bifurcation diagram of solutions of (10.4). The space domain Ω is $(0,1)$, $r_1 = 5$, $r_2 = 2$, $a_1 = 3$, $a_2 = 3$, $b_1 = 1$, $b_2 = 1$, $d_{12} = 3$, and $d_1 = d_2 = d$ is left as the bifurcation parameter. Each blue dot represents a proved solution. The black squares indicate bifurcations, while the other *apparent* crossings are just due to the projection (i.e. $v(0)$) we used to represent the solutions.

Theorem 10.1.1. *Referring to Figure 10.2, each bullet represents a solution of* (10.4)*, for the parameter values* $r_1 = 5$*,* $r_2 = 2$*,* $a_1 = 3$*,* $a_2 = 3$*,* $b_1 = 1$*,* $b_2 = 1$*,* $d_{12} = 3$*, and* $d_1 = d_2 = d$*. In particular there exists at least 13 different solutions when* $d_1 = d_2 = 0.005$*.*

The proof of each steady states also provides precise qualitative informations about the solution, in terms of explicit bounds of the distance (in some function space, see Section 10.4) between the genuine solution that is proved to exist and a numerically computed approximation (more details in Section 10.5).

In [137], the linear stability of the obtained steady states was also studied (still numerically), suggesting that most of the solutions displayed in the bifurcation diagram of Figure 10.2 are

unstable, while others seems to be stable. In this direction, the second contribution of this paper is a rigorous computational approach to the study of the spectral properties of the equilibria presented above.

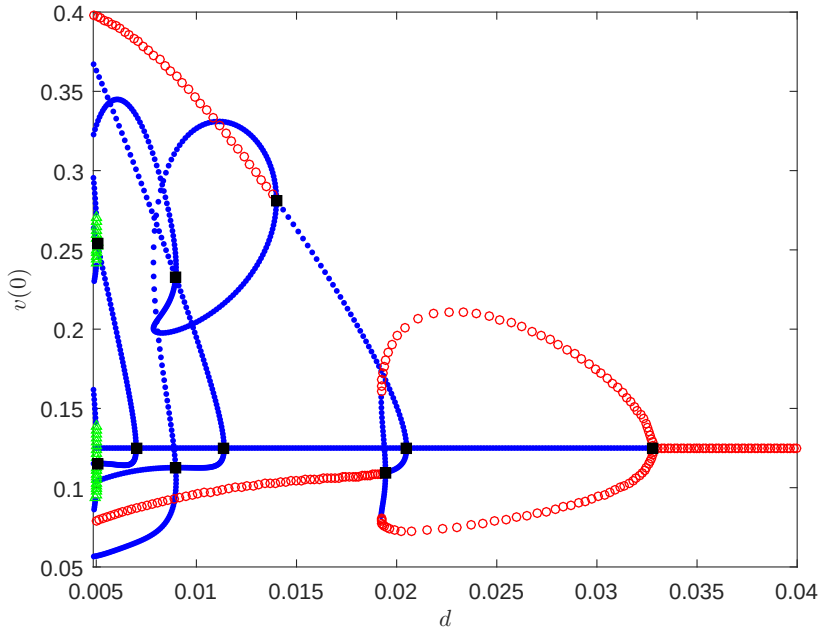


Figure 10.3 – Validated bifurcation diagram of solutions of (10.4). The space domain Ω is $(0,1)$, $r_1 = 5$, $r_2 = 2$, $a_1 = 3$, $a_2 = 3$, $b_1 = 1$, $b_2 = 1$, $d_{12} = 3$, and $d_1 = d_2 = d$ is left as the bifurcation parameter. Each blue dot represents a proved solution, for which we also proved instability. Each green triangle represents a proved solution, that seems unstable numerically but for which we were not able to prove instability. Each red circle represents a proved solution, that seems stable numerically.

Theorem 10.1.2. *Referring to Figure 10.3, each blue bullet represents an unstable steady state. Out of the 13 solutions at parameter value $d = 0.005$, at least 11 are unstable.*

The steady states marked in red in Figure 10.3 and in particular the two solutions out of 13 that are not concerned by the above Theorem seem to be stable. However, at the moment we are not yet able to use our validation method to prove linear stability, as this requires to control the whole spectrum and not just a single eigenvalue. Still, we point out that the straight line of solutions at $v(0) = 0.125$ corresponds to the homogeneous steady state (10.3), for which the linear stability could of course be studied analytically. In particular it could be proven that, before the bifurcation occuring at $d \simeq 0.0328$ the homogeneous steady state is linearly stable. Validated numerics techniques were used successfully to prove stability in other situations ([85, 141], see also [170]), but the adaptation to our problem presents several challenges (mainly due to the cross-diffusion terms, which *muddle* the asymptotic structure of the eigenvalue problem, see Section 10.6) and will be the object of future investigations.

The paper is organized as follows. In Section 10.2, we give a brief exposition of the validated numerics techniques we apply in this work, as well as additional references on the subject. In particular we state the Theorem 10.2.1, that serves as common reference and guideline for both the rigorous computation of the steady states and the rigorous enclosure of the eigenvalues. Section 10.3 is devoted to the introduction of some notations and elementary estimates that are used throughout the paper. In Sections 10.4 to 10.5, we then prove the existence of steady states. More precisely, in Section 10.4.1 we expose how to reformulate the problem of existence of solutions of (10.4) into a framework suitable for Theorem 10.2.1. In Section 10.4.3, we then derive explicit and implementable formulas for the bounds involved in Theorem 10.2.1, and

finally give examples of results in Section 10.5. Sections 10.6 to 10.7 are dedicated to proving the instability of some of these steady states, following the same procedure: suitable reformulation in Section 10.6.1, bounds in Section 10.6.3 and results in Section 10.7.

10.2 Overview of the rigorous computational method

In this section we briefly explain the strategy for both solving (10.4) and computing the linear stability of the steady states by means of validated numerics techniques. Each problem is formulated as solving an equation $F(X) = 0$ defined on a suitable Banach space. The core of the method, first presented in [214], consists in the introduction of an operator T whose fixed points are in one-to-one correspondence with the zeros of $F(X)$. The existence and enclosure of the solution follow by the Banach fixed point theorem once the operator T is proven to be a contraction on some complete set. The explicit determination of the neighborhood on which the operator is a contraction is done efficiently using the *radii polynomial approach* (see [109]), which is reminiscent of the Newton-Kantorovich Theorem. The technique can be summarized in the following statement.

Theorem 10.2.1. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ be Banach spaces and $F : \mathcal{X} \to \mathcal{Y}$ a $\mathcal{C}^1$ function. Let $A : \mathcal{Y} \to \mathcal{X}$ and $A^\dagger : \mathcal{X} \to \mathcal{Y}$ be linear operators, so that AF maps $\mathcal{X}$ into itself. Let $\bar{X} \in \mathcal{X}$ and assume there exist positive constants Y, Z_0, Z_1 and a positive function $r \mapsto Z_2(r)$ such that*

$$\left\|AF(\bar{X})\right\|_{\mathcal{X}} \leq Y \tag{10.5}$$

$$\left\|I - AA^\dagger\right\|_{\mathcal{X}} \leq Z_0 \tag{10.6}$$

$$\left\|A\left(DF(\bar{X}) - A^\dagger\right)\right\|_{\mathcal{X}} \leq Z_1 \tag{10.7}$$

$$\left\|A\left(DF(X) - DF(\bar{X})\right)\right\|_{\mathcal{X}} \leq rZ_2(r) \quad \forall\ X \in \mathcal{B}_{\mathcal{X}}(\bar{X}, r), \tag{10.8}$$

where $\mathcal{B}_{\mathcal{X}}(\bar{X}, r)$ is the closed ball of $\mathcal{X}$, centered at $\bar{X}$ and of radius r, and $|||\cdot|||_{\mathcal{X}}$ denotes the operator norm on $\mathcal{X}$. Define the function P as

$$P(r) = Z_2(r)r^2 - (1 - (Z_0 + Z_1))\, r + Y. \tag{10.9}$$

If there exists $r > 0$ such that $P(r) < 0$, then the operator $T : \mathcal{X} \to \mathcal{X}$ defined as

$$T = I - AF \tag{10.10}$$

has a unique fixed point in $\mathcal{B}_{\mathcal{X}}(\bar{X}, r)$. Moreover if A is injective, then F has a unique zero in $\mathcal{B}_{\mathcal{X}}(\bar{X}, r)$.

We omit the proof of the theorem, that can be found for instance in [109]. Nevertheless, in the next remark we explain the role of the different operators involved in the theorem and some instructions on how to define them. These considerations are detailed and made more explicit in Section 10.4.3 (resp. Section 10.6.3), where we derive the bounds Y, Z_0, Z_1 and $Z_2(r)$ for an F associated to the existence of solutions of (10.4) (resp. their instability). We also mention that in practice, because of the way we define A, its injectivity is in fact implied by the existence of a $r > 0$ such that $P(r) < 0$ (see Proposition 10.4.5).

Remark 10.2.2. • *$\bar{X}$ is chosen as an approximate solution for $F(X) = 0$, to be computed numerically as zero for a finite dimensional approximation of F. The constant Y is the defect bound and measures how far is $\bar{X}$ from being a fixed point of T. Depending on the accuracy of the approximate solution $\bar{X}$, we expect Y to be small.*

- *The $Z_i(r)$, $i = 0, 1, 2$, are meant as bounds for the rate of contraction of the operator T in the ball $\mathcal{B}_{\mathcal{X}}(\bar{X}, r)$. More precisely, $Z_0 + Z_1$ provides a bound for the derivative of T at $\bar{X}$, while $Z_2(r)$ gives a correction for the derivative in the whole ball $\mathcal{B}_{\mathcal{X}}(\bar{X}, r)$. Assume for a moment that Z_2 is constant. Necessary and sufficient conditions for the existence of an $r > 0$ such that $P(r) < 0$ are given by*

$$Z_0 + Z_1 < 1 \quad \textit{and} \quad (1 - (Z_0 + Z_1))^2 > 4Z_2Y.$$

  The two conditions imply that T is a contraction on the ball $\mathcal{B}_{\mathcal{X}}(\bar{X}, r)$. In order to obtain a small Z_1, the operator $A^\dagger$ is conceived as an approximation of $DF(\bar{X})$ (again based on a finite dimensional approximation of F). Similarly, the operator A is constructed as an approximate inverse of $A^\dagger$, which will then make Z_0 small. A key point is to define $A^\dagger$ and A in a smart way, to have good enough approximations while being able to derive tight bounds for Z_1. Typically $\mathcal{Y}$ is a space of functions less regular than $\mathcal{X}$. To this extent, the operator A acts as a smoothing operator.
- *If the nonlinearity of F is a polynomial (say of degree d), Z_2 can be constructed as a polynomial (of degree $d - 2$), and therefore $P(r)$ is indeed a polynomial (of same degree d as F).*
- *All the bounds are obtained through a combination of analytic estimates (because the spaces involved are naturally infinite dimensional) and numerical computations (since they depend on the approximate solution $\bar{X}$). To ensure that all possible round off errors are controlled during the computations, we use an interval arithmetic package (in our case INTLAB [190]).*

As said in the Introduction, this work is far from being the first application of this kind of rigorous computational techniques to solve systems of PDEs, see for instance [60, 92, 96, 109, 120, 119]. Particularly related to our work is the result presented in [79]. In that paper a similar method was used to rigorously validate a bifurcation diagram of steady states of a 3-component reaction diffusion system (without cross-diffusion term). The system considered in [79] depends on a parameter ε and it has the property that its steady states approach the solutions of (10.4) as ε goes to 0, see [138]. However, the proof could only be made for a fixed (small) ε and the limit case $\varepsilon = 0$ is in some sense singular. Therefore the cross-diffusion case could not be handled.

More broadly, techniques similar to the one presented in this work were developed to prove the existence of fixed points, periodic orbits, invariant manifolds and connecting orbits for ordinary differential equations, infinite dimensional maps, partial differential equations and delay differential equations (see for instance [63, 66, 67, 93, 95, 50, 122, 153, 157]). We also mention the existence of comparable techniques where, instead of computing A by using a finite dimensional truncation as we do, a bound on the norm of the inverse of $DF(\bar{X})$ is obtained via spectral estimations (see [187, 161] and the references therein). Instead of the contraction mapping principle, computer assisted proofs in dynamical systems are frequently based on topological tools as covering relations, the Brouwer degree, the fixed point index, the Conley index, see for instance [91, 125, 174, 175, 216]. To conclude this paragraph, we refer the interested reader to [54, 56, 117, 215] for a list, surely not exhaustive, of rigorous computational techniques developed to solve a variety of problems, not necessarily in the area of dynamical systems.

Theorem 10.2.1 is the cornerstone for all the proofs that are presented in this paper. Whatever problem we want to solve, once the system is rephrased as a zero finding problem and the hypotheses of the theorem are verified, then the proof follows as application of the theorem. Thus, for a given problem $(\mathcal{P})$, we proceed as follows:

1. Introduce a Banach space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ and a $\mathcal{C}^1$ function F defined on $\mathcal{X}$ so that the solutions of $F(X) = 0$ correspond to solutions of $(\mathcal{P})$;
2. Compute a numerical approximation $\bar{X} \in \mathcal{X}$ so that $F(\bar{X}) \approx 0$ and define the linear operators A and $A^\dagger$;

3. Define and compute the bounds $Y, Z_i(r)$ satisfying (10.5)-(10.8);
4. Check that $P(r)$ given in (10.9) is negative for some $r > 0$.

If the last condition is met, the existence of a solution X for $F(X) = 0$ is proved in the form specified in the theorem.

In the sequel we detail each step of the above list for the problem of proving existence of steady states (Section 10.4) and for the problem of proving their linear instability (Section 10.6).

10.3 Sequence space, convolutions and norm estimates

The solutions of system (10.4) as well as the eigenfunctions of the linearised system are sought in the form of Fourier series, which is fairly natural given the boundary conditions included in (10.4). This approach also provides a very convenient setting to apply our validated numerics technique. In this section we introduce the sequence space relevant for our analysis and we recall some useful properties. The material presented here is standard and mainly included for the sake of completeness and to fix some notations.

Definition 10.3.1. *Let $\nu > 1$. For any sequence $u = (u_k)_{k\geq 0} \in \mathbb{C}^{\mathbb{N}}$ we define the ν-norm of u as*

$$\|u\|_\nu = |u_0| + 2\sum_{k\geq 1} |u_k|\, \nu^{|k|},$$

and introduce the space

$$\ell^1_\nu = \left\{ u \in \mathbb{C}^{\mathbb{N}},\ \|u\|_\nu < \infty \right\}.$$

We also define $\ell^1_\nu(\mathbb{R})$ the subspace of ℓ^1_ν made of real sequences.

Definition 10.3.2. *For any $u, v \in \ell^1_\nu$, we define the sequences $(u * v)$, $(u \star v)$ and $(u \bullet v)$ as*

$$(u * v)_k = \sum_{\substack{k_1,k_2\in\mathbb{Z}\\ k_1+k_2=k}} u_{|k_1|}v_{|k_2|}, \quad (u \star v)_k = \sum_{\substack{k_1,k_2\in\mathbb{Z}\\ k_1+k_2=k}} \mathrm{sgn}(k_1)u_{|k_1|}v_{|k_2|},$$

$$(u \bullet v)_k = \sum_{\substack{k_1,k_2\in\mathbb{Z}\\ k_1+k_2=k}} \mathrm{sgn}(k_1)\,\mathrm{sgn}(k_2)u_{|k_1|}v_{|k_2|},$$

where $\mathrm{sgn}(k)$ denotes the sign of k and $\mathrm{sgn}(0) = 0$.

The reason for introducing three different convolution products is that we will deal with multiplications of both even and odd functions. The role played by each of the above operations is as follows.

Let u and v in ℓ^1_ν and consider the even functions (still denoted u and v) defined by

$$u(x) = u_0 + 2\sum_{k\geq 1} u_k \cos(kx), \quad v(x) = v_0 + 2\sum_{k\geq 1} v_k \cos(kx),$$

then $(u * v)$ is the sequence of Fourier coefficients of the product function uv, i.e.

$$u(x)v(x) = (u * v)_0 + 2\sum_{k\geq 1} (u * v)_k \cos(kx).$$

If instead, u is the odd function given by

$$u(x) = 2\sum_{k\geq 1} u_k \sin(kx),$$

then $(u \star v)$ provides the sequence of Fourier coefficients of the product function uv, i.e.

$$u(x)v(x) = 2\sum_{k\geq 1}(u\star v)_k \sin(kx).$$

Finally, if both the functions u and v are odd

$$u(x) = 2\sum_{k\geq 1} u_k \sin(kx), \quad v(x) = 2\sum_{k\geq 1} v_k \sin(kx),$$

then $(u \bullet v)$ is the sequence of Fourier coefficients of the product function uv, i.e.

$$u(x)v(x) = (u\bullet v)_0 + 2\sum_{k\geq 1}(u\bullet v)_k \cos(kx).$$

We also recall that ℓ^1_ν equipped with any of the three convolution products $*$, $\star$ or $\bullet$ is a Banach algebra. More precisely we have the following estimate.

Lemma 10.3.3. *Let $u, v \in \ell^1_\nu$ and $\circ \in \{*, \star, \bullet\}$ any of the convolution products. Then*

$$\|u \circ v\|_\nu \leq \|u\|_\nu \|v\|_\nu .$$

Consider now $B : \ell^1_\nu \to \ell^1_\nu$ a bounded linear operator. To the operator B is associated an infinite dimensional matrix (still denoted by B) so that $(Bu)_k = \sum_{j\geq 0} B(k,j)u_j$ for all u in ℓ^1_ν. We denote by $|\!|\!| B |\!|\!|_\nu$ the operator norm of B, i.e.

$$|\!|\!| B |\!|\!|_\nu = \sup_{\|u\|_\nu = 1} \|Bu\|_\nu . \tag{10.11}$$

Lemma 10.3.4. *Let $B : \ell^1_\nu \to \ell^1_\nu$ be a linear operator and consider $B(k,j)$ the matrix representation. Then*

$$|\!|\!| B |\!|\!|_\nu = \sup_{j\geq 0} \frac{1}{\nu^j} \sum_{k\geq 0} |B(k,j)| \nu^k.$$

A linear functional $b : \ell^1_\nu \to \mathbb{R}$ is a particular case of the general operator B given above, when $(Bu)_k = 0$ for any $k \neq 0$. The linear functional b acts on u as $bu = \sum_{j\geq 0} b_j u_j$ and the operator norm of b is then given by , $|\!|\!| b |\!|\!|_\nu = \sup_{j\geq 0} \frac{|b_j|}{\nu^{|j|}}$. We point out that this last formula is linked to the fact that the dual space of ℓ^1 with weight ν is isometric to the space ℓ^∞ with weight ν^{-1}.

When dealing with numerical computations, we need to consider only a finite number of coefficients u_k in the infinite sequence $u \in \ell^1_\nu$, that is we consider a finite dimensional projection of u.

Definition 10.3.5. *Let $u \in \ell^1_\nu$. For $m \in \mathbb{N}$ we denote $\hat{u}^m$ the* truncated part *(i.e. the finite m-dimensional projection) of u and $\check{u}^m$ the* tail part *(i.e. infinite dimensional complement) of u, given as*

$$\hat{u}^m_k = \begin{cases} u_k, & k < m, \\ 0, & k \geq m, \end{cases} \quad \text{and} \quad \check{u}^m_k = \begin{cases} 0, & k < m, \\ u_k, & k \geq m. \end{cases}$$

By a slight abuse of notation, we also refer to $\hat{u}^m$ as the finite dimensional vector $(u_k)_{0\leq k<m}$. Moreover, when there is no possible confusion about the dimension of the projection we may drop the exponent m and simply use $\hat{u}$ and $\check{u}$.

We end this section with an estimate used to bound a convolution with a *tail term*.

Lemma 10.3.6. *Let $m \in \mathbb{N}$ and $u, v \in \ell^1_\nu$ and $\circ \in \{*, \star, \bullet\}$ any of the convolution products. Then, for all $k \geq 0$,*

$$|u \circ \check{v}^m|_k \leq \Phi^m_k(u,\nu) \|v\|_\nu ,$$

where

$$\Phi^m_k(u,\nu) = \sup_{|l|\geq m} \frac{\left|u_{|l-k|}\right|}{\nu^{|l|}}.$$

10.4 Framework for the existence of steady states

We are now concerned with the proof of existence of non homogeneous solutions of system (10.4). According to the algorithm outlined in Section 10.2, the first step is to reformulate the problem in the form $F(X) = 0$, where F is defined on a proper Banach space. Then we introduce the linear operators A and $A^\dagger$, and finally we provide the definition of the bounds $Y, Z_0, Z_1, Z_2(r)$. The latter are then combined to define the radii polynomial $P(r)$.

10.4.1 Existence of steady states: the function F

We preventively need to transform (10.4) into an equivalent system that is more amenable to the application of validated numerics techniques. We introduce further unknown functions and change of coordinates in order to remove the cross-diffusion nonlinearity and to obtain a system with only polynomial nonlinearities (which will be useful when deriving the validation estimates). Then we discretize the obtained system by using Fourier series and finally we introduce F and the Banach space $\mathcal{X}$ in which we look for the solutions.

Auxiliary functions and polynomial system

In order to remove the cross-diffusion nonlinearity we introduce the function w defined as

$$w = (d_1 + d_{12}v)u.$$

Expressing (10.4) in term of the unknowns v and w gives simpler higher order terms, but the nonlinear terms become rational functions. To keep the nonlinearity in the form of polynomials, define the function p as

$$p = \frac{1}{d_1 + d_{12}v}, \tag{10.12}$$

so that $u = pw$. In term of w, v, p, the system (10.4) takes the form

$$\begin{cases} w'' + (r_1 - a_1 pw - b_1 v)pw = 0, & \text{on } (0,1), \\ d_2 v'' + (r_2 - b_2 pw - a_2 v)v = 0, & \text{on } (0,1), \\ v'(0) = v'(1) = 0, \\ w'(0) = w'(1) = 0 \end{cases}$$

with p given above as a function of $v(x)$. However, we want to also treat p as an independent unknown, on the same footing as w and v. For this, it is enough to append a differential equation and some initial conditions uniquely satisfied by the requested function $p(x)$. Let us consider the equation

$$p' = -d_{12}p^2 v',$$

together with the constraint

$$p(0)(d_1 + d_{12}v(0)) = 1.$$

It is straightforward to check that the function $p(x)$ in (10.12) is the only solution of such initial value problem. Finally, introducing a variable s for v', we obtain the system

$$\begin{cases} w'' + (r_1 - a_1 pw - b_1 v)pw = 0, & \text{on } (0,1), \\ d_2 s' + (r_2 - b_2 pw - a_2 v)v = 0, & \text{on } (0,1), \\ p' + d_{12} s p^2 = 0, & \text{on } (0,1), \\ v' - s = 0, & \text{on } (0,1), \\ p(0)(d_1 + d_{12}v(0)) = 1, \\ v'(0) = v'(1) = 0, \\ w'(0) = w'(1) = 0, \end{cases} \tag{10.13}$$

to be solved in the unknowns w, v, p, s. We point out that the usage of the variable p to recover polynomial nonlinearities is inspired from [154], where this technique was introduced in the context of validated numerics.

Algebraic system in Fourier space

We now expand the unknown functions and we project the differential system (10.13) onto the Fourier basis. Because of the boundary conditions and (10.12), $v(x)$, $w(x)$ and $p(x)$ are written as cosine series. On the opposite, since $s(x) = v'(x)$, the function $s(x)$ is expanded on the sines basis. Precisely, consider

$$\begin{aligned} w(x) &= w_0 + 2\sum_{k\geq 1} w_k \cos(k\pi x), & v(x) &= v_0 + 2\sum_{k\geq 1} v_k \cos(k\pi x), \\ p(x) &= p_0 + 2\sum_{k\geq 1} p_k \cos(k\pi x), & s(x) &= s_0 + 2\sum_{k\geq 1} s_k \sin(k\pi x). \end{aligned} \tag{10.14}$$

The addition of the s_0 coefficient (which is clearly zero since $s = v'$) is deliberate, because we see each of the sequence of Fourier coefficients $(v_k)_{k\geq 0}$, $(w_k)_{k\geq 0}$, $(p_k)_{k\geq 0}$ and $(s_k)_{k\geq 0}$ as an element of ℓ^1_ν. Plugging these series expansions into (10.13) and then projecting back onto the cosine/sine basis, we obtain the following (infinite dimensional) algebraic system

$$\begin{cases} -(\pi k)^2 w_k + r_1(p*w)_k - a_1(p*p*w*w)_k - b_1(p*v*w)_k = 0, & \forall\ k\in\mathbb{N}, \\ d_2\pi k s_k + r_2 v_k - a_2(v*v)_k - b_2(p*v*w)_k = 0, & \forall\ k\in\mathbb{N}, \\ -\pi k p_k + d_{12}(s\star p*p)_k = 0, & \forall\ k\in\mathbb{N}, \\ -\pi k v_k - s_k = 0, & \forall\ k\in\mathbb{N}, \\ \left(p_0 + 2\displaystyle\sum_{k\geq 1} p_k\right)\left(d_1 + d_{12}\left(v_0 + 2\sum_{k\geq 1} v_k\right)\right) - 1 = 0, \end{cases} \tag{10.15}$$

to be solved for the unknown sequences $w = \{w_k\}_{k\geq 0}, v = \{v_k\}_{k\geq 0}, p = \{p_k\}_{k\geq 0}, s = \{s_k\}_{k\geq 0}$.

Notice that because of the equation $-\pi k v_k - s_k = 0$ with $k = 0$, any solution of this system does indeed satisfy $s_0 = 0$. Notice also that for $k = 0$, the equation $-\pi k p_k + d_{12}(s\star p*p)_k = 0$ is an identity. Indeed $-0\pi p_0 = 0$, and it follows from the definition of the convolution products $\star$ that $(s\star p*p)_0 = 0$. In other words, since $p(x)$ is even and $s(x)$ is odd, the product sp^2 is odd, hence the 0-th Fourier coefficients vanishes. Therefore this equation can be removed and we are then left with a *square* system.

The $F = 0$ formulation

For $\nu > 1$, we define $\mathcal{X}_\nu = \left(\ell^1_\nu(\mathbb{R})\right)^4$, where ℓ^1_ν is given in Definition 10.3.1, and denote by $X = (v, w, p, s)$ any element in $\mathcal{X}_\nu$. We also use the notation X_k to denote (v_k, w_k, p_k, s_k). We endow $\mathcal{X}_\nu$ with the norm

$$\|X\|_{\mathcal{X}_\nu} = \|v\|_\nu + \|w\|_\nu + \|p\|_\nu + \|s\|_\nu, \tag{10.16}$$

which makes it a Banach space. We then define the function $F = (F^{(v)}, F^{(w)}, F^{(p)}, F^{(s)})$ acting on $\mathcal{X}_\nu$ by

$$F_k^{(v)}(X) = -\pi k v_k - s_k, \qquad \forall\ k \in \mathbb{N}, \quad (10.17)$$

$$F_k^{(w)}(X) = -(\pi k)^2 w_k + r_1(p * w)_k - a_1(p * p * w * w)_k - b_1(p * v * w), \qquad \forall\ k \in \mathbb{N}, \quad (10.18)$$

$$F_0^{(p)}(X) = \left(p_0 + 2\sum_{k\geq 1} p_k\right)\left(d_1 + d_{12}\left(v_0 + 2\sum_{k\geq 1} v_k\right)\right) - 1, \quad (10.19)$$

$$F_k^{(p)}(X) = -\pi k p_k + d_{12}(s \star p * p)_k, \qquad \forall\ k \geq 1, \quad (10.20)$$

$$F_k^{(s)}(X) = d_2 \pi k s_k + r_2 v_k - a_2(v * v)_k - b_2(p * v * w)_k, \qquad \forall\ k \in \mathbb{N}. \quad (10.21)$$

The next Lemma summarizes and justifies in a precise statement all the formal computations and substitutions made previously in this section, with the goal of solving system (10.4).

Lemma 10.4.1. *Let $\nu > 1$. Assume that there exists $X \in \mathcal{X}_\nu$ such that $F(X) = 0$ and consider as in* (10.14) *the functions v, w, p and s. Assume also that the coefficients $(v_k)_{k\geq 0}$ and $(w_k)_{k\geq 0}$ are such that the functions v and w are positive. Define the function $u = pw$. Then u and v are smooth positive functions that solve* (10.4)*.*

Proof. First notice that since $X \in \mathcal{X}_\nu$ with $\nu > 1$, the Fourier coefficients are decaying exponentially fast to 0, and thus the functions v, w, p and s are well defined and smooth (in fact analytic) 2-periodic functions. Then, having $F(X) = 0$ means exactly that the sequences $(v_k)_{k\geq 0}$, $(w_k)_{k\geq 0}$, $(p_k)_{k\geq 0}$ and $(s_k)_{k\geq 0}$ solve (10.15), which in turn implies that the functions v, w, p and s solve (10.13). All the derivatives needed in (10.13) are legitimate thanks to the exponential decay of the coefficients. Besides, since p satisfies the differential equation $p' + d_{12}p^2 v' = 0$ and $p(0)(d_1 + d_{12}v(0)) = 1$, by uniqueness we have

$$p = \frac{1}{d_1 + d_{12}v}.$$

Therefore $w = \frac{u}{p} = (d_1 + d_{12}v)u$ and (u, v) does indeed solve (10.4) (the boundary condition for u is also satisfied since $u'(0) = p'(0)w(0) + p(0)w'(0) = 0$ and $u'(1) = p'(1)w(1) + p(1)w'(1) = 0$). □

10.4.2 Existence of steady states: the operators A and $A^\dagger$

As outlined in Remark (10.2.2), the definition of the operators A and $A^\dagger$ is based on some approximate solution $\bar{X}$, which is computed as numerical zero of a finite dimensional projection of $F(X) = 0$.

Extending the notations introduced in Definition 10.3.5, for $X \in \mathcal{X}_\nu$ we denote $\hat{X}^m$ the vector of truncated sequences, i.e.

$$\hat{X}^m = (\hat{v}^m, \hat{w}^m, \hat{p}^m, \hat{s}^m).$$

Similarly,

$$\hat{F}^m = \left(\left(F_k^{(v)}\right)_{0\leq k<m}, \left(F_k^{(w)}\right)_{0\leq k<m}, \left(F_k^{(p)}\right)_{0\leq k<m}, \left(F_k^{(s)}\right)_{0\leq k<m}\right).$$

We consider $\hat{F}^m$ as acting on truncated sequences $\hat{X}^m$ only, so that we can see it as a function mapping $\mathbb{R}^{4m}$ to itself. Therefore, finding $\hat{X}^m$ such that $\hat{F}^m(\hat{X}^m) = 0$ is a finite dimensional problem that can be solved numerically. We now assume to have computed numerically a zero of $\hat{F}^m$, denoted by $\bar{X}$.

The linear operator $A^\dagger$ is defined as an approximation of $DF(\bar{X})$. However, since we will also need to construct an approximate inverse of $A^\dagger$, $A^\dagger$ is required to have a *simple* structure. In practice we impose that $A^\dagger$ acts diagonally on the *tail* $\{X_k\}_{k\geq m}$. More precisely, we define $A^\dagger$ (acting on $X=(v,w,p,s)\in\mathcal{X}_\nu$), as

$$\widehat{A^\dagger X}^m = D\hat{F}^m(\bar{X})\hat{X}^m, \tag{10.22}$$

and

$$\left(A^\dagger X\right)_k = \left(-\pi k v_k, -(\pi k)^2 w_k, -\pi k p_k, d_2\pi k s_k\right), \quad \forall\ k\geq m.$$

The operator A is then constructed as an approximate inverse of $A^\dagger$. We consider $\hat{A}^m$ a numerically computed inverse of $D\hat{F}^m(\bar{X})$ and define A (acting on $X=(v,w,p,s)\in\mathcal{X}_\nu$), as

$$\widehat{AX}^m = \hat{A}^m\hat{X}^m,$$

and

$$(AX)_k = \left(-(\pi k)^{-1}v_k, -(\pi k)^{-2}w_k, -(\pi k)^{-1}p_k, (d_2\pi k)^{-1}s_k\right), \quad \forall\ k\geq m.$$

The definition of A and the fact that ℓ^1_ν is a algebra for both convolution products $*$ and $\star$ (see Lemma 10.3.3) ensure that AF does map $\mathcal{X}_\nu$ into itself, as requested in the hypothesis of Theorem 10.2.1.

Remark 10.4.2. *To define the action of $A^\dagger$ on the tail space, we simply kept the asymptotically dominant terms of the derivative $DF(\bar{X})$. Since these terms act diagonally, we are able to easily and analytically invert the tail of $A^\dagger$ and hence to define A. However, the fact that the dominant terms of the derivative are diagonal is not a mere happenstance, rather it is the result of the various reformulations performed in Section 10.4.1. Had we not introduced the function w, the Fourier expansion of the cross-diffusion term would create a messy dominant expression that we would not be able to invert analytically.*

10.4.3 Existence of steady states: the bounds Y and $Z_i(r)$

Having the Banach space $\mathcal{X}_\nu$, the function F, the approximate solution $\bar{X}=(\bar{v},\bar{w},\bar{p},\bar{s})$ and the operators A, $A^\dagger$ in hands, we now proceed to derive computable bounds Y, Z_0, Z_1 and Z_2 satisfying (10.5)-(10.8) (for $\mathcal{X}=\mathcal{X}_\nu$).

The bound Y

The definition of the bound Y is rather straightforward, and we just consider

$$Y = \left\|AF(\bar{X})\right\|_{\mathcal{X}_\nu}. \tag{10.23}$$

The key observation here is that Y can be computed explicitly. Indeed, we recall that $\bar{X}\in\hat{X}^m$ is a truncated sequence, i.e. $\bar{X}_k=(\bar{v}_k,\bar{w}_k,\bar{p}_k,\bar{s}_k)=(0,0,0,0)$ for all $k\geq m$. Therefore we have

$$F^{(v)}_k(\bar{X})=0, \quad \forall\ k\geq m, \qquad F^{(w)}_k(\bar{X})=0, \quad \forall\ k\geq 4m-3,$$
$$F^{(p)}_k(\bar{X})=0, \quad \forall\ k\geq 3m-2, \qquad F^{(s)}_k(\bar{X})=0, \quad \forall\ k\geq 3m-2,$$

and thus $F(\bar{X})$ only has a finite number of non zero coefficients. This is also true for $AF(\bar{X})$ (thanks to the diagonal structure of the tail of A), and therefore $\left\|AF(\bar{X})\right\|_{\mathcal{X}_\nu}$ can be evaluated on a computer. To be completely precise, what we mean by (10.23) is that a (sharp) upper bound of $\left\|AF(\bar{X})\right\|_{\mathcal{X}_\nu}$ can be computed using interval arithmetic, and that we define Y to be this upper bound. We are going to repeat this abuse of language whenever we define bounds that involve terms that have to be evaluated on a computer.

The bound Z_0

In this section we focus on getting a bound Z_0 satisfying (10.6). Here and thereafter, when dealing with linear operators on $\mathcal{X}_\nu$, it is convenient to use a *block notation*. For a linear operator $B : \mathcal{X}_\nu \to \mathcal{X}_\nu$, we consider the decomposition

$$B = \begin{pmatrix} B^{(v,v)} & B^{(v,w)} & B^{(v,p)} & B^{(v,s)} \\ B^{(w,v)} & B^{(w,w)} & B^{(w,p)} & B^{(w,s)} \\ B^{(p,v)} & B^{(p,w)} & B^{(p,p)} & B^{(p,s)} \\ B^{(s,v)} & B^{(s,w)} & B^{(s,p)} & B^{(s,s)} \end{pmatrix}, \quad \text{each } B^{(i,j)} : \ell^1_\nu(\mathbb{R}) \to \ell^1_\nu(\mathbb{R})$$

so that, for $X = (v, w, p, s) \in \mathcal{X}_\nu$,

$$(BX)^{(v)} = B^{(v,v)}v + B^{(v,w)}w + B^{(v,p)}p + B^{(v,s)}s,$$

and similarly for the other components. Thus, recalling (10.16) and the operator norm (10.11),

$$\begin{aligned} \|BX\|_{\mathcal{X}_\nu} &= \left\|(BX)^{(v)}\right\|_\nu + \left\|(BX)^{(w)}\right\|_\nu + \left\|(BX)^{(p)}\right\|_\nu + \left\|(BX)^{(s)}\right\|_\nu && (10.24)\\ &\le \Theta_B^{(v)} \|v\|_\nu + \Theta_B^{(w)} \|w\|_\nu + \Theta_B^{(p)} \|p\|_\nu + \Theta_B^{(s)} \|s\|_\nu && (10.25)\\ &\le \max\left[\Theta_B^{(v)}, \Theta_B^{(w)}, \Theta_B^{(p)}, \Theta_B^{(s)}\right] \|X\|_{\mathcal{X}_\nu}, && (10.26) \end{aligned}$$

where

$$\Theta_B^{(i)} = \|\!| B^{(v,i)} |\!\|_\nu + \|\!| B^{(w,i)} |\!\|_\nu + \|\!| B^{(p,i)} |\!\|_\nu + \|\!| B^{(s,i)} |\!\|_\nu, \quad \forall i \in \{v, w, p, s\}.$$

Therefore, we define

$$Z_0 = \max\left[\Theta^{(v)}_{I-AA^\dagger}, \Theta^{(w)}_{I-AA^\dagger}, \Theta^{(p)}_{I-AA^\dagger}, \Theta^{(s)}_{I-AA^\dagger}\right]. \tag{10.27}$$

Notice that, since the tail part of A and $A^\dagger$ are exact inverse of each other by definition, the tail part of $I - AA^\dagger$ is zero. Therefore, each block in the decomposition of $I - AA^\dagger$ has only finitely many non zero coefficients, and each $\Theta^{(i)}_{I-AA^\dagger}$ can be computed using Lemma 10.3.4, the supremum and the sum ranging only over finitely many coefficients.

The bound Z_1

In this section we focus on getting a bound Z_1 satisfying 10.7.

Lemma 10.4.3. *Let $\hat\alpha^m_v, \hat\alpha^m_w, \hat\alpha^m_p, \hat\alpha^m_s$ be vectors in $\mathbb{R}^{4m}$ each, defined as*

$$(\hat\alpha^m_v)_0 = \begin{pmatrix} 0 \\ \Phi^m_0(-b_1(\bar p * \bar w), \nu) \\ \left|d_{12}\left(\bar p_0 + 2\sum_{k\ge1}\bar p_k\right)\right| \frac{2}{\nu^m} \\ \Phi^m_0(-2a_2\bar v - b_2(\bar p * \bar w), \nu) \end{pmatrix}, \quad (\hat\alpha^m_w)_0 = \begin{pmatrix} 0 \\ \Phi^m_0(r_1\bar p - 2a_1(\bar p * \bar p * \bar w) - b_1(\bar p * \bar v), \nu) \\ 0 \\ \Phi^m_0(-b_2(\bar p * \bar v), \nu) \end{pmatrix},$$

$$(\hat\alpha^m_p)_0 = \begin{pmatrix} 0 \\ \Phi^m_0(r_1\bar w - 2a_1(\bar p * \bar w * \bar w) - b_1(\bar v * \bar w), \nu) \\ \left|d + d_{12}\left(\bar v_0 + 2\sum_{k\ge1}\bar v_k\right)\right| \frac{2}{\nu^m} \\ \Phi^m_0(-b_2(\bar v * \bar w), \nu) \end{pmatrix}, \quad (\hat\alpha^m_s)_0 = \begin{pmatrix} 0 \\ 0 \\ \Phi^m_0(d_{12}(\bar p * \bar p), \nu) \\ 0 \end{pmatrix}$$

and for each $1 \le k < m$,

$$(\hat{\alpha}_v^m)_k = \begin{pmatrix} 0 \\ \Phi_k^m(-b_1(\bar{p}*\bar{w}),\nu) \\ 0 \\ \Phi_k^m(-2a_2\bar{v} - b_2(\bar{p}*\bar{w}),\nu) \end{pmatrix}, \quad (\hat{\alpha}_w^m)_k = \begin{pmatrix} 0 \\ \Phi_k^m(r_1\bar{p} - 2a_1(\bar{p}*\bar{p}*\bar{w}) - b_1(\bar{p}*\bar{v}),\nu) \\ 0 \\ \Phi_k^m(-b_2(\bar{p}*\bar{v}),\nu) \end{pmatrix}$$

$$(\hat{\alpha}_p^m)_k = \begin{pmatrix} 0 \\ \Phi_k^m(r_1\bar{w} - 2a_1(\bar{p}*\bar{w}*\bar{w}) - b_1(\bar{v}*\bar{w}),\nu) \\ \Phi_k^m(2d_{12}(\bar{s}\star\bar{p}),\nu) \\ \Phi_k^m(-b_2(\bar{v}*\bar{w}),\nu) \end{pmatrix}, \quad (\hat{\alpha}_s^m)_k = \begin{pmatrix} 0 \\ 0 \\ \Phi_k^m(d_{12}(\bar{p}*\bar{p}),\nu) \\ 0 \end{pmatrix}.$$

Define

$$\begin{aligned} Z_1 = \max &\left[\||A|\hat{\alpha}_v^m\|_{\mathcal{X}_\nu}, \||A|\hat{\alpha}_w^m\|_{\mathcal{X}_\nu}, \left\||A|\hat{\alpha}_p^m\right\|_{\mathcal{X}_\nu}, \||A|\hat{\alpha}_s^m\|_{\mathcal{X}_\nu} \right] \\ &+ \max \left[\left(\frac{\|b_1(\bar{p}*\bar{w})\|_\nu}{(\pi m)^2} + \frac{\|r_2 - 2a_2\bar{v} - b_2(\bar{p}*\bar{w})\|_\nu}{d\pi m} \right), \right. \\ &\left(\frac{\|r_1\bar{p} - 2a_1(\bar{p}*\bar{p}*\bar{w}) - b_1(\bar{p}*\bar{v})\|_\nu}{(\pi m)^2} + \frac{\|b_2(\bar{p}*\bar{v})\|_\nu}{d\pi m} \right), \\ &\left(\frac{\|r_1\bar{w} - 2a_1(\bar{p}*\bar{w}*\bar{w}) - b_1(\bar{v}*\bar{w})\|_\nu}{(\pi m)^2} + \frac{\|2d_{12}(\bar{s}\star\bar{p})\|_\nu}{\pi m} + \frac{\|b_2(\bar{v}*\bar{w})\|_\nu}{d\pi m} \right), \\ &\left. \left(\frac{1}{\pi m} + \frac{\|d_{12}(\bar{p}*\bar{p})\|_\nu}{\pi m} \right) \right]. \end{aligned} \tag{10.28}$$

Then

$$Z_1 \ge \left\| A\left(DF(\bar{X}) - A^\dagger \right) \right\|_{\mathcal{X}_\nu}.$$

Proof. According to (10.7), we need a bound for

$$A\left(DF(\bar{X}) - A^\dagger \right) X,$$

for $X \in B_{\mathcal{X}_\nu}(0,1)$. Denoting $U = \left(DF(\bar{X}) - A^\dagger \right) X$ and using the triangular inequality, we have

$$\begin{aligned} \left\| A\left(DF(\bar{X}) - A^\dagger \right) X \right\|_{\mathcal{X}_\nu} &\le \||A||U|\|_{\mathcal{X}_\nu} \\ &\le \||A|\hat{|U^m|}\|_{\mathcal{X}_\nu} + \||A|\check{|U^m|}\|_{\mathcal{X}_\nu}, \end{aligned}$$

where here and in the sequel, the absolute values must be understood component-wise. We point out that, since A is built as a finite dimensional block $\hat{A}^m$ (acting on $\hat{U}^m$) and a diagonal tail, it follows that $|A|\hat{|U^m|} = |\hat{A}^m|\hat{|U^m|}$ is a finite vector, (it has non zero components only for $k < m$), whereas $|A|\check{|U^m|}$ has non zero components only for $k \ge m$. We provide a bound for both terms separately.

At first, let us compute a bound on $\hat{|U^m|}$. Recalling from (10.22) that $A^\dagger$ is defined so that

$$\widehat{A^\dagger X}^m = D\hat{F}^m(\bar{X})\hat{X}^m,$$

it follows that in computing $\hat{U}^m$ all the linear contributions of X cancel out. Explicitly, using Lemma 10.3.6, a meticulous though straightforward analysis gives

$$\hat{|U^m|} \le (\hat{\alpha}_v^m)\|v\|_\nu + (\hat{\alpha}_w^m)\|w\|_\nu + (\hat{\alpha}_p^m)\|p\|_\nu + (\hat{\alpha}_s^m)\|s\|_\nu.$$

The vectors $\hat{\alpha}_i^m$ can each be seen as an element of $\mathbb{R}^{4m}$, or equivalently of $\mathcal{X}_\nu$ with coefficients equal to 0 for all $k \geq m$. Inserting he previous inequality into $|A||\hat{U}^m|$, we have that

$$\begin{aligned}\||A||\hat{U}^m|\|_{\mathcal{X}_\nu} &\leq \||A|\hat{\alpha}_v^m\|_{\mathcal{X}_\nu} \|v\|_\nu + \||A|\hat{\alpha}_w^m\|_{\mathcal{X}_\nu} \|w\|_\nu + \left\||A|\hat{\alpha}_p^m\right\|_{\mathcal{X}_\nu} \|p\|_\nu + \||A|\hat{\alpha}_s^m\|_{\mathcal{X}_\nu} \|s\|_\nu \\ &\leq \max\left[\||A|\hat{\alpha}_v^m\|_{\mathcal{X}_\nu}, \||A|\hat{\alpha}_w^m\|_{\mathcal{X}_\nu}, \left\||A|\hat{\alpha}_p^m\right\|_{\mathcal{X}_\nu}, \||A|\hat{\alpha}_s^m\|_{\mathcal{X}_\nu}\right] \|X\|_{\mathcal{X}_\nu}, \end{aligned} \tag{10.29}$$

the maximum being taken over terms that can all be evaluated on a computer.

For the tail part (i.e. for modes $k \geq m$), $A^\dagger$ only cancels the diagonal dominant terms, and we have

$$\begin{aligned}|U_k| \leq &\begin{pmatrix} 0 \\ (|b_1(\bar{p}*\bar{w})| * |v|)_k \\ 0 \\ (|r_2 - 2a_2\bar{v} - b_2(\bar{p}*\bar{w})| * |v|)_k \end{pmatrix} + \begin{pmatrix} 0 \\ (|r_1\bar{p} - 2a_1(\bar{p}*\bar{p}*\bar{w}) - b_1(\bar{p}*\bar{v})| * |w|)_k \\ 0 \\ (|b_2(\bar{p}*\bar{v})| * |w|)_k \end{pmatrix} \\ &+ \begin{pmatrix} 0 \\ (|r_1\bar{w} - 2a_1(\bar{p}*\bar{w}*\bar{w}) - b_1(\bar{v}*\bar{w})| * |p|)_k \\ (|2\alpha(\bar{s}\star\bar{p})| * |p|)_k \\ (|b_2(\bar{v}*\bar{w})| * |p|)_k \end{pmatrix} + \begin{pmatrix} |s|_k \\ 0 \\ (|d_{12}(\bar{p}*\bar{p})| \star |s|)_k \\ 0 \end{pmatrix}. \end{aligned}$$

Using Lemma 10.3.3, and the definition of the tail part of A, we get

$$\begin{aligned}\left\||A||\check{U}^m|\right\|_{\mathcal{X}_\nu} &\leq \left(\frac{\|b_1(\bar{p}*\bar{w})\|_\nu}{(\pi m)^2} + \frac{\|r_2 - 2a_2\bar{v} - b_2(\bar{p}*\bar{w})\|_\nu}{d\pi m}\right)\|v\|_\nu \\ &\quad + \left(\frac{\|r_1\bar{p} - 2a_1(\bar{p}*\bar{p}*\bar{w}) - b_1(\bar{p}*\bar{v})\|_\nu}{(\pi m)^2} + \frac{\|b_2(\bar{p}*\bar{v})\|_\nu}{d\pi m}\right)\|w\|_\nu \\ &\quad + \left(\frac{\|r_1\bar{w} - 2a_1(\bar{p}*\bar{w}*\bar{w}) - b_1(\bar{v}*\bar{w})\|_\nu}{(\pi m)^2} + \frac{\|2\alpha(\bar{s}\star\bar{p})\|_\nu}{\pi m} + \frac{\|b_2(\bar{v}*\bar{w})\|_\nu}{d\pi m}\right)\|p\|_\nu \\ &\quad + \left(\frac{1}{\pi m} + \frac{\|d_{12}(\bar{p}*\bar{p})\|_\nu}{\pi m}\right)\|s\|_\nu \\ &\leq \max\left[\left(\frac{\|b_1(\bar{p}*\bar{w})\|_\nu}{(\pi m)^2} + \frac{\|r_2 - 2a_2\bar{v} - b_2(\bar{p}*\bar{w})\|_\nu}{d\pi m}\right),\right. \\ &\qquad \left(\frac{\|r_1\bar{p} - 2a_1(\bar{p}*\bar{p}*\bar{w}) - b_1(\bar{p}*\bar{v})\|_\nu}{(\pi m)^2} + \frac{\|b_2(\bar{p}*\bar{v})\|_\nu}{d\pi m}\right), \\ &\qquad \left(\frac{\|r_1\bar{w} - 2a_1(\bar{p}*\bar{w}*\bar{w}) - b_1(\bar{v}*\bar{w})\|_\nu}{(\pi m)^2} + \frac{\|2d_{12}(\bar{s}\star\bar{p})\|_\nu}{\pi m} + \frac{\|b_2(\bar{v}*\bar{w})\|_\nu}{d\pi m}\right), \\ &\qquad \left.\left(\frac{1}{\pi m} + \frac{\|d_{12}(\bar{p}*\bar{p})\|_\nu}{\pi m}\right)\right]\|X\|_{\mathcal{X}_\nu}. \end{aligned}$$

The sum of the latter estimate and (10.29) provides the bound Z_1 (10.28). □

It is important to remark that all the Φ_k^m functions involved in the definition of $\hat{\alpha}_i^m$, $i \in \{v, w, p, s\}$, take as arguments sequences that only have a finite number of non zero coefficients and that are given in terms of the numerical guess $\bar{X}$. Therefore, all the vectors of coefficients and the bound Z_1 can be rigorously and explicitly computed.

The bound Z_2

In this section we focus on defining a bound Z_2 satisfying (10.8).

Lemma 10.4.4. *Consider the block decomposition of A and the associated coefficients Θ_A, as*

introduced in Section 10.4.3. Define the quantities

$$\alpha_{v,v'} = 2a_2\Theta_A^{(s)}, \quad \alpha_{w,w'} = 2a_1 \left\|\bar{p} * \bar{p}\right\|_\nu \Theta_A^{(w)}, \quad \alpha_{p,p'} = 2a_1 \left\|\bar{w} * \bar{w}\right\|_\nu \Theta_A^{(w)} + 2d_{12} \left\|\bar{s}\right\|_\nu \Theta_A^{(p)},$$
$$\alpha_{v,w'} = b_1 \left\|\bar{p}\right\|_\nu \Theta_A^{(w)} + b_2 \left\|\bar{p}\right\|_\nu \Theta_A^{(s)}, \quad \alpha_{v,p'} = b_1 \left\|\bar{w}\right\|_\nu \Theta_A^{(w)} + d_{12}\Theta_A^{(p)} + b_2 \left\|\bar{w}\right\|_\nu \Theta_A^{(s)},$$
$$\alpha_{w,p'} = \left\|r_1 - 4a_1\bar{w} * \bar{p} - b_1\bar{v}\right\|_\nu \Theta_A^{(w)} + b_2 \left\|\bar{v}\right\|_\nu \Theta_A^{(s)}, \quad \alpha_{p,s'} = 2d_{12} \left\|\bar{p}\right\|_\nu \Theta_A^{(p)},$$

and

$$\alpha_{v,w',p''} = b_1\Theta_A^{(w)} + b_2\Theta_A^{(s)}, \quad \alpha_{w,w',p''} = 4a_1 \left\|\bar{p}\right\|_\nu \Theta_A^{(w)},$$
$$\alpha_{w,p',p''} = 4a_1 \left\|\bar{w}\right\|_\nu \Theta_A^{(w)}, \quad \alpha_{p,p',s''} = 2d_{12}\Theta_A^{(p)},$$

and

$$\alpha_{w,,w',p'',p'''} = 4a_1\Theta_A^{(w)}.$$

Define

$$\begin{aligned} Z_2(r) = &\max\left[\alpha_{v,v'}, \alpha_{w,w'}, \alpha_{p,p'}, \alpha_{v,w'}, \alpha_{v,p'}, \alpha_{w,p'}, \alpha_{p,s'}\right] \\ &+ \frac{1}{2}\max\left[\alpha_{v,w',p''}, \alpha_{w,w',p''}, \alpha_{w,p',p''}, \alpha_{p,p',s''}\right] r \\ &+ \frac{1}{6}\alpha_{w,w',p'',p'''}r^2. \end{aligned} \tag{10.30}$$

Then

$$rZ_2(r) \geq \left\| A\left(DF(X) - DF(\bar{X})\right)\right\|_{\mathcal{X}} \quad \forall\ X \in \mathcal{B}_{\mathcal{X}}(\bar{X}, r).$$

Proof. Consider the expansion:

$$\begin{aligned} A\left(DF(\bar{X} + X') - DF(\bar{X})\right)X = &AD^2F(\bar{X})(X', X) \\ &+ \frac{1}{2}AD^3F(\bar{X})(X', X', X) \\ &+ \frac{1}{6}AD^4F(\bar{X})(X', X', X', X). \end{aligned}$$

We provide bounds for each term on the right hand side, uniform for all $X \in B_{\mathcal{X}_\nu}(0,1)$ and $X' \in B_{\mathcal{X}_\nu}(0,r)$.

For the quadratic term, we have that

$$\begin{aligned} \left\|D^2F^{(v)}(\bar{X})(X', X)\right\|_\nu &= 0, \\ \left\|D^2F^{(w)}(\bar{X})(X', X)\right\|_\nu &\leq b_1 \left\|\bar{w}\right\|_\nu \left(\|v\|_\nu \left\|p'\right\|_\nu + \left\|v'\right\|_\nu \|p\|_\nu\right) \\ &\quad + b_1 \left\|\bar{p}\right\|_\nu \left(\|v\|_\nu \left\|w'\right\|_\nu + \left\|v'\right\|_\nu \|w\|_\nu\right) \\ &\quad + 2a_1 \left\|\bar{w}^2\right\|_\nu \|p\|_\nu \left\|p'\right\|_\nu + 2a_1 \left\|\bar{p}^2\right\|_\nu \|w\|_\nu \left\|w'\right\|_\nu \\ &\quad + \left\|r_1 - 4a_1\bar{w} * \bar{p} - b_1\bar{v}\right\|_\nu \left(\|w\|_\nu \left\|p'\right\|_\nu + \left\|w'\right\|_\nu \|p\|_\nu\right) \\ \left\|D^2F^{(p)}(\bar{X})(X', X)\right\|_\nu &\leq d_{12}\Big(2 \left\|\bar{s}\right\|_\nu \|p\|_\nu \left\|p'\right\|_\nu + 2 \left\|\bar{p}\right\|_\nu \left(\|p\|_\nu \left\|s'\right\|_\nu + \left\|p'\right\|_\nu \|s\|_\nu\right) \\ &\quad + \|v\|_\nu \left\|p'\right\|_\nu + \left\|v'\right\|_\nu \|p\|_\nu\Big), \\ \left\|D^2F^{(s)}(\bar{X})(X', X)\right\|_\nu &\leq 2a_2 \|v\|_\nu \left\|v'\right\|_\nu + b_2\Big(\left\|\bar{p}\right\|_\nu \left(\|v\|_\nu \left\|w'\right\|_\nu + \left\|v'\right\|_\nu \|w\|_\nu\right) \\ &\quad + \left\|\bar{w}\right\|_\nu \left(\|v\|_\nu \left\|p'\right\|_\nu + \left\|v'\right\|_\nu \|p\|_\nu\right) \\ &\quad + \left\|\bar{v}\right\|_\nu \left(\|w\|_\nu \left\|p'\right\|_\nu + \left\|w'\right\|_\nu \|p\|_\nu\right)\Big). \end{aligned}$$

According to (10.24) and by rearrangements of the several terms, it follows

$$\begin{aligned}\left\|AD^2F(\bar{X})(X',X)\right\|_{\mathcal{X}_\nu} &\leq \alpha_{v,v'}\|v\|_\nu\|v'\|_\nu + \alpha_{w,w'}\|w\|_\nu\|w'\|_\nu + \alpha_{p,p'}\|p\|_\nu\|p'\|_\nu \\ &\quad + \alpha_{v,w'}\left(\|v\|_\nu\|w'\|_\nu + \|v'\|_\nu\|w\|_\nu\right) + \alpha_{v,p'}\left(\|v\|_\nu\|p'\|_\nu + \|v'\|_\nu\|p\|_\nu\right) \\ &\quad + \alpha_{w,p'}\left(\|w\|_\nu\|p'\|_\nu + \|w'\|_\nu\|p\|_\nu\right) + \alpha_{p,s'}\left(\|p\|_\nu\|s'\|_\nu + \|p'\|_\nu\|s\|_\nu\right) \\ &\leq \max\left[\alpha_{v,v'},\alpha_{w,w'},\alpha_{p,p'},\alpha_{v,w'},\alpha_{v,p'},\alpha_{w,p'},\alpha_{p,s'}\right] \\ &\quad \times\left(\|v\|_\nu + \|w\|_\nu + \|p\|_\nu + \|s\|_\nu\right)\left(\|v'\|_\nu + \|w'\|_\nu + \|p'\|_\nu + \|s'\|_\nu\right) \\ &\leq \max\left[\alpha_{v,v'},\alpha_{w,w'},\alpha_{p,p'},\alpha_{v,w'},\alpha_{v,p'},\alpha_{w,p'},\alpha_{p,s'}\right] r\|X\|_{\mathcal{X}_\nu}.\end{aligned}$$

The same procedure applied to the higher order derivative gives

$$\left\|AD^3F(\bar{X})(X',X',X)\right\|_{\mathcal{X}_\nu} \leq \max\left[\alpha_{v,w',p''},\alpha_{w,w',p''},\alpha_{w,p',p''},\alpha_{p,p',s''}\right] r^2\|X\|_{\mathcal{X}_\nu},$$

and

$$\left\|AD^4F(\bar{X})(X',X',X',X)\right\|_{\mathcal{X}_\nu} \leq \alpha_{w,w',p'',p'''}r^3\|X\|_{\mathcal{X}_\nu}.$$

Combining the above estimates, it follows that

$$rZ_2(r) \geq \left\|A\left(DF(\bar{X}+X') - DF(\bar{X})\right)X\right\|_{\mathcal{X}_\nu}, \quad \forall\ X\in B_{\mathcal{X}_\nu}(0,1),\ \forall\ X'\in B_{\mathcal{X}_\nu}(0,r).$$

□

Notice that the computation of $\Theta_A^{(i)}$ requires the computation of $\left\|\left\|A^{(i,j)}\right\|\right\|_\nu$. Contrarily to the situation in Section 10.4.3, the tail of $A^{(i,j)}$ is not zero, in case $i=j$. However, since it has a diagonal structure, we can still explicitly compute the operator norm of each block of A. For instance

$$\begin{aligned}\left\|\left\|A^{(v,v)}\right\|\right\|_\nu &= \sup_{j\geq 0}\frac{1}{\nu^j}\sum_{k\geq 0}|A^{(v,v)}(k,j)|\nu^k \\ &= \max\left[\max_{0\leq j<m}\frac{1}{\nu^j}\sum_{0\leq k<m}|A^{(v,v)}(k,j)|\nu^k, \sup_{j\geq m}\frac{1}{\nu^j}|-(\pi j)^{-1}|\nu^j\right] \\ &= \max\left[\max_{0\leq j<m}\frac{1}{\nu^j}\sum_{0\leq k<m}|A^{(v,v)}(k,j)|\nu^k, \frac{1}{\pi m}\right].\end{aligned}$$

10.4.4 Existence of steady States: the radii polynomial

We now collect all the ingredients required to prove the existence of steady states into a unique proposition.

Proposition 10.4.5. *For $\nu>1$, let the space $\mathcal{X}_\nu = \left(\ell^1_\nu(\mathbb{R})\right)^4$ be endowed with the norm* (10.16) *and let F, $\bar{X}$, A, $A^\dagger$ be as defined in section 10.4.1 and section 10.4.2. Let the bounds Y, Z_0, Z_1 and Z_2 be defined in* (10.23), (10.27), (10.28) *and* (10.30) *respectively, and rigorously computed.*

i) If there exists $r>0$ such that

$$P(r) = Z_2(r)r^2 - (1-(Z_0+Z_1))\,r + Y < 0,$$

then there exists a unique zero of F in $\mathcal{B}_{\mathcal{X}_\nu}(\bar{X},r)$.

ii) Let the functions $\bar{v}(x)$, $\bar{w}(x)$ *and* $\bar{u}(x)$ *be*

$$\bar{w}(x) = \bar{w}_0 + 2\sum_{k=1}^{m-1} \bar{w}_k \cos(k\pi x), \quad \bar{v}(x) = \bar{v}_0 + 2\sum_{k=1}^{m-1} \bar{v}_k \cos(k\pi x).$$

$$\bar{u}(x) = \bar{w}(w)\bar{p}(x) = (\bar{p} * \bar{w})_0 + 2\sum_{k=1}^{2m-2} (\bar{p} * \bar{w})_k \cos(k\pi x),$$

If $P(r) < 0$ *and* $\inf_{x\in[0,1]} \bar{w}(x) - r > 0$ *and* $\inf_{x\in[0,1]} \bar{v}(x) - r > 0$*, then there exists of a smooth solution* $(u(x), v(x))$ *to* (10.4) *so that*

$$|v(x) - \bar{v}(x)| < r, \quad |u(x) - \bar{u}(x)| < (\|\bar{w}\|_\nu + \|\bar{p}\|_\nu)\, r + \frac{r^2}{4}, \qquad \forall x \in [0,1].$$

Proof. i) The definition of the bounds implies that the assumptions (10.5)-(10.8) are satisfied. Theorem 10.2.1 then yields the existence and uniqueness of a zero for F in $\mathcal{B}_{\mathcal{X}_\nu}(\bar{X}, r)$. The injectivity of A follows for free from the fact that $P(r) < 0$. Indeed it is necessary that $Z_0 < 1$ which means $\left\|I - AA^\dagger\right\|_{\mathcal{X}} < 1$. Note that the tail parts of A and $A^\dagger$ are analytically defined in a way that $1 > \left\|I - AA^\dagger\right\|_{\mathcal{X}} = \left\|\hat{I}^m - \hat{A}^m \hat{A}^{\dagger^m}\right\|_{\mathcal{X}}$. The latter implies that both $\hat{A}^m$ and $\hat{A}^{\dagger^m}$ are invertible, and thus A is injective because its diagonal tail is made of non-zero coefficients.

ii) Let $X = (v, w, p, s)$ be the unique zero of F in $\mathcal{B}_{\mathcal{X}_\nu}(\bar{X}, r)$. Thus $\|v - \bar{v}\|_\nu \leq r$ and $\|w - \bar{w}\|_\nu < r$. Since for any $a \in \ell^1_\nu$ we have that $\|a\|_\nu \geq \sup_x |a(x)|$, it follows that $v(x) \geq \inf_x \bar{v}(x) - r > 0$. The same holds for $w(x)$. For Lemma 10.4.1 it follows the existence of a smooth solution to (10.4). The error bound between u and $\bar{u}$ is proven in Section 10.6.1. □

10.5 Results about the existence of steady states

In this section we present the computer-assisted proof of existence of steady states solutions stated in Theorem 10.1.1.

Each solution that is represented on Figure 10.2 was validated using the procedure described at the end of Section 10.2. In particular, we computed each solution numerically, implemented the bounds described in Section 10.4.3, and then *successfully* applied Proposition 10.4.5 to validate the numerical solution. By successfully we mean that we found a positive r such that $P(r) < 0$ and checked that $\inf_{x\in[0,1]} \bar{w}(x) - r > 0$ and $\inf_{x\in[0,1]} \bar{v}(x) - r > 0$ (with the notation of Proposition 10.4.5). The numerical data as well as the *Matlab* codes to perform the proofs and some documentation are available at [72]. To make the computation of the bounds rigorous by controlling round-off errors, the interval arithmetic package *Intlab* [190] has been used. The computations presented here have been run on a laptop with a processor Intel Core i7 (2.50Ghz) and 8GB of RAM.

Proof of Theorem 10.1.1. In the script `script_proof_branch_steadystates.m` fix the values of the parameters $r_1 = 5$, $r_2 = 2$, $a_1 = 3$, $a_2 = 3$, $b_1 = 1$, $b_2 = 1$, $d_{12} = 3$. The parameter $d_1 = d_2 = d$ is intended as the bifurcation parameter. Choose a value for the finite dimensional projection m and a value for the norm weight $\nu > 1$. Also select a branch of solutions (for the names of the several branches we refer to the documentation and the `readme` file). The script loads the numerical data, computes the required bounds and verifies the existence of an interval $\mathcal{I} = (r_1, r_2)$ such that $P(r) < 0$ for any $r \in \mathcal{I}$. If $\mathcal{I}$ is not empty then the conditions $\inf_{x\in[0,1]} \bar{w}(x) - r > 0$ and $\inf_{x\in[0,1]} \bar{v}(x) - r > 0$ are checked. In case of successful computation, Proposition 10.4.5 implies the existence of the solutions. The values for m and ν that allow the rigorous computation of all the branches depicted in the Figure 10.2 are available in the documentation.

The script `script_proof_steadystate_and_instability.m` concerns the existence of steady states for a fixed value of d. It is used to prove the existence of 13 solutions at values $d = 0.005$. Figure 10.4 shows the numerical data for the 13 steady states solutions. In Table 10.1 we detail the values for m and ν used in the proof and the resulting validation radius r (the script also aims at computing unstable eigenvalues, see Section 10.7).

Solution (see Figure 10.4)	m used for the proof	ν used for the proof	Validation radius
(a)	500	1.06	2.5968×10^{-11}
(b)	500	1.06	9.8961×10^{-12}
(c)	500	1.06	7.2076×10^{-11}
(d)	500	1.06	7.8228×10^{-11}
(e)	500	1.06	5.7165×10^{-12}
(f)	500	1.06	1.0104×10^{-10}
(g)	700	1.055	7.7146×10^{-11}
(h)	500	1.06	2.9001×10^{-12}
(i)	500	1.06	5.0578×10^{-12}
(j)	500	1.06	6.4651×10^{-12}
(k)	700	1.055	8.1462×10^{-11}
(l)	500	1.06	1.6680×10^{-11}
(m)	500	1.06	4.2505×10^{-11}

Table 10.1 – For each solution displayed in Figure 10.4, we give the dimension m that was used for the finite dimensional projection, the weight ν that was chosen for the space $\mathcal{X}_\nu$, and a validation radius r for which the proof is successful, with those parameters m and ν.

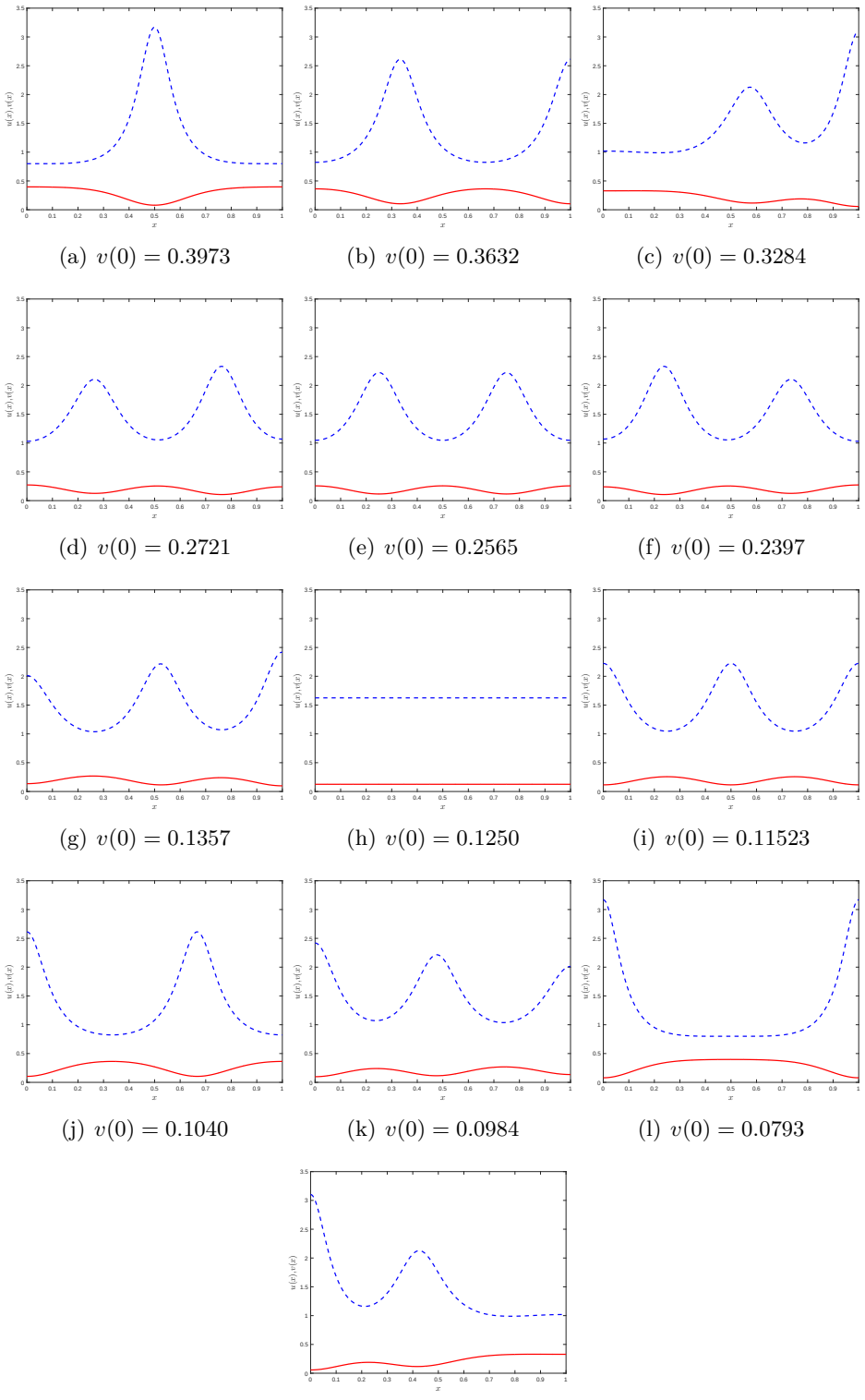
10.6 Framework for the instability of steady states

In this section, we focus on the stability of the steady states whose existence we proved in the Sections 10.4 to 10.5. More precisely, we consider the cross-diffusion system (10.2) in the triangular case $d_{21} = 0$ and with space dimension one ($\Omega = (0,1)$),

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial^2}{\partial x^2}\left((d_1 + d_{12}v)u\right) + (r_1 - a_1 u - b_1 v)u, & \text{on } \mathbb{R}_+ \times (0,1), \\ \dfrac{\partial v}{\partial t} = d_2 \dfrac{\partial^2 v}{\partial x^2} + (r_2 - b_2 u - a_2 v)v, & \text{on } \mathbb{R}_+ \times (0,1), \\ \dfrac{\partial u}{\partial x}(t,0) = 0 = \dfrac{\partial u}{\partial x}(t,1), & \text{on } \mathbb{R}_+, \\ \dfrac{\partial v}{\partial x}(t,0) = 0 = \dfrac{\partial v}{\partial x}(t,1), & \text{on } \mathbb{R}_+. \end{cases} \tag{10.31}$$

Let $(u,v) = (u(x), v(x))$ be a steady state of (10.31). The linearization of (10.31), around the steady state (u,v) yields the eigenvalue problem

$$\begin{cases} d_1\xi'' + d_{12}\left(v''\xi + 2v'\xi' + v\xi'' + \eta''u + 2\eta'u' + \eta u''\right) + r_1\xi - 2a_1u\xi - b_1(v\xi + \eta u) = \lambda\xi, & \text{on } (0,1), \\ d_2\eta'' + r_2\eta - b_2(u\eta + \xi v) - 2a_2v\eta = \lambda\eta, & \text{on } (0,1), \\ \xi'(0) = \xi'(1) = 0, \\ \eta'(0) = \eta'(1) = 0, \end{cases} \tag{10.32}$$



(a) $v(0) = 0.3973$ (b) $v(0) = 0.3632$ (c) $v(0) = 0.3284$

(d) $v(0) = 0.2721$ (e) $v(0) = 0.2565$ (f) $v(0) = 0.2397$

(g) $v(0) = 0.1357$ (h) $v(0) = 0.1250$ (i) $v(0) = 0.11523$

(j) $v(0) = 0.1040$ (k) $v(0) = 0.0984$ (l) $v(0) = 0.0793$

(m) $v(0) = 0.0569$

Figure 10.4 – The 13 solutions announced in Theorem 10.1.1 for $d = 0.005$. They can be replaced on the bifurcation diagram using the value of $v(0)$. u is represented in dashed blue, and v in red. We give additional information about the proof for each of these solutions in Table 10.1, and discuss their stability in Section 10.7.

where the functions $(\xi, \eta) = (\xi(x), \eta(x))$ form the eigenfunction and λ is the eigenvalue. As announced in the introduction, we aim at proving that most of the steady states obtained

in Section 10.5 are unstable, by showing that the eigenproblem (10.32) admits an unstable eigenvalue, i.e. there exists a solution $((\xi,\eta),\lambda)$ of (10.32) such that $\Re(\lambda) > 0$. The approach is similar to the one used in Section 10.4: we first reformulate (10.32) into an equivalent problem more amenable to validated numerics, and then use Theorem 10.2.1 to prove the existence of an unstable eigenvalue. For this, we again follow the algorithm outlined at the end of Section 10.2.

10.6.1 Proof of instability: the function F

As we did for the steady states, we look for the eigenfunctions (ξ,η) as cosine series. We point out that the steady state (u,v) in (10.32) are non constant functions which have been obtained in Sections 10.4 and 10.5. If we directly expand (10.32) on the Fourier basis, the dominant terms would not be diagonal. We take care of this issue by transforming (10.32) into an equivalent *generalized eigenvalue problem* which has an autonomous second order terms.

The two equations in (10.32) can be rewritten as

$$M_1 \begin{pmatrix} \xi'' \\ \eta'' \end{pmatrix} + M_2 \begin{pmatrix} \xi' \\ \eta' \end{pmatrix} + M_3 \begin{pmatrix} \xi \\ \eta \end{pmatrix} = \lambda \begin{pmatrix} \xi \\ \eta \end{pmatrix}, \tag{10.33}$$

where

$$M_1 = \begin{pmatrix} d_1 + d_{12}v & d_{12}u \\ 0 & d_2 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 2d_{12}v' & 2d_{12}u' \\ 0 & 0 \end{pmatrix},$$

$$M_3 = \begin{pmatrix} d_{12}v'' + r_1 - 2a_1u - b_1v & d_{12}u'' - b_1u \\ -b_2v & r_2 - b_2u - 2a_2v \end{pmatrix}.$$

Introducing $p = \frac{1}{d_1 + d_{12}v}$ as in Section 10.4, and knowing that $p(x) > 0$ for any $x \in [0,1]$, we can express the inverse of M_1 as:

$$M_1^{-1} = \begin{pmatrix} \frac{1}{d_1+d_{12}v} & -\frac{d_{12}u}{d_2(d_1+d_{12}v)} \\ 0 & \frac{1}{d_2} \end{pmatrix} = \begin{pmatrix} p & -\frac{d_{12}up}{d_2} \\ 0 & \frac{1}{d_2} \end{pmatrix}.$$

We multiply (10.33) by M_1^{-1} and obtain the following equivalent formulation for (10.32):

$$\begin{cases} \xi'' + c_1\xi' + c_2\eta' + c_3\xi + c_4\eta + c_5\lambda\xi + c_6\lambda\eta = 0, & \text{on } (0,1), \\ \eta'' + c_7\xi + c_8\eta + c_9\lambda\eta = 0, & \text{on } (0,1), \\ \xi'(0) = \xi'(1) = 0, \\ \eta'(0) = \eta'(1) = 0, \end{cases} \tag{10.34}$$

where the functions $(c_j)_{1\le j\le 9}$ depend on the steady state (u,v) (and on the parameters of the cross-diffusion system), and are given by

$$c_1 = 2d_{12}pv', \quad c_2 = 2d_{12}pu', \quad c_3 = (r_1 - 2a_1u - b_1v + d_{12}v'')p + \frac{d_{12}b_2}{d_2}uvp,$$

$$c_4 = (d_{12}u'' - b_1u)p - \frac{d_{12}}{d}up(r_2 - b_2u - 2a_2v), \quad c_5 = -p, \quad c_6 = \frac{d_{12}}{d_2}up,$$

$$c_7 = -\frac{b_2}{d_2}v, \quad c_8 = \frac{1}{d_2}(r_2 - b_2u - 2a_2v), \quad c_9 = -\frac{1}{d_2}. \tag{10.35}$$

The algebraic system in Fourier space and the $F = 0$ formulation

Expanding the eigenfunctions in cosine series

$$\xi(x) = \xi_0 + 2\sum_{k\geq 1} \xi_k \cos(k\pi x), \quad \eta(x) = \eta_0 + 2\sum_{k\geq 1} \eta_k \cos(k\pi x),$$

and inserting these expansions in (10.34), we end up with the following equations, for all $k \in \mathbb{N}$,

$$\begin{cases} -(\pi k)^2 \xi_k - (c_1 \bullet K\xi)_k - (c_2 \bullet K\eta)_k + (c_3 * \xi)_k + (c_4 * \eta)_k + \lambda(c_5 * \xi)_k + \lambda(c_6 * \eta)_k = 0, \\ -(\pi k)^2 \eta_k + (c_7 * \xi)_k + (c_8 * \eta)_k + \lambda(c_9 * \eta)_k = 0. \end{cases} \tag{10.36}$$

Again, we identify the functions ξ, η and c_j with their sequence of Fourier coefficients.

Remark 10.6.1. ***Reset of some notations.*** *To maintain the same notations as in Theorem 10.2.1, we are going to redefine the appropriate $\mathcal{X}$, X, F, $\bar{X}$, A and $A^\dagger$ corresponding to the eigenproblem* (10.34)*. Henceforth, we forget about the definition of this symbols that was given in Section 10.4, and give new ones in the sequel.*

For $\gamma > 1$, we define $\mathcal{X}_\gamma = \left(\ell^1_\gamma\right)^2 \times \mathbb{C}$ and denote by $X = (\xi, \eta, \lambda)$ any element in $\mathcal{X}_\gamma$. We point out that in Section 10.4 we could restrict ourselves to real sequences because we were only looking for real solutions, but this is no longer the case here since we may encounter complex conjugate eigenvalues and eigenvectors. We endow $\mathcal{X}_\gamma$ with the norm

$$\|X\|_{\mathcal{X}_\gamma} = \|\xi\|_\gamma + \|\eta\|_\gamma + |\lambda|\,,$$

which makes it a Banach space. We then fix an index $k_0 \in \mathbb{N}$ and define the function $F = (F^{(\xi)}, F^{(\eta)}, F^{(\lambda)})$ acting on $X = (\xi, \eta, \lambda) \in \mathcal{X}_\gamma$ by

$$\begin{aligned} F_k^{(\xi)}(X) &= -(\pi k)^2 \xi_k - (c_1 \bullet K\xi)_k - (c_2 \bullet K\eta)_k \\ &\qquad + (c_3 * \xi)_k + (c_4 * \eta)_k + \lambda(c_5 * \xi)_k + \lambda(c_6 * \eta)_k, \qquad \forall\ k \in \mathbb{N}, \\ F_k^{(\eta)}(X) &= -(\pi k)^2 \eta_k + (c_7 * \xi)_k + (c_8 * \eta)_k + \lambda(c_9 * \eta)_k, \qquad \forall\ k \in \mathbb{N}, \\ F^{(\lambda)}(X) &= \xi_{k_0} - 1. \end{aligned}$$

Notice that the only difference between $F(X) = 0$ and system (10.36) is the equation $\xi_{k_0} = 1$. The role of this additional constraint is to normalise the eigenfunction and hence to isolate the potential solutions of F. Indeed, we cannot hope to successfully use Theorem 10.2.1, which is based on a contraction argument, if the zeros of F are not isolated. We point out that many different conditions could have been added to isolate the solution, and that this specific choice is rather arbitrary.

We now state the precise link between F and our stability problem.

Lemma 10.6.2. *Assume that (u, v) is a positive stationary solution of* (10.31) *and that there exists $\gamma > 1$ such that the Fourier coefficients of the functions $(c_j)_{1\leq j\leq 9}$, defined in* (10.35)*, belong to $\ell^1_\gamma(\mathbb{R})$. Fix $k_0 \in \mathbb{N}$ and suppose that there exists $X = (\xi, \eta, \lambda) \in \mathcal{X}_\gamma$, with $\Re(\lambda) > 0$, such that $F(X) = 0$. Then the steady state (u, v) is linearly unstable.*

Proof. As for Lemma 10.4.1, the proof just consists in checking that the regularity (i.e. the fact that the Fourier coefficients belongs to ℓ^1_γ) of the solution X and of the data $(c_j)_{1\leq j\leq 9}$ allows to rigorously backtrack the manipulations made to obtain F from the eigenproblem (10.32). □

We point out that the assumption that u and v are positive, is in fact only needed here to ensure that p is well defined. Concerning the assumption on the functions $(c_j)_{1\leq j\leq 9}$, the method developed in Sections 10.4 to 10.5 naturally provides us with steady states (u, v) for which the Fourier coefficients of u, v (and p) belong to ℓ^1_ν, for some $\nu > 1$. However, since some of the c_j involve derivatives of u and v, we can only get that their Fourier coefficients belong to ℓ^1_γ for $\gamma < \nu$. We give more details and explicit estimates right below.

About the functions c_j

The function F depends on the Fourier coefficients of $(c_j)_{1\leq j\leq 9}$, which themselves depend on the steady state (u, v). The method described in Sections 10.4 to 10.5 (in particular Proposition 10.4.5) provides us with an approximate steady state, in the form of Fourier sequences $(\bar{v}, \bar{w}, \bar{p}, \bar{s})$, together with a validation radius r_ν which gives an upper bound of the distance (in the ℓ^1_ν norm) between the approximate steady state and the genuine one.

Remark 10.6.3. *From now on, we denote r_ν the validation radius obtained in the computation of the steady states. This validation radius was simply denoted by r in Section 10.4 and 10.5, but this new notation should avoid possible confusions with the validation radius that we are going to consider for the eigenvalue problem.*

More precisely, the steady state (in the (v, w, p, s) coordinates) is proved to exist in the form

$$v = \bar{v} + \varepsilon_v, \quad w = \bar{w} + \varepsilon_w, \quad p = \bar{p} + \varepsilon_p, \quad s = \bar{s} + \varepsilon_s,$$

where the $(\bar{v}, \bar{w}, \bar{p}, \bar{s})$ are finite Fourier sequences that we have explicitly on our computer, and $\|\varepsilon_v\|_\nu + \|\varepsilon_w\|_\nu + \|\varepsilon_p\|_\nu + \|\varepsilon_s\|_\nu \leq r_\nu$, where r_ν is the radius provided by the proof.

Therefore, we can also represent each c_j as

$$c_j = \bar{c}_j + \varepsilon_j,$$

where $\bar{c}_j$ is a finite sequence of Fourier coefficients and $\|\varepsilon_j\|_\gamma \leq \epsilon_j(\gamma)$. In this section, we provide formulas for the finite sequences $\bar{c}_j$ and the upper bounds $\epsilon_j(\gamma)$ on the distance (in ℓ^1_γ) between $\bar{c}_j$ and c_j .

The first step is to provide an enclosure for u in the form $u = \bar{u} + \varepsilon_u$. Remembering that $u(x) = p(x)w(x)$, and hence $u = p * w$, we have

$$u = \bar{p} * \bar{w} + \bar{p} * \varepsilon_w + \varepsilon_p * \bar{w} + \varepsilon_p * \varepsilon_w.$$

Therefore, we can define $\bar{u} = \bar{p} * \bar{w}$, and $\varepsilon_u = \bar{p} * \varepsilon_w + \varepsilon_p * \bar{w} + \varepsilon_p * \varepsilon_w$. By Lemma 10.3.3, a bound for the norm of ε_u is given as $\|\varepsilon_u\|_\nu \leq (\|\bar{p}\|_\nu + \|\bar{w}\|_\nu) r_\nu + \frac{r_\nu^2}{4}$, where we used that $\|\varepsilon_p\|_\nu \|\varepsilon_w\|_\nu \leq \frac{1}{4}(\|\varepsilon_p\|_\nu + \|\varepsilon_w\|_\nu)^2$. For further use, we define $\epsilon_u = (\|\bar{p}\|_\nu + \|\bar{w}\|_\nu) r_\nu + \frac{r_\nu^2}{4}$.

The second step is to derive estimates on derivatives.

Definition 10.6.4. *Denote by K the (unbounded) linear operator such that, for all z in ℓ^1_ν*

$$(Kz)_k = \pi k z_k, \quad \forall k \geq 0.$$

Up to a change of sign (depending on whether we consider the cosine or sine expansion), Kz is nothing but the sequence of Fourier coefficients of the derivative of the function $z(x)$.

Lemma 10.6.5. *Let $1 < \gamma < \nu$ and $z \in \ell^1_\nu$. Then*

$$\|Kz\|_\gamma \leq \Upsilon^1_{\gamma,\nu} \|z\|_\nu, \quad \left\|K^2 z\right\|_\gamma \leq \Upsilon^2_{\gamma,\nu} \|z\|_\nu,$$

where

$$\Upsilon^1_{\gamma,\nu} = \begin{cases} \dfrac{\gamma}{\nu}, & \text{if } \gamma < e^{-1}\nu \\ \dfrac{e^{-1}}{\ln \frac{\nu}{\gamma}}, & \text{otherwise,} \end{cases} \quad \text{and} \quad \Upsilon^2_{\gamma,\nu} = \begin{cases} \dfrac{\gamma}{\nu}, & \text{if } \gamma < e^{-2}\nu \\ \left(\dfrac{2e^{-1}}{\ln \frac{\nu}{\gamma}}\right)^2, & \text{otherwise.} \end{cases}$$

Proof. We estimate

$$\begin{aligned}\|Kz\|_\gamma &= 2\sum_{k\geq 1} k\,|z_k|\,\gamma^k \\ &= 2\sum_{k\geq 1} k\left(\frac{\gamma}{\nu}\right)^k |z_k|\,\nu^k \\ &\leq \|z\|_\nu \sup_{k\geq 1} k\left(\frac{\gamma}{\nu}\right)^k.\end{aligned}$$

The constant $\Upsilon^1_{\gamma,\nu}$ is the maximum (on $[1,+\infty)$) of the function $k\mapsto k\left(\frac{\gamma}{\nu}\right)^k$. Similarly, $\Upsilon^2_{\gamma,\nu}$ is the maximum (on $[1,+\infty)$) of the function $k\mapsto k^2\left(\frac{\gamma}{\nu}\right)^k$. □

For $v=\bar v+\varepsilon_v$ we have $Kv=K\bar v+K\varepsilon_v$. The sequence $K\bar v$ can be rigorously computed, being finite dimensional, while by the Lemma 10.6.5 $\|K\varepsilon_v\|_\gamma \leq \Upsilon^1_{\gamma,\nu} r_\nu$. The same argument used to derive $\bar u$ and ϵ_u and the application of the Lemma 10.6.5 when requested, provide the following expressions for $\bar c_j$, and $\epsilon_j(\gamma)$:

$$\bar c_1=-2d_{12}(K\bar v\star\bar p),\quad \bar c_2=-2d_{12}(K\bar u\star\bar p),\quad \bar c_3=((r_1-2a_1\bar u-b_1\bar v+d_{12}K\bar s)*\bar p)+\frac{d_{12}b_2}{d_2}(\bar u*\bar v*\bar p),$$

$$\bar c_4=((-d_{12}K^2\bar u-b_1\bar u)*\bar p)-\frac{d_{12}}{d}(\bar u*\bar p*(r_2-b_2\bar u-2a_2\bar v)),\quad \bar c_5=-\bar p,\quad \bar c_6=\frac{d_{12}}{d_2}(\bar u*\bar p),$$

$$\bar c_7=-\frac{b_2}{d_2}\bar v,\quad \bar c_8=\frac{1}{d_2}(r_2-b_2\bar u-2a_2\bar v),\quad \bar c_9=-\frac{1}{d_2},$$

together with

$$\epsilon_1(\gamma)=2d_{12}(\|K\bar v\|_\gamma\, r_\nu+\|\bar p\|_\gamma\,\Upsilon^1_{\gamma,\nu}r_\nu+\Upsilon^1_{\gamma,\nu}r_\nu^2),\quad \epsilon_2(\gamma)=2d_{12}(\|K\bar u\|_\gamma\, r_\nu+\|\bar p\|_\gamma\,\Upsilon^1_{\gamma,\nu}\epsilon_u+\Upsilon^1_{\gamma,\nu}\epsilon_u r_\nu),$$

$$\begin{aligned}\epsilon_3(\gamma)=&\|r_1-2a_1\bar u-b_1\bar v+d_{12}K\bar s\|_\gamma\, r_\nu+(2a_1\epsilon_u+b_1r_\nu+d_{12}\Upsilon^1_{\gamma,\nu}r_\nu)\,\|\bar p\|_\gamma\\ &+(2a_1\epsilon_u+b_1r_\nu+d_{12}\Upsilon^1_{\gamma,\nu}r_\nu)r_\nu\\ &+\frac{d_{12}b_2}{d_2}(\|\bar u*\bar v\|_\gamma\, r_\nu+\|\bar u*\bar p\|_\gamma\, r_\nu+\|\bar v*\bar p\|_\gamma\,\epsilon_u+\|\bar u\|_\gamma\, r_\nu^2+\|\bar v\|_\gamma\,\epsilon_u r_\nu+\|\bar p\|_\gamma\,\epsilon_u r_\nu+\epsilon_u r_\nu^2),\end{aligned}$$

$$\begin{aligned}\epsilon_4(\gamma)=&\left\|-d_{12}K^2\bar u-b_1\bar u\right\|_\gamma\, r_\nu+(d_{12}\Upsilon^2_{\gamma,\nu}\epsilon_u+b_1\epsilon_u)\,\|\bar p\|_\gamma+(d_{12}\Upsilon^2_{\gamma,\nu}\epsilon_u+b_1\epsilon_u)r_\nu\\ &+\frac{d_{12}}{d_2}\Big(\|\bar u*\bar p\|_\gamma\,(b_2\epsilon_u+2a_2r_\nu)+\|\bar u*(r_2-b_2\bar u-2a_2\bar v)\|_\gamma\, r_\nu\\ &\qquad+\|\bar p*(r_2-b_2\bar u-2a_2\bar v)\|_\gamma\,\epsilon_u+\epsilon_u r_\nu(b_2\epsilon_u+2a_2r_\nu)\Big),\end{aligned}$$

$$\epsilon_5=r_\nu,\quad \epsilon_6(\gamma)=\frac{d_{12}}{d_2}(\|\bar u\|_\gamma\, r_\nu+\|\bar u*\bar p\|_\gamma\,\epsilon_u+\epsilon_u r_\nu),\quad \epsilon_7=\frac{b_2}{d_2}r_\nu,\quad \epsilon_8=\frac{1}{d_2}(b_2\epsilon_u+2a_2r_\nu),\quad \epsilon_9=0.$$

10.6.2 Proof of instability: the operators A and $A^\dagger$

We now introduce the approximate solution and linear operators needed to apply Theorem 10.2.1 in the context of the eigenproblem (10.32).

Compared to the situation of Section 10.4.2, we have here an additional difficulty due to the fact that the function F depends on the coefficients $(c_j)_{1\leq j\leq 9}$, which (as detailed above) are only known *up to an error bound*. This motivates the splitting of the function F into two parts,

one containing the known terms $\bar{c}_j$ and the other one containing the remainder terms ε_j. More precisely, we define $\bar{F}$ by

$$\begin{aligned}
\bar{F}_k^{(\xi)}(X) &= -(\pi k)^2\xi_k - (\bar{c}_1 \bullet K\xi)_k - (\bar{c}_2 \bullet K\eta)_k \\
&\quad + (\bar{c}_3 * \xi)_k + (\bar{c}_4 * \eta)_k + \lambda(\bar{c}_5 * \xi)_k + \lambda(\bar{c}_6 * \eta)_k, && \forall\ k \in \mathbb{N}, \\
\bar{F}_k^{(\eta)}(X) &= -(\pi k)^2\eta_k + (\bar{c}_7 * \xi)_k + (\bar{c}_8 * \eta)_k + \lambda(\bar{c}_9 * \eta)_k, && \forall\ k \in \mathbb{N}, \\
\bar{F}^{(\lambda)}(X) &= \xi_{k_0} - 1, && (10.37)
\end{aligned}$$

and $\mathcal{E}_F$ as

$$\begin{aligned}
\left(\mathcal{E}_F\right)_k^{(\xi)}(X) &= -(\varepsilon_1 \bullet K\xi)_k - (\varepsilon_2 \bullet K\eta)_k \\
&\quad + (\varepsilon_3 * \xi)_k + (\varepsilon_4 * \eta)_k + \lambda(\varepsilon_5 * \xi)_k + \lambda(\varepsilon_6 * \eta)_k, && \forall\ k \in \mathbb{N}, \\
\left(\mathcal{E}_F\right)_k^{(\eta)}(X) &= (\varepsilon_7 * \xi)_k + (\varepsilon_8 * \eta)_k + \lambda(\varepsilon_9 * \eta)_k, && \forall\ k \in \mathbb{N}, \\
\left(\mathcal{E}_F\right)^{(\lambda)}(X) &= 0, && (10.38)
\end{aligned}$$

so that $F = \bar{F} + \mathcal{E}_F$.

Then, extending again the notations introduced in Definition 10.3.5, we denote

$$\hat{X}^n = (\hat{\xi}^n, \hat{\eta}^n, \lambda).$$

Notice that the truncation parameter n does not need to be (and in practice is not) the same as the truncation parameter m used for the steady states. However, we require $n > k_0$, so that the isolating condition that we imposed is incorporated in the finite dimensional projection. We also define

$$\hat{F}^n = \left(\left(\bar{F}_k^{(\xi)}\right)_{0\le k<n}, \left(\bar{F}_k^{(\eta)}\right)_{0\le k<n}, \bar{F}^{(\lambda)}\right).$$

We consider $\hat{F}^n$ as acting on truncated sequences $\hat{X}^n$ only, so that we can see it as a function mapping $\mathbb{C}^{2n+1}$ to itself. Therefore, finding $\hat{X}^n$ such that $\hat{F}^n(\hat{X}^n) = 0$ is a finite dimensional problem that can be solved numerically. Notice that crucially, $\hat{F}$ is a finite dimensional projection of $\bar{F}$ rather than of F, so it only depends on coefficients that are known explicitly.

We now assume that we have computed numerically a zero of $\hat{F}^n$, and denote it $\bar{X}$. The next step is to define $A^\dagger$ and A. Again, we are going to take for $A^\dagger$ an approximation of $DF(\bar{X})$, with a diagonal tail. More precisely, we define $A^\dagger$ (acting on $X = (\xi, \eta, \lambda) \in \mathcal{X}_\gamma$), as

$$\widehat{A^\dagger X}^n = D\hat{F}^n(\bar{X})\hat{X}^n,$$

and

$$\left(A^\dagger X\right)_k = \left(-(\pi k)^2\xi_k, -(\pi k)^2\eta_k\right), \quad \forall\ k \ge n.$$

Then, we consider $\hat{A}^n$ a numerically computed inverse of $D\hat{F}^n(\bar{X})$ and define A (acting on $X = (\xi, \eta, \lambda) \in \mathcal{X}_\gamma$), as

$$\widehat{AX}^n = \hat{A}^n\hat{X}^n,$$

and

$$(AX)_k = \left(-(\pi k)^{-2}\xi_k, -(\pi k)^{-2}\eta_k\right), \quad \forall\ k \ge n.$$

Remark 10.6.6. *As in Remark* (10.4.2)*, we point out the diagonal dominant behaviour of the derivative $DF(\bar{X})$ is the result of some preliminary manipulations done on the differential system, in this case the multiplication of the eigenproblem* (10.32) *by M_1^{-1}.*

The definition of the tail part of A and the fact that ℓ^1_γ is an algebra for both convolution products $*$ and $\bullet$ (see Lemma 10.3.3) ensure that AF does map $\mathcal{X}_\gamma$ into itself as requested to apply Theorem 10.2.1.

Adopting a similar bloc notation as introduced in Section 10.4.3, we write

$$A = \begin{pmatrix} A^{(\xi,\xi)} & A^{(\xi,\eta)} & A^{(\xi,\lambda)} \\ A^{(\eta,\xi)} & A^{(\eta,\eta)} & A^{(\eta,\lambda)} \\ A^{(\lambda,\xi)} & A^{(\lambda,\eta)} & A^{(\lambda,\lambda)} \end{pmatrix},$$

and define

$$\Theta_A^{(\xi)} = |\!|\!| A^{(\xi,\xi)} |\!|\!|_\gamma + |\!|\!| A^{(\eta,\xi)} |\!|\!|_\gamma + |\!|\!| A^{(\lambda,\xi)} |\!|\!|_\gamma, \quad \Theta_A^{(\eta)} = |\!|\!| A^{(\xi,\eta)} |\!|\!|_\gamma + |\!|\!| A^{(\eta,\eta)} |\!|\!|_\gamma + |\!|\!| A^{(\lambda,\eta)} |\!|\!|_\gamma,$$

$$\Theta_A^{(\lambda)} = \|A^{(\xi,\lambda)}\|_\gamma + \|A^{(\eta,\lambda)}\|_\gamma + |A^{(\lambda,\lambda)}|.$$

10.6.3 Proof of instability: the bounds Y and $Z_i(r)$

Consider $\mathcal{X}_\gamma$, F, $\bar{X} = (\bar\xi, \bar\eta, \bar\lambda)$, A, $A^\dagger$ as defined in Sections 10.6.1-10.6.2. Now we derive computable bounds Y, Z_0, Z_1 and Z_2 satisfying (10.5)-(10.8) (for $\mathcal{X} = \mathcal{X}_\gamma$).

The bound Y

Lemma 10.6.7. *Define*

$$\begin{aligned} Y = {} & \left\| A\bar F(\bar X) \right\|_{\mathcal{X}_\gamma} + \Theta_A^{(\eta)} \left(\left\| \bar\xi \right\|_\gamma \epsilon_7(\gamma) + \|\bar\eta\|_\gamma \left(\epsilon_8(\gamma) + |\bar\lambda|\epsilon_9(\gamma)\right) \right) \\ & + \Theta_A^{(\xi)} \left(\left\| K\bar\xi \right\|_\gamma \epsilon_1(\gamma) + \|K\bar\eta\|_\gamma \epsilon_2(\gamma) + \left\| \bar\xi \right\|_\gamma \left(\epsilon_3(\gamma) + |\bar\lambda|\epsilon_5(\gamma)\right) + \|\bar\eta\|_\gamma \left(\epsilon_4(\gamma) + |\bar\lambda|\epsilon_6(\gamma)\right) \right). \end{aligned} \tag{10.39}$$

Then $Y \geq \left\| AF(\bar X) \right\|_\gamma$.

Proof. Using the splitting $F = \bar F + \mathcal{E}_F$ introduced in (10.37)-(10.38), we bound separately $\left\| A\bar F(\bar X) \right\|_\gamma$ and $\left\| A\mathcal{E}_F(\bar X) \right\|_\gamma$. $A\bar F(\bar X)$ only has finitely many non zero coefficients, therefore $\left\| A\bar F(\bar X) \right\|_\gamma$ can be evaluated on a computer (using interval arithmetic to control the round-off errors).

Concerning the second term, we have that

$$\begin{aligned} \left\| \mathcal{E}_F^{(\xi)}(\bar X) \right\|_\gamma &\leq \left\| K\bar\xi \right\|_\gamma \epsilon_1(\gamma) + \|K\bar\eta\|_\gamma \epsilon_2(\gamma) + \left\| \bar\xi \right\|_\gamma \left(\epsilon_3(\gamma) + |\bar\lambda|\epsilon_5(\gamma)\right) + \|\bar\eta\|_\gamma \left(\epsilon_4(\gamma) + |\bar\lambda|\epsilon_6(\gamma)\right), \\ \left\| \mathcal{E}_F^{(\eta)}(\bar X) \right\|_\gamma &\leq \left\| \bar\xi \right\|_\gamma \epsilon_7(\gamma) + \|\bar\eta\|_\gamma \left(\epsilon_8(\gamma) + |\bar\lambda|\epsilon_9(\gamma)\right). \end{aligned}$$

Thus

$$\begin{aligned} & \left\| A\mathcal{E}_F(\bar X) \right\|_{\mathcal{X}_\gamma} \leq \\ & \quad \Theta_A^{(\xi)} \left(\left\| K\bar\xi \right\|_\gamma \epsilon_1(\gamma) + \|K\bar\eta\|_\gamma \epsilon_2(\gamma) + \left\| \bar\xi \right\|_\gamma \left(\epsilon_3(\gamma) + |\bar\lambda|\epsilon_5(\gamma)\right) + \|\bar\eta\|_\gamma \left(\epsilon_4(\gamma) + |\bar\lambda|\epsilon_6(\gamma)\right) \right) \\ & \quad + \Theta_A^{(\eta)} \left(\left\| \bar\xi \right\|_\gamma \epsilon_7(\gamma) + \|\bar\eta\|_\gamma \left(\epsilon_8(\gamma) + |\bar\lambda|\epsilon_9(\gamma)\right) \right). \end{aligned}$$

The sum of the last expression and $\left\| A\bar F(\bar X) \right\|_\gamma$ gives Y. □

The bound Z_0

Arguing exactly as in Section 10.4.3, we define

$$Z_0 = \max\left[\Theta^{(\xi)}_{I-AA^\dagger}, \Theta^{(\eta)}_{I-AA^\dagger}, \Theta^{(\lambda)}_{I-AA^\dagger}\right]. \tag{10.40}$$

The bound Z_1

Now we focus on providing a bound Z_1 satisfying (10.7).

Lemma 10.6.8. *Let $\hat{\alpha}^n_\xi, \hat{\alpha}^n_\eta$ be vectors in $\mathbb{C}^{2n+1}$ each, defined as*

$$\left(\hat{\alpha}^n_\xi\right)_0 = \begin{pmatrix} \Phi^n_0(\bar{c}_3 - K\bar{c}_1 + \bar{\lambda}\bar{c}_5, \gamma) \\ \Phi^n_0(\bar{c}_7, \gamma) \\ 0 \end{pmatrix}, \quad \left(\hat{\alpha}^n_\eta\right)_0 = \begin{pmatrix} \Phi^n_0(\bar{c}_4 - K\bar{c}_2 + \bar{\lambda}\bar{c}_6\gamma) \\ \Phi^n_0(\bar{c}_8 + \bar{\lambda}\bar{c}_9, \gamma) \\ 0 \end{pmatrix},$$

and for all $1 \le k < n$

$$\left(\hat{\alpha}^n_\xi\right)_k = \begin{pmatrix} k\Phi^n_k(\bar{c}_1, \gamma) + \Phi^n_k(\bar{c}_3 - K\bar{c}_1 + \bar{\lambda}\bar{c}_5, \gamma) \\ \Phi^n_k(\bar{c}_7, \gamma) \end{pmatrix},$$

$$\left(\hat{\alpha}^n_\eta\right)_k = \begin{pmatrix} k\Phi^n_k(\bar{c}_2, \gamma) + \Phi^n_k(\bar{c}_4 - K\bar{c}_2 + \bar{\lambda}\bar{c}_6\gamma) \\ \Phi^n_k(\bar{c}_8 + \bar{\lambda}\bar{c}_9, \gamma) \end{pmatrix}.$$

Let the operator $\tilde{K}$ acting on $\mathcal{X}_\gamma$ defined as

$$\tilde{K} = \begin{pmatrix} K & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

where K is the same operator as in Definition 10.6.4.

Choose $\tilde{\gamma} \in (\gamma, \nu)$ and define

$$\begin{aligned} Z_1 = \max&\left[\left\||A|\hat{\alpha}^n_\xi\right\|_{\mathcal{X}_\gamma}, \left\||A|\hat{\alpha}^n_\eta\right\|_{\mathcal{X}_\gamma}\right] \\ &+ \max\left[\frac{\|\bar{c}_1\|_\gamma}{\pi n} + \frac{\left\|\bar{c}_3 - K\bar{c}_1 + \bar{\lambda}\bar{c}_5\right\|_\gamma + \|\bar{c}_7\|_\gamma}{(\pi n)^2}, \frac{\|\bar{c}_2\|_\gamma}{\pi n} + \frac{\left\|\bar{c}_4 - K\bar{c}_2 + \bar{\lambda}\bar{c}_6\right\|_\gamma + \left\|\bar{c}_8 + \bar{\lambda}\bar{c}_9\right\|_\gamma}{(\pi n)^2},\right. \\ &\qquad \left.\frac{\left\|\bar{c}_5 * \bar{\xi}\right\|_\gamma + \|\bar{c}_6 * \bar{\eta}\|_\gamma + \|\bar{c}_9 * \bar{\eta}\|_\gamma}{(\pi n)^2}\right] \\ &+ \max\left[\Theta^{(\xi)}_A\left(\Upsilon^1_{\gamma,\tilde{\gamma}}\epsilon_1(\tilde{\gamma}) + \epsilon_3(\gamma) + |\bar{\lambda}|\epsilon_5(\gamma)\right) + \Theta^{(\eta)}_A\epsilon_7(\gamma),\right. \\ &\qquad \Theta^{(\xi)}_A\left(\Upsilon^1_{\gamma,\tilde{\gamma}}\epsilon_2(\tilde{\gamma}) + \epsilon_4(\gamma) + |\bar{\lambda}|\epsilon_6(\gamma)\right) + \Theta^{(\eta)}_A\left(\epsilon_8(\gamma) + |\bar{\lambda}|\epsilon_9(\gamma)\right), \\ &\qquad \left.\Theta^{(\xi)}_A\left(\left\|\bar{\xi}\right\|_\gamma\epsilon_5(\gamma) + \|\bar{\eta}\|_\gamma\epsilon_6(\gamma)\right) + \Theta^{(\eta)}_A\|\bar{\eta}\|_\gamma\epsilon_9(\gamma)\right] \\ &+ \Theta^{(\xi)}_{A\tilde{K}}\max\left[\epsilon_1(\gamma), \epsilon_2(\gamma)\right]. \end{aligned} \tag{10.41}$$

then

$$Z_1 \ge \left\|A\left(DF(\bar{X}) - A^\dagger\right)\right\|_{\mathcal{X}_\gamma}$$

Proof. Using the splitting of F introduced in in (10.37)-(10.38) we have

$$\left\| A\left(DF(\bar{X}) - A^{\dagger}\right)\right\|_{\mathcal{X}_\gamma} \leq \left\| A\left(D\bar{F}(\bar{X}) - A^{\dagger}\right)\right\|_{\mathcal{X}_\gamma} + \left\| AD\mathcal{E}_F(\bar{X})\right\|_{\mathcal{X}_\gamma}.$$

The procedure to obtain a bound for the first part on the r.h.s is similar to the one followed in Section 10.7 (since the remainder terms $(\varepsilon_j)_{1\leq j\leq 9}$ are not involved), therefore we skip most of the details. Let $X \in \mathcal{B}_{\mathcal{X}_\gamma}(0,1)$ and introduce $U = \left(D\bar{F}(\bar{X}) - A^{\dagger}\right)X$. We have

$$\begin{aligned}\left\| A\left(D\bar{F}(\bar{X}) - A^{\dagger}\right)X\right\|_{\mathcal{X}_\gamma} &\leq |||A||U|||_{\mathcal{X}_\gamma} \\ &\leq |||A||\hat{U}^n|||_{\mathcal{X}_\gamma} + |||A||\check{U}^n|||_{\mathcal{X}_\gamma},\end{aligned}$$

and we provide a bound for each term separately. For both of them, it is helpful to notice that

$$(\bar{c}_1 \bullet K\xi) = -K(\bar{c}_1 \star \xi) + (K\bar{c}_1 * \xi), \tag{10.42}$$

and similarly for $(\bar{c}_2 \bullet K\eta)$, this identity being nothing but $\bar{c}_1\xi' = (\bar{c}_1\xi)' - \bar{c}_1'\xi$ written for the Fourier sequences. Using (10.42) and proceeding exactly as in Section 10.7, we obtain

$$|||A||\hat{U}^n|||_{\mathcal{X}_\gamma} \leq \max\left[\left\||A|\hat{\alpha}^n_\xi\right\|_{\mathcal{X}_\gamma}, \left\||A|\hat{\alpha}^n_\eta\right\|_{\mathcal{X}_\gamma}\right]\|X\|_{\mathcal{X}_\gamma}. \tag{10.43}$$

For the tail part, using Lemma 10.3.3, and the definition of the tail part of A, we compute

$$\begin{aligned}\left\||A||\check{U}^n|\right\|_{\mathcal{X}_\gamma} \leq \max\Bigg[&\frac{\|\bar{c}_1\|_\gamma}{\pi n} + \frac{\left\|\bar{c}_3 - K\bar{c}_1 + \bar{\lambda}\bar{c}_5\right\|_\gamma + \|\bar{c}_7\|_\gamma}{(\pi n)^2}, \\ &\frac{\|\bar{c}_2\|_\gamma}{\pi n} + \frac{\left\|\bar{c}_4 - K\bar{c}_2 + \bar{\lambda}\bar{c}_6\right\|_\gamma + \left\|\bar{c}_8 + \bar{\lambda}\bar{c}_9\right\|_\gamma}{(\pi n)^2}, \\ &\frac{\left\|\bar{c}_5 * \bar{\xi}\right\|_\gamma + \|\bar{c}_6 * \bar{\eta}\|_\gamma + \|\bar{c}_9 * \bar{\eta}\|_\gamma}{(\pi n)^2}\Bigg]\|X\|_{\mathcal{X}_\gamma}.\end{aligned} \tag{10.44}$$

It remains to estimate $\left\|AD\mathcal{E}_F(\bar{X})\right\|_{\mathcal{X}_\gamma}$. We want to proceed in a similar fashion as in Section 10.6.3 where we computed a bound for $\left\|A\mathcal{E}_F(\bar{X})\right\|_{\mathcal{X}_\gamma}$. However, we have to be slightly more careful, since $\left\|D\mathcal{E}_F(\bar{X})\right\|_{\mathcal{X}_\gamma}$ is not finite (because of the πk terms coming from the first order derivatives). Therefore, using again (10.42), we separate the unbounded contributions and decompose $\mathcal{E}_F$ into $\mathcal{E}_F^1 + \mathcal{E}_F^2$, where

$$\begin{aligned}\left(\mathcal{E}_F^1\right)_k^{(\xi)}(X) &= -(K\varepsilon_1 * \xi)_k - (K\varepsilon_2 * \eta)_k \\ &\quad + (\varepsilon_3 * \xi)_k + (\varepsilon_4 * \eta)_k + \lambda(\varepsilon_5 * \xi)_k + \lambda(\varepsilon_6 * \eta)_k, && \forall\ k \in \mathbb{N}, \\ \left(\mathcal{E}_F^1\right)_k^{(\eta)}(X) &= (\varepsilon_7 * \xi)_k + (\varepsilon_8 * \eta)_k + \lambda(\varepsilon_9 * \eta)_k, && \forall\ k \in \mathbb{N}, \\ \left(\mathcal{E}_F^2\right)^{(\lambda)}(X) &= 0,\end{aligned}$$

and

$$\begin{aligned}\left(\mathcal{E}_F^2\right)_k^{(\xi)}(X) &= \pi k(\varepsilon_1 \star \xi)_k + \pi k(\varepsilon_2 \star \eta)_k, && \forall\ k \in \mathbb{N}, \\ \left(\mathcal{E}_F^2\right)_k^{(\eta)}(X) &= 0, && \forall\ k \in \mathbb{N}, \\ \left(\mathcal{E}_F^1\right)^{(\lambda)}(X) &= 0.\end{aligned}$$

and we provide bounds on $\left\|AD\mathcal{E}_F^1(\bar{X})X\right\|_{\mathcal{X}_\gamma}$ and $\left\|AD\mathcal{E}_F^2(\bar{X})X\right\|_{\mathcal{X}_\gamma}$, for $X=(\xi,\eta,\lambda)\in\mathcal{B}_{\mathcal{X}_\gamma}(0,1)$. The smoothing effect of A will make the second bound finite.

First, consider

$$\begin{aligned}\left\|D\left(\mathcal{E}_F^1\right)^{(\xi)}(\bar{X})X\right\|_{\mathcal{X}_\gamma} &\le \left(\|K\varepsilon_1\|_\gamma+\epsilon_3(\gamma)+|\bar{\lambda}|\epsilon_5(\gamma)\right)\|\xi\|_\gamma+\left(\|K\varepsilon_2\|_\gamma+\epsilon_4(\gamma)+|\bar{\lambda}|\epsilon_6(\gamma)\right)\|\eta\|_\gamma\\ &\quad+\left(\left\|\bar{\xi}\right\|_\gamma\epsilon_5(\gamma)+\|\bar{\eta}\|_\gamma\epsilon_6(\gamma)\right)|\lambda|,\\ \left\|D\left(\mathcal{E}_F^1\right)^{(\eta)}(\bar{X})X\right\|_{\mathcal{X}_\gamma} &\le \epsilon_7(\gamma)\|\xi\|_\gamma+\left(\epsilon_8(\gamma)+|\bar{\lambda}|\epsilon_9(\gamma)\right)\|\eta\|_\gamma+\|\bar{\eta}\|_\gamma\epsilon_9(\gamma)|\lambda|.\end{aligned}$$

Note that explicit upper bound for the $\|K\varepsilon_1\|_\gamma$ and $\|K\varepsilon_2\|_\gamma$ are required. For this, let $\tilde{\gamma}$ be such that $\gamma<\tilde{\gamma}<\nu$ and use Lemma 10.6.5 to obtain

$$\|K\varepsilon_1\|_\gamma\le\Upsilon^1_{\gamma,\tilde{\gamma}}\epsilon_1(\tilde{\gamma}),\qquad \|K\varepsilon_2\|_\gamma\le\Upsilon^1_{\gamma,\tilde{\gamma}}\epsilon_2(\tilde{\gamma}).$$

Finally, again using the bloc notation, we have

$$\begin{aligned}\left\|AD\mathcal{E}_F^1(\bar{X})X\right\|_{\mathcal{X}_\gamma} &\le \Theta_A^{(\xi)}\left\|D\left(\mathcal{E}_F^1\right)^{(\xi)}(\bar{X})X\right\|_{\mathcal{X}_\gamma}+\Theta_A^{(\eta)}\left\|D\left(\mathcal{E}_F^1\right)^{(\eta)}(\bar{X})X\right\|_{\mathcal{X}_\gamma}\\ &\quad+\Theta_A^{(\lambda)}\left\|D\left(\mathcal{E}_F^1\right)^{(\lambda)}(\bar{X})X\right\|_{\mathcal{X}_\gamma}\\ &\le\left(\Theta_A^{(\xi)}\left(\Upsilon^1_{\gamma,\tilde{\gamma}}\epsilon_1(\tilde{\gamma})+\epsilon_3(\gamma)+|\bar{\lambda}|\epsilon_5(\gamma)\right)+\Theta_A^{(\eta)}\epsilon_7(\gamma)\right)\|\xi\|_\gamma\\ &\quad+\left(\Theta_A^{(\xi)}\left(\Upsilon^1_{\gamma,\tilde{\gamma}}\epsilon_2(\tilde{\gamma})+\epsilon_4(\gamma)+|\bar{\lambda}|\epsilon_6(\gamma)\right)+\Theta_A^{(\eta)}\left(\epsilon_8(\gamma)+|\bar{\lambda}|\epsilon_9(\gamma)\right)\right)\|\eta\|_\gamma\\ &\quad+\left(\Theta_A^{(\xi)}\left(\left\|\bar{\xi}\right\|_\gamma\epsilon_5(\gamma)+\|\bar{\eta}\|_\gamma\epsilon_6(\gamma)\right)+\Theta_A^{(\eta)}\|\bar{\eta}\|_\gamma\epsilon_9(\gamma)\right)|\lambda|\\ &\le\max\Big[\Theta_A^{(\xi)}\left(\Upsilon^1_{\gamma,\tilde{\gamma}}\epsilon_1(\tilde{\gamma})+\epsilon_3(\gamma)+|\bar{\lambda}|\epsilon_5(\gamma)\right)+\Theta_A^{(\eta)}\epsilon_7(\gamma),\\ &\qquad\Theta_A^{(\xi)}\left(\Upsilon^1_{\gamma,\tilde{\gamma}}\epsilon_2(\tilde{\gamma})+\epsilon_4(\gamma)+|\bar{\lambda}|\epsilon_6(\gamma)\right)+\Theta_A^{(\eta)}\left(\epsilon_8(\gamma)+|\bar{\lambda}|\epsilon_9(\gamma)\right),\\ &\qquad\Theta_A^{(\xi)}\left(\left\|\bar{\xi}\right\|_\gamma\epsilon_5(\gamma)+\|\bar{\eta}\|_\gamma\epsilon_6(\gamma)\right)+\Theta_A^{(\eta)}\|\bar{\eta}\|_\gamma\epsilon_9(\gamma)\Big]\|X\|_{\mathcal{X}_\gamma}.\end{aligned}\tag{10.45}$$

$$\tag{10.46}$$

To deal with $\left\|AD\mathcal{E}_F^2(\bar{X})X\right\|_{\mathcal{X}_\gamma}$, we introduce $\tilde{\mathcal{E}}_F^2$ defined as

$$\begin{aligned}\left(\tilde{\mathcal{E}}_F^2\right)_k^{(\xi)}(X)&=(\varepsilon_1\star\xi)_k+(\varepsilon_2\star\eta)_k, &&\forall\ k\in\mathbb{N},\\ \left(\tilde{\mathcal{E}}_F^2\right)_k^{(\eta)}(X)&=0, &&\forall\ k\in\mathbb{N},\\ \left(\tilde{\mathcal{E}}_F^2\right)^{(\lambda)}(X)&=0,\end{aligned}$$

to get

$$AD\mathcal{E}_F^2(\bar{X})X=A\tilde{K}D\tilde{\mathcal{E}}_F^2(\bar{X})X.$$

Now we can estimate

$$\left\|D\left(\tilde{\mathcal{E}}_F^2\right)^{(\xi)}(\bar{X})X\right\|_{\mathcal{X}_\gamma}\le\epsilon_1(\gamma)\|\xi\|_\gamma+\epsilon_2(\gamma)\|\eta\|_\gamma,$$

so to obtain

$$\left\|AD\mathcal{E}_F^2(\bar{X})X\right\|_{\mathcal{X}_\gamma}\le\Theta^{(\xi)}_{A\tilde{K}}\max\left[\epsilon_1(\gamma),\epsilon_2(\gamma)\right]\|X\|_{\mathcal{X}_\gamma}.\tag{10.47}$$

Notice that $\Theta^{(\xi)}_{A\tilde{K}}$ is finite and can be compute explicitly, because the tail of A is diagonal an decreases like $(\pi k)^{-2}$. For instance

$$\begin{aligned}\left\|(A\tilde{K})^{(\xi,\xi)}\right\|_\gamma &= \sup_{j\geq 0}\frac{1}{\gamma^j}\sum_{k\geq 0}\pi j|A^{(\xi,\xi)}(k,j)|\gamma^k \\ &= \max\left[\max_{0\leq j<m}\frac{\pi j}{\gamma^j}\sum_{0\leq k<m}|A^{(\xi,\xi)}(k,j)|\gamma^k, \sup_{j\geq m}\frac{1}{\gamma^j}\pi j| - (\pi j)^{-2}|\gamma^j\right] \\ &= \max\left[\max_{0\leq j<m}\frac{\pi j}{\gamma^j}\sum_{0\leq k<m}|A^{(\xi,\xi)}(k,j)|\gamma^k, \frac{1}{\pi m}\right].\end{aligned}$$

The sum of all contributions (10.43)-(10.47) gives the required Z_1. □

The bound Z_2

Since F is quadratic, we have that for all $X' \in \mathcal{B}_{\mathcal{X}_\gamma}(0,r)$ and $X \in \mathcal{B}_{\mathcal{X}_\gamma}(0,1)$

$$A\left(DF(\bar{X}+X') - DF(\bar{X})\right)X = AD^2F(\bar{X})(X,X').$$

Direct computations give

$$\begin{aligned}D^2F^{(\xi)}(\bar{X})(X,X') &= \lambda(c_5*\xi') + \lambda'(c_5*\xi) + \lambda(c_6*\eta') + \lambda'(c_6*\eta),\\ D^2F^{(\eta)}(\bar{X})(X,X') &= \lambda(c_9*\eta') + \lambda'(c_9*\eta),\\ D^2F^{(\lambda)}(\bar{X})(X,X') &= 0,\end{aligned}$$

therefore

$$\begin{aligned}&\left\|AD^2F(\bar{X})(X,X')\right\|_{\mathcal{X}_\gamma}\\ &\quad\leq \Theta^{(\xi)}_A\left(\|\bar{c}_5\|_\gamma + \epsilon_5(\gamma)\right)\left(\|\xi\|_\gamma|\lambda'| + \|\xi'\|_\gamma|\lambda|\right)\\ &\qquad + \left(\Theta^{(\xi)}_A\left(\|\bar{c}_6\|_\gamma + \epsilon_6(\gamma)\right) + \Theta^{(\eta)}_A\left(\|\bar{c}_9\|_\gamma + \epsilon_9(\gamma)\right)\right)\left(\|\eta\|_\gamma|\lambda'| + \|\eta'\|_\gamma|\lambda|\right)\\ &\quad\leq \max\left[\Theta^{(\xi)}_A\left(\|\bar{c}_5\|_\gamma + \epsilon_5(\gamma)\right), \Theta^{(\xi)}_A\left(\|\bar{c}_6\|_\gamma + \epsilon_6(\gamma)\right) + \Theta^{(\eta)}_A\left(\|\bar{c}_9\|_\gamma + \epsilon_9(\gamma)\right)\right]\\ &\qquad\times\left(\|\xi\|_\gamma + \|\eta\|_\gamma + |\lambda|\right)\left(\|\xi'\|_\gamma + \|\eta'\|_\gamma + |\lambda'|\right)\\ &\quad= \max\left[\Theta^{(\xi)}_A\left(\|\bar{c}_5\|_\gamma + \epsilon_5(\gamma)\right), \Theta^{(\xi)}_A\left(\|\bar{c}_6\|_\gamma + \epsilon_6(\gamma)\right) + \Theta^{(\eta)}_A\left(\|\bar{c}_9\|_\gamma + \epsilon_9(\gamma)\right)\right] r\|X\|_{\mathcal{X}_\gamma}.\end{aligned}$$

Thus, we define

$$Z_2 = \max\left[\Theta^{(\xi)}_A\left(\|\bar{c}_5\|_\gamma + \epsilon_5(\gamma)\right), \Theta^{(\xi)}_A\left(\|\bar{c}_6\|_\gamma + \epsilon_6(\gamma)\right) + \Theta^{(\eta)}_A\left(\|\bar{c}_9\|_\gamma + \epsilon_9(\gamma)\right)\right]. \tag{10.48}$$

10.6.4 Proof of instability: The radii polynomial

We now collect all the bounds developed above into a statement about the stability of the steady states.

Proposition 10.6.9. *Let $\nu > 1$. Assume to have computed finite sequences of Fourier coefficients $\bar{v}$, $\bar{w}$, $\bar{p}$, $\bar{s}$ and $r_\nu > 0$ such that there exists a unique $(v,w,p,s) \in \left(\ell^1_\nu(\mathbb{R})\right)^4$ that solves* (10.15) *and satisfies*

$$\|v-\bar{v}\|_\nu + \|w-\bar{w}\|_\nu + \|p-\bar{p}\|_\nu + \|s-\bar{s}\|_\nu \leq r_\nu.$$

Choose $1 < \gamma < \nu$ *and let* $\mathcal{X}_\gamma$, F, $\bar{X}$, A, $A^\dagger$ *be as in Sections 10.6.1-10.6.2. Suppose to have computed the bounds* Y, Z_0, Z_1 *and* Z_2, *defined in* (10.39), (10.40), (10.41) *and* (10.48) *respectively. If there exists* $r > 0$ *such that*

$$P(r) = Z_2(r)r^2 - (1 - (Z_0 + Z_1))\, r + Y < 0,$$

then there exists a unique zero of F *in* $\mathcal{B}_{\mathcal{X}_\gamma}(\bar{X}, r)$. *If moreover* $\Re(\bar{\lambda}) > r$ *then the steady state* (u, v), $u = pw$, *is unstable.*

Proof. It follows as application of Theorem 10.2.1 and Lemma 10.6.2 □

10.7 Results about the instability of steady states

In this section, we give some details about the proof of Theorem 10.1.2. We recall that the parameters of (10.4) are fixed as $\Omega = (0, 1)$, $r_1 = 5$, $r_2 = 2$, $a_1 = 3$, $a_2 = 3$, $b_1 = 1$, $b_2 = 1$, $d_{12} = 3$, and that $d_1 = d_2 = d$ is left as the bifurcation parameter. For each solution represented by a blue dot on Figure 10.3, we proved the existence of an unstable eigenvalue, using the procedure described at the end of Section 10.2. In particular, for each of these steady states we computed numerically an eigenvalue with positive real part, implemented the bounds described in Section 10.6.3, and then *successfully* applied Proposition 10.6.9 to validate the numerical eigenvalue. By successfully we mean that we found a positive r such that $P(r) < 0$ and checked that $\Re(\bar{\lambda}) > r$. For the steady states displayed previously in Figure 10.4, we detail in Table 10.2 what is the value of the unstable eigenvalue, what dimension n was used for the finite dimensional projection, what γ was chosen for the space $\mathcal{X}_\gamma$, and give a validation radius r for which the proof is succesfull with those parameters (for the steady states where an unstable eigenvalue was actually found).

Solution (see Figure 10.4)	unstable eigenvalue λ	n	γ	Validation radius
(a)	no unstable eigenvalue found	–	–	–
(b)	0.0153	1000	1.0001	3.2807×10^{-7}
(c)	$0.2050 \pm 0.1673i$	1000	1.0001	7.8894×10^{-6}
(d)	$0.0463 \pm 0.0524i$	1000	1.0001	8.0803×10^{-6}
(e)	0.0844	1000	1.0001	2.5572×10^{-7}
(f)	$0.0463 \pm 0.0524i$	1100	1.0001	1.2238×10^{-5}
(g)	$0.0570 \pm 0.0390i$	1400	1.0001	1.0099×10^{-5}
(h)	0.2743	1000	1.0001	4.3919×10^{-9}
(i)	0.0844	1000	1.0001	2.2639×10^{-7}
(j)	0.0153	1000	1.0001	2.1499×10^{-7}
(k)	$0.0570 \pm 0.0390i$	1500	1.0001	1.093×10^{-5}
(l)	no unstable eigenvalue found	–	–	–
(m)	$0.2050 \pm 0.1673i$	1000	1.0001	4.0795×10^{-6}

Table 10.2 – For each steady state displayed in Figure 10.4, when an unstable eigenvalue is found we give the dimension n that was used for the finite dimensional projection, the weight γ that was chosen for the space $\mathcal{X}_\gamma$, and a validation radius r for which the proof of the eigenvalue is successful, with those parameters n and γ.

Proof of Theorem 10.1.2. In the script `script_proof_branch_instability.m` fix the values of the parameters $r_1 = 5$, $r_2 = 2$, $a_1 = 3$, $a_2 = 3$, $b_1 = 1$, $b_2 = 1$, $d_{12} = 3$. The parameter

$d_1 = d_2 = d$ is intended as the bifurcation parameter. Choose a value for the finite dimensional projection n and a value for the norm weight $\gamma > 1$. Also select a branch of solutions (for the names of the several branches we refer to the documentation and the `readme` file). The script loads the numerical data, computes the required bounds and verifies the existence of an interval $\mathcal{I} = (r_1, r_2)$ such that $P(r) < 0$ for any $r \in \mathcal{I}$. If $\mathcal{I}$ is not empty then the condition $\Re(\bar{\lambda}) > r$ is checked. In case of successful computation, Proposition 10.6.9 implies that the concerned steady state is unstable. The values for n and γ that allow the rigorous computation of all the branches depicted in the Figure 10.2 are available in the documentation.

The script `script_proof_steadystate_and_instability.m` concerns the existence of steady states for a fixed value of d. It is used to prove the existence of 13 solutions at values $d = 0.005$. Figure 10.4 shows the numerical data for the 13 steady states solutions. In Table 10.2 we detail the values for n and γ used in the proof and the resulting validation radius r. □

Part IV

Bibliography

Chapter 11

About coagulation-fragmentation equations (Parts I and II)

[1] R.A. Adams and J.J.F. Fournier. *Sobolev Spaces*. Academic press, second edition, 2003.

[2] H. Amann. Coagulation-fragmentation processes. *Arch. Rational Mech. Anal.*, 151: 339–366, 2000.

[3] J.M. Ball and J. Carr. The discrete coagulation-fragmentation equations: Existence, uniqueness, and density conservation. *Journal of Statistical Physics*, 61(1): 203–234, 1990.

[4] J.M. Ball, J. Carr and O. Penrose. The Becker-Döring cluster equations: basic properties and asymptotic behaviour of solutions. *Communications in Mathematical Physics*, 104(4): 657–692, 1986.

[5] R. Becker and W. Döring. Kinetische Behandlung der Keimbildung in übersättigten Dämpfen. *Annalen der Physik*, 416(8): 719–752, 1935.

[6] M. Breden. Applications of improved duality lemmas to the discrete coagulation-fragmentation equations with diffusion. *To appear in Kinetic and Related Models*, arXiv:1606.07661v2, 2017.

[7] M. Breden, L. Desvillettes and K. Fellner. Smoothness of moments of the solutions of discrete coagulation equations with diffusion. *Monatshefte für Mathematik*, 183(3): 437–463, 2017.

[8] J. A. Cañizo. *Some problems related to the study of interaction kernels: coagulation, fragmentation and diffusion in kinetic and quantum equations.* PhD thesis, Universidad de Granada, June 2006.

[9] J. A. Cañizo, L. Desvillettes and K. Fellner. Absence of gelation for models of coagulation-fragmentation with degenerate diffusion. *Il Nuovo cimento della Società italiana di fisica. C*, 33(1): 79, 2010.

[10] J. A. Cañizo, L. Desvillettes and K. Fellner. Regularity and mass conservation for discrete coagulation–fragmentation equations with diffusion. *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, 27(2): 639–654, 2010.

[11] J. A. Cañizo, L. Desvillettes and K. Fellner. Improved duality estimates and applications to reaction-diffusion equations. *Communications in Partial Differential Equations*, 39(6): 1185–1204, 2014.

[12] J. Carr. Asymptotic behaviour of solutions to the coagulation-fragmentation equations. I. The strong fragmentation case. *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, 121(3-4): 231–244, 1992.

[13] J. A. Carrillo, L. Desvillettes and K. Fellner. Rigorous derivation of a nonlinear diffusion equation as fast-reaction limit of a continuous coagulation-fragmentation model with diffusion. *Communications in Partial Differential Equations*, 34(11): 1338–1351, 2009.

[14] F.P. da Costa. Existence and uniqueness of density conserving solutions to the coagulation-fragmentation equations with strong fragmentation. *Journal of mathematical analysis and applications*, 192(3): 892–914, 1995.

[15] L. Desvillettes. About Entropy Methods for Reaction-Diffusion Equations. *Rivista di Matematica dell'Università di Parma*, 7(7): 81–123, 2007.

[16] L. Desvillettes, K. Fellner. Duality and Entropy Methods in Coagulation-Fragmentation Models. *Revista di Matematica della Universita di Parma*, 4(2): 215–263, 2013.

[17] L. Desvillettes, K. Fellner, M. Pierre and J. Vovelle. Global existence for quadratic systems of reaction-diffusion. *Advanced Nonlinear Studies*, 7(3): 491–511, 2007.

[18] R.L. Drake. A general mathematical survey of the coagulation equation. *International Reviews in Aerosol Physics and Chemistry*, Oxford, 203–376, 1972.

[19] M. Escobedo, P. Laurençot and S. Mischler. Fast reaction limit of the discrete diffusive coagulation-fragmentation equation. *Communications in Partial Differential Equations*, 28(5-6): 1113–1133, 2003.

[20] M. Escobedo, P. Laurençot, S. Mischler and B. Perthame. Gelation and mass conservation in coagulation and fragmentation models. *Journal of Differential Equations*, 195(1): 143–174, 2003.

[21] M. Escobedo, S. Mischler and B. Perthame. Gelation in coagulation and fragmentation models. *Communications in Mathematical Physics*, 231(1): 157–188, 2002.

[22] R. Haller-Dintelmann, H. Heck and M. Hieber. L_p–L_q-estimates for parabolic systems in non-divergence form with VMO coefficients. *J. London Math. Soc. (2)*, 74(3): 717–736, 2006.

[23] A. Hammond and F. Rezakhanlou. Moment Bounds for the Smoluchowski Equation and their Consequences. *Communications in Mathematical Physics*, 276(3): 645–670, 2007.

[24] E.M. Hendriks, M.H. Ernst and R.M. Ziff. Coagulation equations with gelation. *Journal of Statistical Physics*, 31(3): 519–563, 1983.

[25] N.V. Krylov. Parabolic equations with VMO coefficients in Sobolev spaces with mixed norms. *Journal of Functional Analysis*, 250(2): 521–558, 2007.

[26] O. A. Ladyzenskaja, V. A. Solonnikov and N. N. Uralceva. *Linear and quasilinear equations of parabolic type.* Translations of Mathematical Monographs, Vol. 23, American Mathematical Society, 1968.

[27] D. Lamberton. Equations d'évolution linéaires associées à des semi-groupes de contractions dans les espaces L^p. *Journal of Functional Analysis*, 72(2): 252–262, 1987.

[28] P. Laurençot and S. Mischler. Global existence for the discrete diffusive coagulation-fragmentation equations in L^1. *Revista Matemática Iberoamericana*, 18(3): 731–745, 2002.

[29] P. Laurençot and S. Mischler. From the discrete to the continuous coagulation-fragmentation equations. *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, 132(05): 1219–1248, 2002.

[30] P. Laurençot and S. Mischler. On coalescence equations and related models. In Modeling and computational methods for kinetic equations, Model. Simul. Sci. Eng. Technol. Birkhäuser Boston, Boston, MA, 321–356, 2004.

[31] A. Lushnikov, Some new aspects of coagulation theory. *Izv. Akad. Nauk SSSR, Ser. Fiz. Atmosfer. I Okeana*, 14: 738–743, 1978.

[32] A. Marcus. Stochastic coalescence. *Technometrics*, 10: 133–143, 1968.

[33] A. Maugeri, D. Palagachev and L. Softova. *Elliptic and Parabolic Equations with Discontinuous Coefficients.* Wiley-VCH, 2002.

[34] H. Müller. Zur allgemeinen Theorie der raschen Koagulation. *Kolloidchemische Beihefte*, 27(6): 223–250, 1928.

[35] J.R. Norris. Smoluchowski's coagulation equation : uniqueness, nonuniqueness and a hydrodynamic limit for the stochastic coalescent. *Ann. Appl. Probab.*, 9: 78–109, 1999.

[36] J. R. Norris. Brownian coagulation. *Communications in Mathematical Sciences*, 2(Supplemental Issue): 93–101, 2004.

[37] M. Pierre. Global existence in reaction-diffusion systems with control of mass: a survey. *Milan J. Math.*, 78(2): 416–455, 2010.

[38] M. Pierre and D. Schmitt. Blowup in reaction-diffusion systems with dissipation of mass. *SIAM J. Math. Anal.*, 28(2): 259–269, 1997.

[39] F. Rezakhanlou. Moment bounds for the solutions of the Smoluchowski equation with coagulation and fragmentation. *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, 140(5): 1041–1059, 2010.

[40] F. Rezakhanlou. Pointwise bounds for the solutions of the Smoluchowski equation with diffusion. *Archive for Rational Mechanics and Analysis*, 212(3): 1011–1035, 2014.

[41] M. Slemrod. The Becker-Döring equations. *Modeling in applied sciences*. Birkhäuser Boston, 149–171, 2000.

[42] M. Smoluchowski. Drei Vorträge über Diffusion, Brownsche Molekularbewegung und Koagulation von Kolloidteilchen. *Physik. Zeitschr.*, 17: 557–599, 1916.

[43] M. Smoluchowski. Versuch einer mathematischen Theorie der Koagulationskinetik kolloider Lösungen. *Zeitschrift f. physik. Chemie*, 92: 129–168, 1917.

[44] W. H. White. A global existence theorem for Smoluchowski's coagulation equations. *Proc. Am. Math. Soc*, 80: 273–276, 1980.

[45] D. Wrzosek. Existence of solutions for the discrete coagulation-fragmentation model with diffusion. *Topological Methods in Nonlinear Analysis*, 9(2): 279–296, 1997.

[46] D. Wrzosek, Mass-conserving solutions to the discrete coagulation fragmentation model with diffusion. *Nonlinear Anal.*, 49: 297–314, 2002.

[47] D. Wrzosek. Weak solutions to the Cauchy problem for the diffusive discrete coagulation–fragmentation system. *J. Math. Anal. Appl.*, 289(2): 405–418, 2004.

Chapter 12

About validated numerics and their applications (Parts I and III)

[48] H. Amann. Dynamic theory of quasilinear parabolic systems. III. Global existence. *Math. Z.*, 202(2): 219–250, 1989.

[49] D. Ambrosi, G. Arioli, and H. Koch. A homoclinic solution for excitation waves on a contractile substratum. *SIAM J. Appl. Dyn. Syst.*, 11(4):1533–1542, 2012.

[50] L. D'Ambrosio, J.-P. Lessard and A. Pugliese, Blow-up profile for solutions of a fourth order nonlinear equation *Nonlinear Anal.*, 121(7): 280–335, 2015.

[51] Z. Arai, W. Kalies, H. Kokubu, K. Mischaikow, H. Oka, and P. Pilarczyk. A database schema for the analysis of global dynamics of multiparameter systems. *SIAM J. Appl. Dyn. Syst.*, 8(3): 757–789, 2009.

[52] G. Arioli, V. Barutello, and S. Terracini. A new branch of Mountain Pass solutions for the choreographical 3-body problem. *Comm. Math. Phys.*, 268(2):439–463, 2006.

[53] G. Arioli and H. Koch. Computer-assisted methods for the study of stationary solutions in dissipative systems, applied to the Kuramoto-Sivashinski equation. *Archive for rational mechanics and analysis*, 197(3): 1033–1051, 2010.

[54] G. Arioli and H. Koch. Existence and stability of traveling pulse solutions of the FitzHugh-Nagumo equation. *Nonlinear Anal.*, 113:51–70, 2015.

[55] V. Arnold. Sur la topologie des écoulements stationaires des fluides parfaits. *C. R. Acad. Sci. Paris*, 261:17–20, 1965.

[56] W. Bahsoun and C. Bose. Invariant densities and escape rates: Rigorous and computable approximations in the L^∞-norm. *Nonlinear Anal*, 74(13):4481–4495, 2011

[57] A. W. Baker, M.l Dellnitz, and O. Junge. A topological method for rigorously computing periodic orbits using Fourier modes. *Discrete Contin. Dyn. Syst.*, 13(4):901–920, 2005.

[58] J.B. van den Berg, M. Breden, J.-P. Lessard, and M. Murray. Continuation of homoclinic orbits in the suspension bridge equation: a computer-assisted proof. *Submitted*, arXiv:1702.07412, 2017.

[59] J.B. van den Berg, M. Breden, J.-P. Lessard, and M. Murray. MATLAB code for "Continuation of homoclinic orbits in the suspension bridge equation: a computer-assisted proof", 2017. `http://www.math.vu.nl/~janbouwe/code/suspensionbridge/`.

[60] J.B. van den Berg, A. Deschênes, J.-P. Lessard and J.D. Mireles James. Stationary coexistence of hexagons and rolls via rigorous computations. *SIAM Journal on Applied Dynamical Systems*, 14(2): 942–979, 2015.

[61] J. B. van den Berg, C. M. Groothedde, and J. F. Williams. Rigorous computation of a radially symmetric localized solution in a Ginzburg-Landau problem. *SIAM J. Appl. Dyn. Syst.*, 14(1):423–447, 2015.

[62] J. B. van den Berg and J.-P. Lessard. Chaotic braided solutions via rigorous numerics: chaos in the Swift-Hohenberg equation. *SIAM Journal on Applied Dynamical Systems*, 7(3): 988–1031, 2008.

[63] J. B. van den Berg and J.-P. Lessard. Rigorous Numerics in Dynamics. *Notices of the AMS*, 62(9), 2015.

[64] J.B. van den Berg, J.-P. Lessard, and K. Mischaikow. Global smooth solution curves using rigorous branch following. *Math. Comp.*, 79(271):1565–1584, 2010.

[65] J.B. van den Berg, J.-P. Lessard, J.D. Mireles James and K. Mischaikow. Rigorous numerics for symmetric connecting orbits: even homoclinics of the Gray-Scott equation. *SIAM Journal on Mathematical Analysis*, 43 (4): 1557–1594, 2011.

[66] J.B. van den Berg, J.D. Mireles James and C. Reinhardt. Computing (un)stable manifolds with validated error bounds: non-resonant and resonant spectra. *Journal of Nonlinear Science*, 26(4): 1055–1095, 2016.

[67] J.B. van den Berg and R. Sheombarsing. Rigorous numerics for ODEs using Chebyshev series and domain decomposition. Submitted, 2015.

[68] J.B. van den Berg and D. Smets. Homoclinic solutions for Swift-Hohenberg and suspension bridge type equations. *J. Differential Equations*, 184(1):78–96, 2002.

[69] M. Berz and K. Makino. Verified integration of ODEs and flows using differential algebraic methods on high-order Taylor models. *Reliab. Comput.*, 4(4):361–369, 1998.

[70] J. P. Boyd. *Chebyshev and Fourier spectral methods*. Dover Publications Inc., Mineola, NY, second edition, 2001.

[71] M. Breden and R. Castelli. Existence and instability of steady states for a triangular cross-diffusion system: a computer-assisted proof. *Submitted*, arXiv:1704.03827, 2017.

[72] M. Breden and R. Castelli. MATLAB code for "Existence and instability of steady states for a triangular cross-diffusion system: a computer-assisted proof", 2017. `http://www.few.vu.nl/~rci270/publications.php`.

[73] M. Breden, L. Desvillettes and J.-P. Lessard. Rigorous numerics for nonlinear operators with tridiagonal dominant linear part *Discrete and Continuous Dynamical Systems - Series A*, 35(10): 4765–4789, 2015.

[74] M. Breden, L. Desvillettes and J.-P. Lessard. MATLAB codes for "Rigorous numerics for nonlinear operators with tridiagonal dominant linear part", 2015. `http://archimede.mat.ulaval.ca/jplessard/PseudoInverse/`.

[75] M. Breden and J.-P. Lessard. Polynomial interpolation and a priori bootstrap for computer-assisted proofs in nonlinear ODEs. *Submitted*, arXiv:1704.03128, 2017.

[76] M. Breden and J.-P. Lessard. MATLAB codes for "Polynomial interpolation and a priori bootstrap for computer-assisted proofs in nonlinear ODEs", 2017. `http://archimede.mat.ulaval.ca/jplessard/AprioriBootstrap/`.

[77] M. Breden, J.-P. Lessard and J.D. Mireles James. Computation of maximal local (un)stable manifold patches by the parameterization method. *Indagationes Mathematicae*, 27(1): 340–367, 2016.

[78] M. Breden, J.-P. Lessard, and J.D. Mireles James, MATLAB codes for "Computation of maximal local (un)stable manifold patches by the parameterization method", 2016. `http://archimede.mat.ulaval.ca/jplessard/MaximizingManifold/`.

[79] M. Breden, J.-P. Lessard and M. Vanicat. Global bifurcation diagrams of steady states of systems of PDEs via rigorous numerics: a 3-component reaction-diffusion system. *Acta applicandae mathematicae*, 128(1): 113–152, 2013.

[80] B. Breuer, J. Horák, P.J. McKenna, and M. Plum. A computer-assisted existence and multiplicity proof for travelling waves in a nonlinearly supported beam. *J. Differential Equations*, 224(1):60–97, 2006.

[81] B. Buffoni, A. R. Champneys, and J. F. Toland. Bifurcation and coalescence of a plethora of homoclinic orbits for a Hamiltonian system. *J. Dynam. Differential Equations*, 8(2):221–279, 1996.

[82] X. Cabré, E. Fontich, and R. de la Llave. The parameterization method for invariant manifolds I: manifolds associated to non-resonant subspaces. *Indiana University mathematics journal*, 52(2):283–328, 2003.

[83] X. Cabré, E. Fontich, and R. de la Llave. The parameterization method for invariant manifolds. II. Regularity with respect to parameters. *Indiana Univ. Math. J.*, 52(2):329–360, 2003.

[84] X. Cabré, E. Fontich, and R. de la Llave. The parameterization method for invariant manifolds. III. Overview and applications. *J. Differential Equations*, 218(2):444–515, 2005.

[85] S. Cai, J. Zeng. A Computer-Assisted Stability Proof for a Stationary Solution of Reaction-Diffusion Equations. *preprint*, arXiv:1408.4678, 2014.

[86] CAPD: Computer assisted proofs in dynamics, a package for rigorous numerics. `http://capd.ii.uj.edu.pl/`.

[87] M. J. Capiński. Covering relations and the existence of topologically normally hyperbolic invariant sets. *Discrete Contin. Dyn. Syst.*, 23(3):705–725, 2009.

[88] M. J. Capiński and P. Roldán. Existence of a center manifold in a practical domain around L_1 in the restricted three-body problem. *SIAM J. Appl. Dyn. Syst.*, 11(1):285–318, 2012.

[89] M. J. Capiński and C. Simó. Computer assisted proof for normally hyperbolic invariant manifolds. *Nonlinearity*, 25(7):1997–2026, 2012.

[90] M. J. Capiński and P. Zgliczyński. Cone conditions and covering relations for topologically normally hyperbolic invariant manifolds. *Discrete Contin. Dyn. Syst.*, 30(3):641–670, 2011.

[91] M. J. Capiński and P. Zgliczyński. Geometric proof for normally hyperbolic invariant manifolds. *J. Differential Equations*, 259(11):6215–6286, 2015.

[92] R. Castelli. Rigorous computation of non-uniform patterns for the 2-dimensional Gray-Scott reaction-diffusion equation. *Under review*, 2016.

[93] R. Castelli and J.-P. Lessard. Rigorous Numerics in Floquet Theory: Computing Stable and Unstable Bundles of Periodic Orbits. *SIAM J. Appl. Dyn. Syst.*, 12(1):204–245, 2013.

[94] R. Castelli, J.-P. Lessard, and J. D. Mireles James. Parameterization of invariant manifolds for periodic orbits I: efficient numerics via the Floquet normal form. *SIAM Journal on Applied Dynamical Systems*, 14(1):132–167, 2015.

[95] R. Castelli, J.-P. Lessard and J.D. Mireles James. Parameterization of invariant manifolds for periodic orbits (II): a-posteriori analysis and computer-assisted error bounds. *Submitted*, 2016.

[96] R. Castelli and H. Teismann. Rigorous numerics for NLS: bound states, spectra, and controllability. *Physica D: Nonlinear Phenomena*, 334: 158–173, 2016.

[97] J. Chen and P.J. McKenna. Travelling waves in a suspension bridge. *SIAM Journal on Applied Mathematics*, 50(3):703–715, 1990.

[98] J. Chen and P.J. McKenna. Travelling waves in a nonlinearly suspended beam: theoretical results and numerical observations. *J. Differential Equations*, 136:325–355, 1997.

[99] E. W. Cheney. *Introduction to approximation theory*. AMS Chelsea Publishing, Providence, RI, 1998. Reprint of the second (1982) edition.

[100] Y. Choi, R. Lui and Y. Yamada. Existence of global solutions for the Shigesada-Kawasaki-Teramoto model with weak cross-diffusion. *Discrete and Continuous Dynamical Systems*, 9(5): 1193–1200, 2003.

[101] S.N. Chow and J.K. Hale. *Methods of bifurcation theory*, volume 251 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Science]*. Springer-Verlag, New York, 1982.

[102] Philippe G. Ciarlet. *Introduction to numerical linear algebra and optimisation*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 1989. With the assistance of Bernadette Miara and Jean-Marie Thomas, Translated from the French by A. Buttigieg.

[103] B.A. Coomes, H. Koçak, and K. J. Palmer. Homoclinic shadowing. *J. Dynam. Differential Equations*, 17(1):175–215, 2005.

[104] B. A. Coomes, H. Koçak, and K. J. Palmer. Transversal connecting orbits from shadowing. *Numer. Math.*, 106(3):427–469, 2007.

[105] A. Correc and J.-P. Lessard. Coexistence of nontrivial solutions of the one-dimensional ginzburg-landau equation: a computer-assisted proof. *European Journal of Applied Mathematics*, 26(1):33–60, 2015.

[106] J. Cyranka and T. Wanner. Computer-assisted proof of heteroclinic connections in the one-dimensional Ohta-Kawasaki model. *arXiv preprint*, arXiv:1703.01022, 2017.

[107] S. Day, O. Junge, and K. Mischaikow. A rigorous numerical method for the global analysis of infinite-dimensional discrete dynamical systems. *SIAM J. Appl. Dyn. Syst.*, 3(2):117–160 (electronic), 2004.

[108] S. Day and W. D. Kalies. Rigorous computation of the global dynamics of integrodifference equations with smooth nonlinearities. *SIAM J. Numer. Anal.*, 51(6):2957–2983, 2013.

[109] S. Day, J.-P. Lessard and K. Mischaikow. Validated continuation for equilibria of PDEs. *SIAM J. Numer. Anal.*, 45(4): 1398–1424 (electronic), 2007.

[110] L. Desvillettes, T. Lepoutre, A. Moussa and A. Trescases. On the entropic structure of reaction-cross-diffusion systems *Communications in Partial Differential Equations*, 40(9): 1705–1747, 2015.

[111] L. Desvillettes and A. Trescases. New results for triangular reaction cross-diffusion system. *Journal of Mathematical Analysis and Applications*, 430(1): 32–59, 2015.

[112] T. Dombre, U. Frisch, J. M. Greene, M. Hénon, A. Mehr, and A. M. Soward. Chaotic streamlines in the ABC flows. *J. Fluid Mech.*, 167:353–391, 1986.

[113] J.-P. Eckmann, H. Koch, and P. Wittwer. A computer-assisted proof of universality for area-preserving maps. *Mem. Amer. Math. Soc.*, 47(289):vi+122, 1984.

[114] H. Ehlich and K. Zeller. Auswertung der Normen von Interpolationsoperatoren. *Math. Ann.*, 164:105–112, 1966.

[115] J.-L. Figueras and À. Haro. Reliable computation of robust response tori on the verge of breakdown. *SIAM J. Appl. Dyn. Syst.*, 11(2):597–628, 2012.

[116] J.-L. Figueras, A. Haro, and A. Luque. Rigorous computer assisted application of KAM theory: a modern approach. *Foundations of Computational Mathematics*, 2017.

[117] S. Galatolo and I. Nisoli. Rigorous computation of invariant measures and fractal dimension for maps with contracting fibers: 2D Lorenz-like maps, *Ergodic Theory and Dynamical Systems*, 36(6):1865–1891, 2016.

[118] M. Gameiro and J.-P. Lessard. Analytic estimates and rigorous continuation for equilibria of higher-dimensional PDEs. *J. Differential Equations*, 249(9): 2237–2268, 2010.

[119] M. Gameiro and J.-P. Lessard. Existence of secondary bifurcations or isolas for PDEs. *Nonlinear Anal.*, 74(12): 4131–4137, 2011.

[120] M. Gameiro and J.-P. Lessard. Rigorous computation of smooth branches of equilibria for the three dimensional Cahn-Hilliard equation. *Numer. Math.*, 117(4): 753–778, 2011.

[121] M. Gameiro and J.-P. Lessard. Efficient Rigorous Numerics for Higher-Dimensional PDEs via One-Dimensional Estimates. *SIAM J. Numer. Anal.*, 51(4):2063–2087, 2013.

[122] M. Gameiro and J.-P. Lessard. A posteriori verification of invariant objects of evolution equations: periodic orbits in the Kuramoto-Sivashinsky PDE. *SIAM Journal on Applied Dynamical Systems*, 16(1): 687–728, 2017.

[123] M. Gameiro, J.-P. Lessard, and A. Pugliese. Computation of smooth manifolds via rigorous multi-parameter continuation in infinite dimensions. *Found. Comput. Math.*, 16(2):531–575, 2016.

[124] F. Gazzola. Nonlinearity in oscillating bridges. *Electron. J. Differential Equations*, pages No. 211, 47, 2013.

[125] M. Gidea and P. Zgliczyński. Covering relations for multidimensional dynamical systems. *Journal of Differential Equations*, 202(1): 32–58, 2004.

[126] A.D. Gilbert, S. Friedlander and M. Vishik. Hydrodynamic instability for certain abc flows. *Geophysical and Astrophysical Fluid Dynamics*, 73(1-4):97–107, 1993.

[127] R. H. Goodman and J. K. Wróbel. High-order bisection method for computing invariant manifolds of two-dimensional maps. *Internat. J. Bifur. Chaos Appl. Sci. Engrg.*, 21(7):2017–2042, 2011.

[128] A. Guillamon and G. Huguet. A computational and geometric approach to phase resetting curves and surfaces. *SIAM J. Appl. Dyn. Syst.*, 8(3):1005–1042, 2009.

[129] A. Haro. Automatic differentiation methods in computational dynamical systems. IMA New Directions short course, 2011.

[130] A. Haro, M. Canadell, J.-L. Figueras, A. Luque, and J.-M. Mondelo. *The parameterization method for invariant manifolds*, volume 195 of *Applied Mathematical Sciences*. Springer, [Cham], 2016. From rigorous results to effective computations.

[131] À. Haro and R. de la Llave. A parameterization method for the computation of invariant tori and their whiskers in quasi-periodic maps: numerical algorithms. *Discrete Contin. Dyn. Syst. Ser. B*, 6(6):1261–1300 (electronic), 2006.

[132] Y. Hiraoka. Rigorous numerics for symmetric homoclinic orbits in reversible dynamical systems. *Kybernetika (Prague)*, 43(6):797–806, 2007.

[133] Y. Hiraoka and T. Ogawa. Rigorous numerics for localized patterns to the quintic Swift-Hohenberg equation. *Japan J. Indust. Appl. Math.*, 22(1):57–75, 2005.

[134] G. Huguet and R. de la Llave. Computation of limit cycles and their isochrons: fast algorithms and their convergence. *SIAM J. Appl. Dyn. Syst.*, 12(4):1763–1802, 2013.

[135] G. Huguet, R. de la Llave, and Y. Sire. Computation of whiskered invariant tori and their associated manifolds: new fast algorithms. *Discrete Contin. Dyn. Syst.*, 32(4):1309–1353, 2012.

[136] A. Hungria, J.-P. Lessard, and J. D. Mireles James. Radii polynomial approach for analytic solutions of differential equations: Theory, examples, and comparisons. *Math. Comp.*, 85 (299): 1427–1459, 2016.

[137] M. Iida, M. Mimura and H. Ninomiya. Diffusion, cross-diffusion and competitive interaction. *Journal of Mathematical Biology*, 53(4): 617–641, 2006.

[138] H. Izuhara. and M. Mimura. Reaction-diffusion system approximation to the cross-diffusion competition system. *Hiroshima Mathematical Journal*, 38(2): 315–347, 2008.

[139] À. Jorba and M. Zou. A software package for the numerical integration of ODEs by means of high-order Taylor methods. *Experiment. Math.*, 14(1):99–117, 2005.

[140] A. Jüngel. Cross-Diffusion Systems. *Entropy Methods for Diffusive Partial Differential Equations*, Springer International Publishing, 69–108, 2016.

[141] T. Kinoshita, Y.Watanabe and M. T. Nakao. An improvement of the theorem of a posteriori estimates for inverse elliptic operators. *Nonlinear Theory and Its Applications, IEICE*, 5(1): 47–52, 2014.

[142] K. Kishimoto. The diffusive Lotka-Volterra system with three species can have a stable non-constant equilibrium solution. *Journal of Mathematical Biology*, 16(1): 103–112, 1982.

[143] Gábor Kiss and Jean-Philippe Lessard. Computational fixed-point theory for differential delay equations with multiple time lags. *J. Differential Equations*, 252(4):3093–3115, 2012.

[144] K. Kishimoto and H. F. Weinberger. The Spatial Homogeneity of Stable Equilibria of Some Reaction-Diffusion Systems on Convex Domains. *Journal of Differential Equations*, 58(1): 15–21, 1985.

[145] Donald E. Knuth. *The art of computer programming. Vol. 2.* Addison-Wesley Publishing Co., Reading, Mass., second edition, 1981. Seminumerical algorithms, Addison-Wesley Series in Computer Science and Information Processing.

[146] H. Koçak, K. Palmer, and B. Coomes. Shadowing in ordinary differential equations. *Rend. Semin. Mat. Univ. Politec. Torino*, 65(1):89–113, 2007.

[147] H. Koch, A. Schenkel, and P. Wittwer. Computer-assisted proofs in analysis and programming in logic: a case study. *SIAM Rev.*, 38(4):565–604, 1996.

[148] V. R. Korostyshevskiy and Thomas Wanner. A Hermite spectral method for the computation of homoclinic orbits and associated functionals. *J. Comput. Appl. Math.*, 206(2):986–1006, 2007.

[149] V. R. Korostyshevskiy. *A Hermite spectral approach to homoclinic solutions of ordinary differential equations.* ProQuest LLC, Ann Arbor, MI, 2005. Thesis (Ph.D.)–University of Maryland, Baltimore County.

[150] B. Krauskopf and H. Osinga. Two-dimensional global manifolds of vector fields. *Chaos*, 9(3):768–774, 1999.

[151] B. Krauskopf, H. M. Osinga, E. J. Doedel, M. E. Henderson, J. Guckenheimer, A. Vladimirsky, M. Dellnitz, and O. Junge. A survey of methods for computing (un)stable manifolds of vector fields. *Internat. J. Bifur. Chaos Appl. Sci. Engrg.*, 15(3):763–791, 2005.

[152] O. E. Lanford, III. A computer-assisted proof of the Feigenbaum conjectures. *Bull. Amer. Math. Soc. (N.S.)*, 6(3):427–434, 1982.

[153] J.-P. Lessard. Recent advances about the uniqueness of the slowly oscillating periodic solutions of Wright's equation. *Journal of Differential Equations*, 248 (5): 992–1016, 2010.

[154] J.-P. Lessard, J.D. Mireles James and J. Ransford. Automatic differentiation for Fourier series and the radii polynomial approach. *Physica D: Nonlinear Phenomena*, 334: 174–186, 2016.

[155] J.-P. Lessard, J.D. Mireles James and C. Reinhardt. Computer assisted proof of transverse saddle-to-saddle connecting orbits for first order vector fields. *Journal of Dynamics and Differential Equations*, 26(2): 267–313, 2014.

[156] J.-P. Lessard and C. Reinhardt. Rigorous numerics for nonlinear differential equations using Chebyshev series. *SIAM J. Numer. Anal.*, 52(1):1–22, 2014.

[157] R. de la Llave and J. D. Mireles James. Connecting orbits for compact infinite dimensional maps: Computer assisted proofs of existence. *SIAM Journal on Applied Dynamical Systems*, 15(2): 1268–1323, 2016.

[158] Y. Lou, W.-M. Ni and S. Yotsutani. On a limiting system in the Lotka-Volterra competition with cross-diffusion. *Discrete and Continuous Dynamical Systems*, 10(1/2): 435–458, 2004.

[159] K. Makino and M. Berz. Taylor models and other validated functional inclusion methods. *Int. J. Pure Appl. Math.*, 4(4):379–456, 2003.

[160] H. Matano and M. Mimura. Pattern formation in competition-diffusion systems in nonconvex domains. *Publications of the Research Institute for Mathematical Sciences*, 19(3): 1049–1079.

[161] P. J. McKenna, F. Pacella, M. Plum, and D. Roth. A Uniqueness Result for a Semilinear Elliptic Problem: A Computer-assisted Proof. *Journal of Differential Equations*, 247: 2140–2162, 2009

[162] P.J. McKenna and W. Walter. Travelling waves in a suspension bridge. *SIAM J. Appl. Math.*, 50:703–715, 1990.

[163] T. McMillen, J. Xin, Y. Yu, and A. Zlatos. Ballistic orbits and front speed enhancement for ABC flows. *SIAM J. Appl. Dyn. Syst.*, 15(3):1753–1782, 2016.

[164] K. R. Meyer, G. R. Hall, and D. Offin. *Introduction to Hamiltonian dynamical systems and the N-body problem*, volume 90 of *Applied Mathematical Sciences*. Springer, New York, second edition, 2009.

[165] M. Mimura, S.-I. Ei and Q. Fang. Effect of domain-shape on coexistence problems in a competition-diffusion system *Journal of Mathematical Biology*, 29(3): 219–237, 1991.

[166] M. Mimura and K. Kawasaki. Spatial segregation in competitive interaction-diffusion equations. *Journal of Mathematical Biology*, 9(1): 49–64, 1980.

[167] M. Mimura, Y. Nishiura, A. Tesei and T. Tsujikawa. Coexistence problem for two competing species models with density-dependent diffusion. *Hiroshima Mathematical Journal*, 14(2): 425–449, 1984.

[168] J. D. Mireles James. Quadratic volume-preserving maps: (un)stable manifolds, hyperbolic dynamics, and vortex-bubble bifurcations. *J. Nonlinear Sci.*, 23(4):585–615, 2013.

[169] J. D. Mireles James. Polynomial approximation of one parameter families of (un)stable manifolds with rigorous computer-assisted error bounds. *Indag. Math. (N.S.)*, 26(1):225–265, 2015.

[170] J. D. Mireles James. Fourier-Taylor Approximation of Unstable Manifolds for Compact Maps: Numerical Implementation and Computer-Assisted Error Bounds. *Foundations of Computational Mathematics*, 1–57, 2016.

[171] J. D. Mireles James and H. Lomelí. Computation of heteroclinic arcs with application to the volume preserving Hénon family. *SIAM J. Appl. Dyn. Syst.*, 9(3):919–953, 2010.

[172] J.D. Mireles James and K. Mischaikow. Rigorous a posteriori computation of (un)stable manifolds and connecting orbits for analytic maps. *SIAM J. Appl. Dyn. Syst.*, 2:957–1006, 2013.

[173] J.D. Mireles James and K. Mischaikow. Computational proofs in dynamics. *Encyclopedia of Applied and Computational Mathematics*, 288–295, 2015.

[174] K. Mischaikow, M. Mrozek and P. Pilarczyk. Graph approach to the computation of the homology of continuous maps *Found. Comp. Math.*, 5:199–229, 2005

[175] M. Mrozek. Topological invariants, multivalued maps and computer assisted proofs in dynamics. *Comput. Math. Appl.* 32: 83–104, 1996

[176] M. T. Nakao. A numerical approach to the proof of existence of solutions for elliptic problems. *Japan J. Appl. Math.*, 5(2): 313–332, 1988.

[177] M. T. Nakao. Numerical verification methods for solutions of ordinary and partial differential equations. *Numer. Funct. Anal. Optim.*, 22(3-4): 321–356, 2001.

[178] A. Neumaier and T. Rage. Rigorous chaos verification in discrete dynamical systems. *Phys. D*, 67(4):327–346, 1993.

[179] L.A. Peletier and W.C. Troy. Multibump periodic travelling waves in suspension bridges. *Proc. Roy. Soc. Edinburgh Sect. A*, 128(3):631–659, 1998.

[180] L.A. Peletier and W.C. Troy. *Spatial patterns*. Progress in Nonlinear Differential Equations and their Applications, 45. Birkhäuser Boston, Inc., Boston, MA, 2001. Higher order models in physics and mechanics.

[181] S. Oishi. Numerical verification method of existence of connecting orbits for continuous dynamical systems. *J.UCS*, 4(2):193–201 (electronic), 1998. SCAN-97 (Lyon).

[182] H. Osinga. Non-orientable manifolds of periodic orbits. In *International Conference on Differential Equations, Vol. 1, 2 (Berlin, 1999)*, pages 922–924. World Sci. Publ., River Edge, NJ, 2000.

[183] H. M. Osinga. Nonorientable manifolds in three-dimensional vector fields. *Internat. J. Bifur. Chaos Appl. Sci. Engrg.*, 13(3):553–570, 2003.

[184] J. M. Ortega. The Newton-Kantorovich theorem. *The American Mathematical Monthly*, 75(6): 658–660, 1968.

[185] K. J. Palmer. Exponential dichotomies, the shadowing lemma and transversal homoclinic points. In *Dynamics reported, Vol. 1*, volume 1 of *Dynam. Report. Ser. Dynam. Systems Appl.*, pages 265–306. Wiley, Chichester, 1988.

[186] L. Perko. *Differential equations and dynamical systems (Vol. 7)*. Springer Science & Business Media, 2013.

[187] M. Plum. Computer-assisted enclosure methods for elliptic differential equations *Linear Algebra and its Applications*, 324(1-3): 147–187, 2001.

[188] W. C. Rheinboldt. A Unified Convergence Theory for a Class of Iterative Processes. *SIAM Journal on Numerical Analysis*, 5(1): 42–63, 1968.

[189] N. Robertson, D. Sanders, P. Seymour and R. Thomas. The four-colour theorem. *J. Combin. Theory Ser. B*, 70(1): 2–44, 1997.

[190] S.M. Rump. *INTLAB - INTerval LABoratory.* Developments in Reliable Computing, Kluwer Academic Publishers, Dordrecht, pp. 77–104, 1999. `http://www.ti3.tuhh.de/rump/`.

[191] S.M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numer.*, 19: 287–449, 2010.

[192] K. Ryu and I. Ahn. Coexistence theorem of steady states for nonlinear self-cross-diffusion systems with competitive dynamics. *Journal of mathematical analysis and applications*, 283(1): 46–65, 2003.

[193] S. Santra and J. Wei. Homoclinic solutions for fourth order traveling wave equations. *SIAM J. Math. Anal.*, 41(5):2038–2056, 2009.

[194] N. Shigesada, K. Kawasaki and E. Teramoto. Spatial segregation of interacting species. *Journal of Theoretical Biology*, 79(1): 83–99, 1979.

[195] L. P. Shil'nikov. A case of the existence of a denumerable set of periodic motions. *Dokl. Akad. Nauk SSSR*, 160: 558–561, 1965.

[196] C. Simó. On the analytical and numerical approximation of invariant manifolds. In D. Benest and C. Froeschle, editors, *Modern Methods in Celestial Mechanics*, page 285, 1990.

[197] D. Stoffer and K.J. Palmer. Rigorous verification of chaotic behaviour of maps using validated shadowing. *Nonlinearity*, 12(6):1683–1698, 1999.

[198] R. Szczelina and P. Zgliczyński. A homoclinic orbit in a planar singular ode-a computer-assisted proof. *SIAM J. Appl. Dyn. Syst.*, 12(3):1541–1565, 2013.

[199] L. N. Trefethen. *Approximation theory and approximation practice*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2013.

[200] W. Tucker. A Rigorous ODE Solver and Smale's 14th Problem. *Found. Comput. Math.*, 2:53–117, 2002.

[201] W. Tucker. *Validated numerics: a short introduction to rigorous computations*. Princeton University Press, 2011.

[202] Y. Watanabe, K. Nagatou, M. Plum and M.T. Nakao. A computer-assisted stability proof for the Orr-Sommerfeld problem with Poiseuille flow. *Nonlinear Theory and Its Applications, IEICE*, 2(1): 123–127, 2011.

[203] D. Wilczak. Symmetric heteroclinic connections in the michelson system: a computer-assisted proof. *SIAM J. Appl. Dyn. Syst.*, 4(3)(electronic):489–514, 2005.

[204] D. Wilczak. The existence of shilnikov homoclinic orbits in the michelson system: a computer-assisted proof. *Found. Comput. Math.*, 6(4):495–535, 2006.

[205] Daniel Wilczak. Symmetric homoclinic solutions to the periodic orbits in the Michelson system. *Topol. Methods Nonlinear Anal.*, 28(1):155–170, 2006.

[206] Daniel Wilczak. Abundance of heteroclinic and homoclinic orbits for the hyperchaotic Rössler system. *Discrete Contin. Dyn. Syst. Ser. B*, 11(4):1039–1055, 2009.

[207] D. Wilczak and P. Zgliczyński. Heteroclinic connections between periodic orbits in planar restricted circular three-body problem - a computer-assisted proof. *Comm. Math. Phys.*, 234(1):37–75, 2003.

[208] D. Wilczak and P. Zgliczyński. Period doubling in the Rössler system—a computer-assisted proof. *Found. Comput. Math.*, 9(5):611–649, 2009.

[209] A. Wittig. *Rigorous High-Precision Enclosures of Fixed Points and Their Invariant Manifolds.* PhD thesis, Michigan State University, 2011.

[210] A. Wittig, M. Berz, J. Grote, K. Makino, and S. Newhouse. Rigorous and accurate enclosure of invariant manifolds on surfaces. *Regul. Chaotic Dyn.*, 15(2-3):107–126, 2010.

[211] K. Wójcik and P. Zgliczyński. On existence of infinitely many homoclinic solutions. *Monatsh. Math.*, 130(2):155–160, 2000.

[212] J. K. Wróbel and R. H. Goodman. High-order adaptive method for computing two-dimensional invariant manifolds of three-dimensional maps. *Commun. Nonlinear Sci. Numer. Simul.*, 18(7):1734–1745, 2013.

[213] J. Xin, Y. Yu, and A. Zlatos. Periodic orbits of the ABC flow with $A = B = C = 1$. *SIAM J. Math. Anal.*, 48(6):4087–4093, 2016.

[214] N. Yamamoto. A numerical verification method for solutions of boundary value problems with local uniqueness by Banach's fixed-point theorem. *SIAM J. Numer. Anal.*, 35: 2004–2013, 1998.

[215] N. Yamamoto. A simple method for error bounds of eigenvalues of symmetric matrices. *Linear Algebra and its Applications*, 324(1):227–234, 2001.

[216] P. Zgliczyński, On periodic points for systems of weakly coupled 1-dim maps. *Nonlinear Anal.*, 46(7): 1039–1062, 2001

[217] P. Zgliczynski. Rigorous numerics for dissipative partial differential equations. II. Periodic orbit for the Kuramoto-Sivashinsky PDE - a computer-assisted proof. *Found. Comput. Math.*, 4(2): 157– 185, 2004.

[218] P. Zgliczynski. $\mathcal{C}^1$ Lohner algorithm. *Foundations of Computational Mathematics*, 2(4): 429–465, 2008.

[219] P. Zgliczynski. Covering relations, cone conditions and the stable manifold theorem. *J. Differential Equations*, 246(5): 1774–1819, 2009.

[220] P. Zgliczyński and K. Mischaikow. Rigorous numerics for partial differential equations: the Kuramoto-Sivashinsky equation. *Found. Comput. Math.*, 1(3):255–288, 2001.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

UNIVERSITÉ LAVAL

Titre : Equations aux dérivées partielles et systèmes dynamiques appliqués à des problèmes issus de la physique et de la biologie

Mots clefs : Equations aux dérivées partielles • Systèmes dynamiques • Equations de coagulation-fragmentation • Estimations de moments • Preuves assistées par ordinateur • Validation a posteriori

Résumé : Cette thèse s'inscrit dans le vaste domaine des équations aux dérivées partielles et des systèmes dynamiques, et s'articule autour de deux sujets distincts. Le premier est relié à l'étude des équations de coagulation-fragmentation discrètes avec diffusion. En utilisant des lemmes de dualité, on établit de nouvelles estimations L^p pour des moments polynomiaux associés aux solutions, sous une hypothèse de convergence des coefficients de diffusion. Ces estimations sur les moments permettent ensuite d'obtenir de nouveaux résultats de régularité, et de démontrer qu'une fragmentation suffisamment forte peut empêcher la gelation dans le modèle incluant la diffusion. Le second sujet est celui des preuves assistées par ordinateur dans le domaine des systèmes dynamiques. On améliore et on applique une méthode basée sur le théorème du point fixe de Banach, permettant de valider a posteriori des solutions numériques. Plus précisément, on élargit le cadre d'application de cette méthode pour inclure des opérateurs avec un terme dominant linéaire tridiagonal, on perfectionne une technique permettant de calculer et de valider des variétés invariantes, et on introduit une nouvelle technique qui améliore de manière significative l'utilisation de l'interpolation polynomiale dans le cadre de ces méthodes de preuves assistées par ordinateur. Ensuite, on applique ces techniques pour démontrer l'existence d'ondes progressives pour l'équation du pont suspendu, et pour étudier les états stationnaires non homogènes d'un système de diffusion croisée.

Title: Partial differential equations and dynamical systems applied to problems coming from physics and biology

Keys words: Partial differential equations • Dynamical systems • Coagulation-fragmentation equations • Moments estimates • Computer-assisted proofs • A posteriori validation

Abstract: This thesis falls within the broad framework of partial differential equations and dynamical systems, and focuses more specifically on two independent topics. The first one is the study of the discrete coagulation-fragmentation equations with diffusion. Using duality lemma we establish new L^p estimates for polynomial moments of the solutions, under an assumption of convergence of the diffusion coefficients. These moment estimates are then used to obtain new results of smoothness and to prove that strong enough fragmentation can prevent gelation even in the diffusive case. The second topic is the one of computer-assisted proofs for dynamical systems. We improve and apply a method enabling to a posteriori validate numerical solutions, which is based on Banach's fixed point theorem. More precisely, we extend the range of applicability of the method to include operators with a dominant linear tridiagonal part, we improve an existing technique allowing to compute and validate invariant manifolds, and we introduce an new technique that significantly improves the usage of polynomial interpolation for a posteriori validation methods. Then, we apply those techniques to prove the existence of traveling waves for the suspended bridge equation, and to study inhomogeneous steady states of a cross-diffusion system.