


--	--	--

		
	<b><u>PROJET</u></b>  2 <sup>ème</sup> année	N° du projet : 7249  2 <sup>ème</sup> version Date : 12/06/2006
	Rapport intermédiaire <input type="checkbox"/>  Rapport final : <input checked="" type="checkbox"/>	Promo : 2007

Titre du projet: Méthodes de séparation d'opérateur pour la simulation numérique de milieux réactifs en évolution impliquant un large spectre d'échelles de temps.

Élèves participant au projet :  Adrien Auclert Miguel Gillot		Enseignant responsable :  Marc Massot
---	--	---

# Méthodes de séparation d'opérateurs pour la simulation numérique de milieux réactifs en évolution impliquant un large spectre d'échelles de temps

Adrien Auclert, Miguel Gillot

12 juin 2006

## Rapport de projet de 2ème année

### Résumé

Des domaines aussi divers que la combustion, la modélisation de l'environnement, la dynamique des populations utilisent couramment des modèles de réaction - diffusion impliquant un large spectre d'échelles de temps. La résolution numérique complète de ces derniers par des méthodes classiques, comme la méthode des lignes, est souvent très coûteuse en temps et en ressources de calcul. Pour pallier à ce problème, les méthodes de séparation d'opérateur, en anglais *splitting*, sont couramment utilisées, mais leur capacité à approcher la solution exacte doit être analysée. Nous présenterons l'analyse mathématique classique qui en est faite, puis mettrons en évidence la faillite de cette analyse pour les pas de temps élevés, lorsqu'interviennent de forts gradients spatiaux. Nous essaierons alors de comprendre la raison de cette "perte d'ordre", en particulier en analysant la structure de l'erreur. Dans notre étude, le modèle de génétique des populations de Kolmogorov-Petrovskii-Piskunov nous sert de référence, les résultats sont ensuite appliqués à un modèle simple de combustion.

### Abstract

Such diverse domains as combustion, environmental modeling or population dynamics all commonly use reaction-diffusion models implying a large time-scale range. Solving these completely using traditional numerical techniques - like the method of lines - often proves to be very time- and memory-consuming. In order to get around this difficulty, operator-splitting methods are of common use. However, their capacity to approximate the exact solution of the system must be analyzed. We will present the classical mathematical analysis, and then we will show how it fails for high splitting time steps, when the solution wave has steep slopes. We will then try to understand this "order reduction", notably by analysing the error structure. All along our study, the Kolmogorov-Petrovskii-Piskunov population genetics model will be used as a reference; the results will then be applied to a simple combustion model.

# Introduction

## Déroulement du projet

### Organisation du binôme

Le travail dans l'équipe s'est organisé de la manière suivante :

- Nous convenions chaque semaine d'une réunion, avec ou sans notre encadrant. Lors des réunions avec ce dernier, nous lui faisions un compte-rendu des résultats obtenus, nous discussions des difficultés rencontrées, tant théoriques que pratique, et nous décidions des orientations futures à donner au projet. Lorsque nous nous voyions en binôme, nous uniformisions nos connaissances des sujets sur lesquels nous avions travaillé individuellement, afin que l'un d'entre nous ne soit pas en retard par rapport à l'autre.
- En plus des réunions hebdomadaires, nous organisions des rendez-vous exceptionnels lorsque cela s'imposait.
- A chaque résultat trouvé, calcul réalisé ou partie du rapport finalisé, nous faisions un petit compte-rendu informel, transmis par mail, destiné à présenter à l'autre et à l'encadrant le travail accompli.
- Les tâches ont été réparties de manière à ce que chacun d'entre nous puisse avoir une vision globale du problème. Ainsi, chacun a travaillé sur une partie théorique et une partie numérique, et le binôme entier a participé à la réalisation du code servant à faire les calculs. Par exemple, Adrien a réalisé le code servant à réaliser les fonctions de phase, et Miguel en a fait le compte-rendu.

### Compétences acquises

Les compétences apprises au cours du projet sont nombreuses et sont à la fois d'ordre techniques et managériales.

- **Compétences techniques :**
  - Nous avons appris à utiliser des langages de programmation toujours très utilisés en sciences, à savoir le Fortran 77 et le Fortran 90, ainsi que l'écriture en  $\text{\LaTeX}$ , et la réalisation de graphiques avec GnuPlot
  - Nous avons beaucoup pratiqué la programmation dans des environnements de travail UNIX. Nous avons appris, entre autres, à utiliser les scripts UNIX.
  - Nous avons dorénavant une bonne expérience dans le domaine du calcul numérique, et nous comprenons bien ses problématiques et ses enjeux (estimation des erreurs d'arrondis, des erreurs d'approximation...)
  - Nous avons appris à effectuer des recherches bibliographiques et à analyser des articles scientifiques
  - Nous avons bien sûr une solide connaissance de la théorie des équations différentielles ordinaires.
  - Nous avons une vision globale des nombreux aspects d'un problème de recherche actuel (Raideur des systèmes réaction-diffusion, méthodes de splitting, perte d'ordre de ces méthodes, tentatives d'explication de la cause des pertes d'ordres...).

– **Compétences de gestion de projet :**

- Nous avons appris à nous adapter rapidement pour travailler avec des outils nouveaux (notamment les outils informatiques).
- Nous avons pu voir les difficultés du travail en groupe sur des sujets très techniques. Nous avons noté l'importance de la communication régulière et claire afin que tout le groupe soit au même niveau de connaissance.

## **Problèmes rencontrés**

Dans la mesure où nous travaillions sur des sujets de recherches actuels, les notions abordées étaient parfois difficiles à appréhender. Nous avons été confrontés à une façon de pratiquer les mathématiques très différente de celle que nous connaissions, d'une part parce que les domaines abordés (analyse numérique, équations aux dérivées partielles, géométrie différentielle) ne nous étaient pas connus, et d'autre part parce que les domaines de recherche, par essence, n'ont pas de théorie unificatrice et cohérente permettant de tout expliquer. Nous avons été, à de nombreuses reprises, déroutés par cet aspect.

Par ailleurs, nous avons rencontré des difficultés dans l'implémentation de nos codes de simulation numérique, qui étaient notamment dus au fait que nous n'avions pas de base théorique en programmation et aucune expérience du langage - FORTRAN - que nous avons utilisé. Nous avons dû comprendre le langage au fur et à mesure, à partir d'un code qui nous était fourni, ce qui n'a pas été une tâche aisée. La question des formats de fichier de sortie nous a, à cet égard, causé de très nombreux problèmes résultant en un très grand nombre d'heures perdues.

Si l'objectif principal qui figure dans le rapport intermédiaire du 14/12/2005, à savoir la découverte du monde de la recherche en mathématiques appliquées par l'étude d'un problème concret, a été pleinement réalisé, nous n'avons pas eu le temps de réaliser l'ensemble des points prévus avec l'encadrant (il nous manque en fait le dernier point, à savoir les applications à la chimie complexe) dans la mesure où il aurait parfois fallu y travailler à plein temps... ce qui n'est pas toujours possible avec les exigences de la scolarité. La rédaction des rapports en particulier est extrêmement longue, si l'on veut obtenir quelque chose de présentable.

## **Bilan des objectifs**

– **Objectifs prévus et réalisés :**

- Bibliographie réalisée sur le sujet des équations différentielles ordinaires et sur les méthodes de splitting
- Prise en main des logiciels LaTeX et du langage de programmation Fortran 77
- Maîtrise complète de la résolution de l'équation KPP (étude des erreurs de splitting, de vitesse, étude dans le plan de phase, perte d'ordre)
- Étude des méthodes de splitting appliquées à la combustion (obtention de résultats satisfaisants)

– **Objectifs non réalisés :**

- Étude des méthodes de splitting appliquées à la chimie complexe.

## **Les projets de recherche à l'interface entre plusieurs disciplines**

### **Introduction**

Au cours de notre étude de la résolution d'équations de combustion par des méthodes de splitting, nous avons été confrontés au monde de la recherche à l'interface entre les mathématiques appliquées et la combustion. Nous nous sommes alors demandés quelles sont les conditions de réussite d'un tel projet. Cette étude tente d'en faire l'inventaire. Elle montre particulièrement l'importance de la communication entre les différentes parties. Elle se fonde sur les entretiens réalisés avec des chercheurs travaillant à l'interface : notre encadrant Marc Massot, professeur à l'ECP, Frédérique Laurent-Nègre, chargée de recherche au CNRS, basée au laboratoire EM2C de l'ECP, et Violaine Louvet, Ingénieur de recherche à l'interface mathématiques appliquées - applications informatiques au CNRS, basée à l'institut Camille Jordan (Université de Lyon).

### **La présentation des projets et l'expression des besoins**

Le besoin de la recherche à l'interface naît de la complexité croissante des connaissances existantes dans chacun des grands domaines de recherche actuels. Les chercheurs ont donc tout à gagner à travailler avec des personnes d'autres disciplines.

Ces projets ne sont profitables à tous que si les chercheurs qui y travaillent ont un intérêt commun. L'expérience montre que les collaborations où les parties travaillent sur des problématiques qui ne sont pas communes n'aboutissent pas en général. Un exemple de collaboration manquée est celui du mathématicien zélé qui, dans un tel projet de recherche, voit dans le sujet la possibilité d'établir des résultats de mathématiques très puissants mais qui n'ont aucune utilité directe pour le spécialiste en combustion ; les rencontres se font alors rares dans la mesure où ce dernier n'est pas en mesure de contribuer au travail de mathématiques, non en adéquation avec ses propres recherches.

Lors de l'initiation d'un projet, les chercheurs à l'interface doivent donc s'assurer qu'ils vont travailler sur un sujet commun, et qu'ils ont une certaine curiosité pour les domaines des autres.

### **L'évolution des projets**

Au cours de l'avancée des projets, des problèmes peuvent se poser de part et d'autres de l'interface.

A l'interface médecine-mathématiques par exemple, les mathématiciens ont des impératifs de modélisation larges (pour garantir la généralité des résultats qu'ils calculeront, ils ont besoin de beaucoup de données) tandis que les médecins veulent surtout des résultats visuellement jolis (des représentations en trois dimension par exemple) sans se rendre compte qu'on perd forcément en rapidité de calcul machine et que de tels développements nécessitent beaucoup de temps pour les réaliser. Des réunions d'information et de synthèse régulières sont primordiales pour que les parties discutent des problèmes rencontrés dans l'avancement.

## **Structures favorisant l'établissement des projets à interfaces multiples**

Il existe à l'heure actuelle des initiatives qui favorisent les travaux de recherche à l'interface (certains laboratoires de la section 10 du CNRS orientés milieux fluides et réactifs accueillent des chercheurs spécialisés en mathématiques appliquées, comme par exemple le laboratoire EM2C de l'Ecole Centrale Paris, et réciproquement). L'évolution des projets à l'interface est nettement aidée lorsque que les chercheurs sont réunis dans le même laboratoire. Les échanges et rencontres sont ainsi facilitées, comme le rappelle Frédérique Laurent. Les rencontres entre les chercheurs de différents domaines se font lors des séminaires pluridisciplinaires, des visites de laboratoire, de conférences ou de journées organisées pour faire se rencontrer deux communautés (Ex : quand mathématiciens et financiers se rencontrent, organisé par la SMAI, Société de Mathématiques Appliquées et Industrielles).

## **Conclusion**

La motivation commune et la curiosité envers le projet de l'autre partie sont donc des facteurs clés de l'initiation d'un projet à l'interface. Au cours de la réalisation, des rencontres de mises au point régulières sont nécessaires pour permettre de recentrer les intérêts de chacun et de déterminer les limites dans lesquelles doivent rester le projet. Le cadre idéal pour le succès de la recherche à l'interface est le travail dans une même infrastructure, afin que les échanges soient simplifiés.

## Remerciements

Nous souhaitons adresser nos remerciements à Stéphane Descombes pour le temps qu'il nous a accordé à préparer et à nous présenter des exposés très clairs, et pour sa patience et sa pédagogie.

Nous voulons également remercier Violaine Louvet qui nous a expliqué ses recherches, et grâce à qui nous avons pu comprendre ce qu'était un programme "bien écrit".

Nous remercions également Frédérique Laurent qui nous a fait part de son expérience de recherche au laboratoire EM2C.

Nous tenons enfin à remercier chaleureusement notre encadrant, Marc Massot, pour nous avoir permis de réaliser ce projet, pour sa disponibilité et tout le temps qu'il nous a consacré, pour ses conseils et remarques ainsi que pour nous avoir permis à tous les deux de préciser notre projet professionnel.

## Plan

Dans une première partie, nous présenterons les systèmes de réaction-diffusion dans leur généralité, ainsi que les deux modèles que nous utiliserons : le modèle de Kolmogorov-Petrovskii-Piskunov et un modèle simplifié de combustion. Nous introduirons aussi la notion de raideur.

Dans une seconde partie, nous présenterons les méthodes numériques que nous avons utilisées : la méthode des lignes pour la discrétisation spatiale, la méthode BDF utilisée par le solveur LSODE. Nous présenterons ensuite les méthodes de séparation d'opérateurs, en détaillant les résultats théoriques dans le cas linéaire et en présentant une introduction à la théorie de la perte d'ordre dans ce même cas. Enfin, nous présenterons le code de calcul numérique que nous avons écrit en Fortran 77 en insistant sur les algorithmes les plus importants.

Dans une troisième partie, nous présenterons les résultats de ces calculs en ce qui concerne l'ordre de l'erreur et la vitesse de l'onde.

Dans une dernière partie, nous présenterons la synthèse de ces résultats et ferons l'étude des diagrammes de phase, ce qui nous permettra d'aborder l'étude de la structure de l'erreur.

## Table des matières

<b>Introduction</b>	<b>2</b>
Déroulement du projet . . . . .	2
Les projets de recherche à l'interface entre plusieurs disciplines . . . . .	4
Remerciements . . . . .	6
Plan . . . . .	7
<b>I   Systèmes réaction - diffusion</b>	<b>10</b>
<b>1   Modélisation d'une réaction chimique autocatalytique</b>	<b>10</b>
1.1 Présentation du problème . . . . .	10
1.2 Mise en équation . . . . .	11
1.3 Résolution . . . . .	11
1.4 Etude dans le plan de phase . . . . .	12
<b>2   Modèle simple de combustion</b>	<b>14</b>
2.1 Modèle et hypothèses . . . . .	14
2.2 Equations et adimensionnement . . . . .	15
2.3 Une solution numérique . . . . .	17
<b>3   Raideur des systèmes réaction - diffusion</b>	<b>20</b>
3.1 Notion de raideur . . . . .	20
3.2 Dispersion des valeurs propres du laplacien discrétisé . . . . .	21
3.3 Influence de la condition initiale sur la raideur du système . . . . .	21
<b>II  Méthodes numériques de résolution</b>	<b>26</b>



<b>4</b>	<b>La méthode des lignes (MOL)</b>	<b>26</b>
4.1	Principe . . . . .	26
4.2	Exemple . . . . .	26
<b>5</b>	<b>Le solveur LSODE</b>	<b>27</b>
5.1	Introduction . . . . .	27
5.2	La méthode BDF . . . . .	28
5.3	Formulation canonique du problème . . . . .	28
5.4	Schémas prédicteur-correcteur . . . . .	29
5.5	Méthode de Newton-Raphson . . . . .	29
5.6	Résumé de la méthode de Newton-Raphson . . . . .	30
5.7	Formulation matricielle . . . . .	30
5.8	Résumé de la formulation matricielle . . . . .	31
5.9	Estimation et contrôle de l'erreur locale . . . . .	31
<b>6</b>	<b>Les méthodes de splitting</b>	<b>32</b>
6.1	Motivations . . . . .	32
6.2	Première approche . . . . .	32
6.2.1	Méthodes de Lie . . . . .	33
6.2.2	Méthodes de Strang . . . . .	33
6.3	Splitting et calcul numérique . . . . .	34
6.4	Splitting et perte d'ordre . . . . .	34
6.4.1	Expression intégrale de l'erreur locale . . . . .	34
6.4.2	Evaluation de la perte d'ordre quand la condition initiale présente de forts gradients . . . . .	35
6.5	Passage de l'erreur locale à l'erreur globale . . . . .	39
6.5.1	Introduction . . . . .	39
6.5.2	Schéma de splitting . . . . .	40
6.5.3	Convergence du schéma de splitting . . . . .	40
6.5.4	Cumul des erreurs locales . . . . .	41
<b>7</b>	<b>Description des algorithmes</b>	<b>41</b>
7.1	Déclaration des variables . . . . .	41
7.2	Intégration quasi-exacte du problème discrétisé avec le solveur LSODE . . . . .	42
7.3	Intégration par des méthodes de splitting du problème discrétisé avec le solveur LSODE . . . . .	43
7.4	Calcul de la vitesse . . . . .	47
7.4.1	Principe de l'algorithme . . . . .	47
7.4.2	Algorithme complet . . . . .	47
7.5	Calcul des erreurs . . . . .	48
7.5.1	Norme infinie . . . . .	48
7.5.2	Norme $L_2$ . . . . .	48
7.5.3	Algorithme du calcul des erreurs . . . . .	48
7.6	Etude dans le plan de phase . . . . .	49

7.6.1	Les variables utilisées . . . . .	49
7.6.2	Calcul des dérivées . . . . .	49
7.6.3	Calcul des différences entre les diagrammes de phase . . . . .	50
7.6.4	Algorithme complet . . . . .	50
<b>III</b>	<b>Resultats</b>	<b>52</b>
<b>8</b>	<b>Modèle KPP</b>	<b>52</b>
8.1	Présentation du programme . . . . .	52
8.2	Convergence vers la solution analytique . . . . .	52
8.3	Influence de la discrétisation par la méthode des lignes . . . . .	55
8.4	Etude des résultats obtenus en résolvant par splitting (cas non raide) . . . . .	58
8.4.1	Splitting de Lie RD . . . . .	58
8.4.2	Splitting de Lie DR . . . . .	62
8.4.3	Splitting de Strang RDR . . . . .	65
8.4.4	Splitting de Strang DRD . . . . .	68
8.4.5	Conclusion . . . . .	70
8.5	Introduction de raideur dans le système : pertes d'ordre en splitting . . . . .	70
8.5.1	Splitting de Lie . . . . .	72
8.5.2	Splitting de Strang . . . . .	78
8.5.3	Conclusion . . . . .	84
<b>9</b>	<b>Modèle de combustion</b>	<b>84</b>
9.1	Procédure . . . . .	84
9.2	Etude de l'erreur de splitting - cas RDR . . . . .	85
9.3	Etude de l'erreur de splitting - cas RD . . . . .	88
9.4	Conclusion . . . . .	93
<b>IV</b>	<b>Synthèse</b>	<b>94</b>
<b>10</b>	<b>Etude du plan de phase</b>	<b>94</b>
10.1	Synthèse des résultats acquis . . . . .	94
10.2	Etude du plan de phase de KPP non raide . . . . .	95
10.2.1	Méthode RDR . . . . .	95
10.2.2	Autres méthodes de splitting . . . . .	97
10.3	Etude du plan de phase de KPP raide . . . . .	101
10.4	Etude du plan de phase du modèle de combustion . . . . .	106
10.5	Conclusion . . . . .	109
	<b>Conclusion</b>	<b>110</b>
	<b>Annexes : théorie des EDO, méthodes numériques de résolution</b>	<b>111</b>

# Première partie

## Systèmes réaction - diffusion

Dans ce projet, l'étude porte sur des systèmes réaction-diffusion dont l'équation la plus générale est :

$$\frac{\partial U}{\partial t}(x, t) = M \Delta U(x, t) + f(U(x, t)), \quad x \in \Omega, \quad t \in \mathbb{R}^+ \quad (1)$$

Où  $U : \Omega \times \mathbb{R}^+ \rightarrow \mathbb{R}^n$ ,  $M \in \mathcal{M}_n(\mathbb{R})$  et  $f$  est une fonction, en général non linéaire, de  $\mathbb{R}^n$  dans lui-même. L'équation est posée sur un ouvert  $\Omega$  de  $\mathbb{R}^d$  et est complétée par des conditions sur le bord.

Ces systèmes décrivent des phénomènes très variés : dès 1937, ils ont été utilisés par Fisher et par Kolmogorov, Petrovski et Piskunov pour décrire la propagation d'un gène mutant au sein d'une population (cf. [5]), mais on les retrouve aussi en chimie complexe, en combustion, en modélisation de l'environnement, en dynamique des populations, etc.

Sous certaines conditions, dont l'étude est faite dans [2] ou [9] par exemple, ces systèmes - paraboliques - peuvent posséder des solutions de type *onde progressive*, c'est-à-dire du type

$$U(x, t) = \varphi(x \cdot e - ct)$$

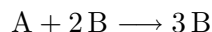
Ce sont des ondes se propageant à la vitesse  $c$  dans la direction  $e$ . Les systèmes que nous étudions entrent dans cette catégorie.

## 1 Modélisation d'une réaction chimique autocatalytique

L'objectif de ce paragraphe est d'étudier en détail l'évolution des concentrations de deux substances dans un milieu réactif dans lequel a lieu une réaction chimique ainsi que de la diffusion. Nous verrons que les concentrations du réactif et du produit vérifient une équation de réaction-diffusion dont la solution est une onde progressive.

### 1.1 Présentation du problème

On considère un milieu réactif contenu dans un tube, pouvant contenir deux substances A et B dont on notera les concentrations respectivement  $a$  et  $b$ . Pour des commodités de modélisation, on considérera que le milieu est unidimensionnel et de dimension infinie (approximation valable si le tube est suffisamment grand). Les substances A et B peuvent réagir suivant le schéma



La cinétique est décrite par la constante de réaction  $k$ . La vitesse de cette réaction est donc  $kab^2$ . A et B diffusent également dans le milieu. On suppose qu'il s'agit de substances ayant un même coefficient de diffusion  $D$ .

## 1.2 Mise en équation

Une étude de la cinétique du problème nous montre que les équations d'évolution temporelle des concentrations s'écrivent comme la somme des contributions de la diffusion et de la réaction

$$\frac{\partial a}{\partial t} = D \frac{\partial^2 a}{\partial r^2} - kab^2 \quad (2)$$

$$\frac{\partial b}{\partial t} = D \frac{\partial^2 b}{\partial r^2} + kab^2 \quad (3)$$

On suppose des conditions aux limites au bord du tube (autrement dit pour  $|x| = \infty$ )

$$\lim_{x \rightarrow -\infty} a = a_0 ; \quad \lim_{x \rightarrow -\infty} b = 0 \quad (4)$$

$$\lim_{x \rightarrow +\infty} a = 0 ; \quad \lim_{x \rightarrow +\infty} b = a_0 \quad (5)$$

En sommant l'équation (2) et l'équation (3) on obtient

$$\frac{\partial(a+b)}{\partial t} = D \frac{\partial^2(a+b)}{\partial r^2} \quad (6)$$

La somme des concentrations vérifie donc l'équation de la chaleur à une dimension. Si l'on ajoute aux conditions aux limites (4) et (5) une condition initiale telle que  $a(0, r) + b(0, r) = a_0$  pour tout  $r \in \mathbb{R}$ , la solution de (6) est alors la constante :

$$\forall(t, r) \quad a + b = a_0$$

En remplaçant  $a$  dans l'équation (2), en divisant par  $a_0$  et en posant  $\beta = \frac{b}{a_0}$  on obtient :

$$\frac{\partial \beta}{\partial t} = D \frac{\partial^2 \beta}{\partial r^2} + k\beta^2(1 - \beta) \quad (7)$$

Cette équation est connue sous le nom d'équation de Kolmogorov-Petrovskii-Piskunov (KPP).

## 1.3 Résolution

Tous d'abord, réécrivons l'équation (7) sous une forme adimensionnée :

$$\frac{\partial \beta}{\partial \tau} = \frac{\partial^2 \beta}{\partial x^2} + \beta^2(1 - \beta) \quad (8)$$

avec  $\tau = kt$  et  $x = (\frac{k}{D})^{1/2}r$  Pour déterminer l'équation que vérifie le profil de l'onde, on cherche une solution de sous la forme d'une onde progressive de vitesse constante  $c$ . On pose pour cela  $\beta(x, \tau) = \beta(z)$  avec  $z = x - c\tau$ . On trouve alors pour la concentration réduite  $\beta$  l'équation suivante :

$$\frac{d^2 \beta}{dz^2} + \frac{d\beta}{dz} + \beta^2(1 - \beta) = 0$$

avec les conditions aux limites

$$\lim_{x \rightarrow -\infty} \beta = 0 ; \quad \lim_{x \rightarrow +\infty} \beta = 1$$

Notons qu'il s'agit d'une équation différentielle ordinaire. On la résout en remarquant qu'une solution doit être comprise entre 0 et 1 (pour éviter des concentrations négatives), que le gradient de concentration doit certainement tendre vers zéro aux infinis (car les concentrations tendent vers des valeurs finies), et que par conséquent le gradient doit être négatif pour toute valeur finie de  $z$ . Une forme simple de  $\beta$  qui satisfait toutes ces conditions est l'équation parabolique suivante

$$\frac{d\beta}{dz} = -\alpha\beta(1 - \beta) \quad (9)$$

Un calcul long mais simple permet alors de déterminer la vitesse  $c$

$$c = \alpha = \frac{1}{\sqrt{2}}$$

ainsi que la solution exacte en onde progressive

$$\beta(z) = \frac{\exp(-\frac{1}{\sqrt{2}}(z - z_0))}{1 + \exp(-\frac{1}{\sqrt{2}}(z - z_0))}$$

En revenant aux données initiales et prenant de plus  $a_0 = 1$ , on en déduit :

$$v = \frac{dr}{dt} = \frac{1}{\sqrt{2}}(kD)^{\frac{1}{2}} \quad (10)$$

$$\beta(r, t) = \frac{\exp(-\frac{1}{\sqrt{2}}(\frac{k}{D})^{\frac{1}{2}}(r - r_0 - vt))}{1 + \exp(-\frac{1}{\sqrt{2}}(\frac{k}{D})^{\frac{1}{2}}(r - r_0 - vt))}$$

Un calcul montre à partir de cette expression que la pente la plus forte de la solution exacte s'écrit

$$(\frac{\partial\beta}{\partial r})_{max} = -\frac{1}{\sqrt{32}}(\frac{k}{D})^{\frac{1}{2}} \quad (11)$$

La figure 1 montre l'allure de cette solution à huit intervalles de temps régulièrement espacés, dans le cas  $k = 1$  et  $D = 1$ .

#### 1.4 Etude dans le plan de phase

L'équation aux dérivées partielles (7) devient, lorsqu'on cherche à l'écrire à l'aide de  $z = x - ct$ , une équation différentielle ordinaire :

$$D\frac{d^2\beta}{dz^2} + c\frac{d\beta}{dz} + k\beta^2(1 - \beta) = 0 \quad (12)$$

Il est intéressant d'étudier cette équation dans son plan de phase  $(\beta, \frac{d\beta}{dz})$ . Si l'on pose :

$$x_1 = \beta \quad x_2 = \frac{d\beta}{dz}$$

on a d'après (12) :

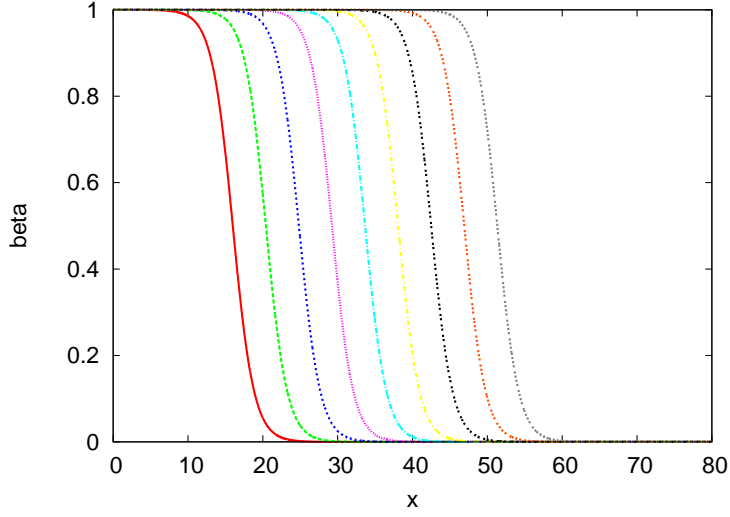


FIG. 1 – Solution exacte de l'équation de KPP ( $k = 1$ ,  $D = 1$ )

$$\frac{dx_1}{dz} = x_2 \quad \frac{dx_2}{dz} = -\frac{c}{D}x_2 - \frac{k}{D}x_1^2(1 - x_1)$$

Le système admet donc deux points stationnaires qui sont

$$(x_1 = 0, x_2 = 0) \quad \text{et} \quad (x_1 = 1, x_2 = 0)$$

Intéressons-nous au linéarisé tangent en ces points : on a

$$J_{(0,0)} = \begin{pmatrix} 0 & 1 \\ 0 & -\frac{c}{D} \end{pmatrix} \quad \text{et} \quad J_{(0,1)} = \begin{pmatrix} 0 & 1 \\ \frac{k}{D} & -\frac{c}{D} \end{pmatrix}$$

$J_{(0,0)}$  a donc pour valeurs propres  $-\frac{c}{D} < 0$  et 0. La direction du vecteur propre associée à  $-\frac{c}{D}$ , à savoir celle du vecteur  ${}^t \begin{pmatrix} 1 & -\frac{c}{D} \end{pmatrix}$  est donc attractive. Or, il se trouve que l'on connaît l'expression explicite de la phase : d'après (9)

$$x_2 = -\left(\frac{k}{2D}\right)^{1/2}(x_1)(1 - x_1)$$

La dérivée de cette fonction en 0 étant  $-(\frac{k}{2D})^{1/2}$ , la courbe arrive sur le point 0 dans la direction  ${}^t \begin{pmatrix} 1 & -(\frac{k}{2D})^{1/2} \end{pmatrix} = {}^t \begin{pmatrix} 1 & -\frac{c}{D} \end{pmatrix}$  car  $c = (\frac{kD}{2})^{1/2}$  selon (10). Le point (0,0) est donc un attracteur de la dynamique du système.

De la même manière,  $J_{(1,0)}$  a pour valeurs propres

$$-\frac{c}{2D} - \left(\frac{c^2}{4D^2} + \frac{k}{D}\right)^{1/2} < 0 \quad \text{et} \quad -\frac{c}{2D} + \left(\frac{c^2}{4D^2} + \frac{k}{D}\right)^{1/2} > 0$$

C'est donc un col hyperbolique, et l'on prouverait de même que le point (1,0) est en fait un point répulsif.

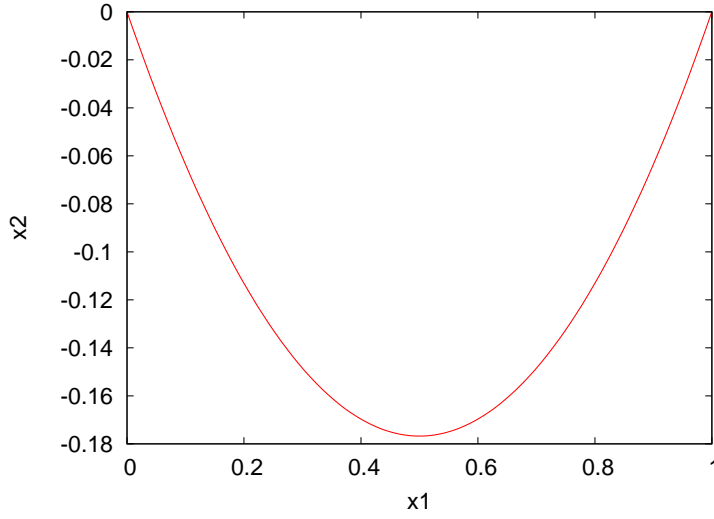


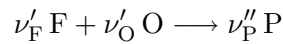
FIG. 2 – Diagramme de phase de l'onde progressive de KPP ( $k = 1, D = 1$ ). Le point  $(0, 0)$  est un attracteur, le point  $(1, 0)$  un répulseur.

## 2 Modèle simple de combustion

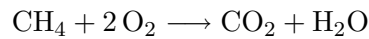
Nous nous intéressons maintenant à un modèle unidimensionnel de combustion, de réaction chimique très simple. Nous allons voir qu'au moyen de certaines hypothèses, la température et les concentrations de fuel et d'oxydant vérifient un système de réaction-diffusion dont la solution est encore une onde progressive. Ce modèle nous permettra d'obtenir une idée des caractéristiques principales de la flamme. Le but est de pouvoir appliquer les résultats à une combustion avec chimie complexe.

### 2.1 Modèle et hypothèses

On s'intéresse à une réaction entre fuel  $F$  et oxydant  $O$  de type :



Un exemple en est la réaction globale de combustion du méthane (sans les intermédiaires réactionnels) :



On appelle  $\phi$  la richesse de la flamme :

$$\phi = \left( \frac{Y_F^1}{Y_O^1} \right) / \left( \frac{Y_F}{Y_O} \right)_{st} = \frac{\nu'_O W_O Y_F^1}{\nu'_F W_F Y_O^1} = \frac{\nu'_O X_F^1}{\nu'_F X_O^1} \quad (13)$$

Où  $X_i^1, Y_i^1, W_i$  désignent respectivement les fractions molaires et massiques initiales, et la masse molaire de l'espèce  $i$  ( $Y_i = X_i \frac{W_i}{W_{tot}}$ ).

Les hypothèses effectuées sont les suivantes :

- On s'intéresse à un modèle de flamme unidimensionnel à faible nombre de Mach et pression constante.
- Les capacités calorifiques sont prises égales à  $c_p$  pour toutes les espèces.
- On suppose  $c_p$  et la conductivité thermique  $\lambda$  indépendantes de la température.
- Le terme de source extérieure de chaleur est nul

A ces hypothèses classiques on ajoute le fait que :

- La densité est constante

Ceci est bien entendu une grosse approximation (l'équation d'état des gaz parfaits impose normalement pour ce modèle isobare  $\rho_1 T_1 = \rho_2 T_2$ ) mais ce premier modèle permet déjà de trouver des résultats intéressants.

## 2.2 Equations et adimensionnement

Les hypothèses précédentes conduisent aux équations (voir [6]) :

$$\begin{aligned} \rho c_p \frac{\partial T}{\partial t} - \lambda \frac{\partial^2 T}{\partial x^2} &= -Q W_F \dot{\omega}_F \\ \rho \frac{\partial Y_F}{\partial t} - \rho D_F \frac{\partial^2 Y_F}{\partial x^2} &= W_F \dot{\omega}_F \\ \rho \frac{\partial Y_O}{\partial t} - \rho D_O \frac{\partial^2 Y_O}{\partial x^2} &= W_O \dot{\omega}_O \end{aligned}$$

Avec  $\dot{\omega}_i$  le taux de production molaire de l'espèce  $i$ ,  $D_i$  sa diffusivité d'espèce et  $Q$  la chaleur spécifique de la réaction. On obtient donc un système de réaction diffusion (la matrice de diffusion est ici diagonale). Effectuons-en un adimensionnement.

On définit les fractions massiques et la température adimensionnées :

$$\overline{Y}_F = \frac{Y_F}{Y_F^1} \quad \overline{Y}_O = \frac{Y_O}{Y_O^1} \quad \Theta = \frac{T - T_1}{\frac{Q}{c_p} Y_F^1} = \frac{T - T_1}{T_2 - T_1}$$

Où  $T_2$  est la température adiabatique de flamme déterminée par l'équation :

$$c_p(T_2 - T_1) = Q Y_F^1$$

Sachant que

$$\dot{\omega}_O = \frac{\nu'_O}{\nu'_F} \dot{\omega}_F$$

on a donc en posant  $a = \lambda/(\rho c_p)$  la diffusivité thermique :



$$\begin{aligned}
\frac{\partial \Theta}{\partial t} - a \frac{\partial^2 \Theta}{\partial x^2} &= -\frac{W_F \omega_F}{\rho Y_F^1} \\
\frac{\partial \overline{Y_F}}{\partial t} - D_F \frac{\partial^2 \overline{Y_F}}{\partial x^2} &= \frac{W_F \omega_F}{\rho Y_F^1} \\
\frac{\partial \overline{Y_O}}{\partial t} - D_O \frac{\partial^2 \overline{Y_O}}{\partial x^2} &= \frac{W_F \omega_F}{\rho Y_F^1} \phi
\end{aligned}$$

Où la richesse du mélange  $\phi$  est définie en (13).

Adimensionnons maintenant le temps et l'espace : on note  $L$  la distance de diffusion de chaleur dans la flamme et on définit :

$$\bar{x} = \frac{x}{L} \quad \bar{t} = \frac{t}{L^2/a}$$

Posons aussi  $Le_i = a/D_i$  le nombre de Lewis de l'espèce  $i$ . Les équations deviennent :

$$\frac{\partial \Theta}{\partial \bar{t}} - \frac{\partial^2 \Theta}{\partial \bar{x}^2} = -\frac{W_F \omega_F}{\rho Y_F^1} \frac{L^2}{a} \quad (14)$$

$$\frac{\partial \overline{Y_F}}{\partial \bar{t}} - \frac{1}{Le_F} \frac{\partial^2 \overline{Y_F}}{\partial \bar{x}^2} = \frac{W_F \omega_F}{\rho Y_F^1} \frac{L^2}{a} \quad (15)$$

$$\frac{\partial \overline{Y_O}}{\partial \bar{t}} - \frac{1}{Le_O} \frac{\partial^2 \overline{Y_O}}{\partial \bar{x}^2} = \frac{W_F \omega_F}{\rho Y_F^1} \frac{L^2}{a} \phi \quad (16)$$

Sachant que  $\omega_F = \nu'_F \dot{\tau}$  où  $\dot{\tau}$ , taux d'avancement de la réaction, suit une loi de type Arrhénius :

$$\dot{\tau} = B_1 T^\beta e^{-\frac{T_a}{T}} \left( \frac{\rho \overline{Y_F} Y_F^1}{W_F} \right) \left( \frac{\rho \overline{Y_O} Y_O^1}{W_O} \right)^{1/2}$$

Où

$$T = T_1 + \frac{Q}{c_p} Y_F^1 \Theta$$

On voit donc que le terme source n'est pas du tout linéaire.

Les conditions aux limites sont

$$\begin{array}{llll}
\overline{Y_F} = 1 & \overline{Y_O} = 1 & \Theta = 0 & \text{en } \bar{x} = -\infty \\
\overline{Y_F} = 0 & \overline{Y_O} = 0 & \Theta = 1 & \text{en } \bar{x} = +\infty
\end{array}$$

Notons quelques cas particuliers :

- Si  $Le_F = 1$ , en effectuant la somme des équations (14) et (15), on voit que  $\Theta + \overline{Y_F}$  vérifie l'équation de la chaleur avec la condition initiale  $\forall x \in \mathbb{R}, \Theta(x, t=0) + \overline{Y_F}(x, t=0) = 1$ . Il résulte de ceci et de la compatibilité des conditions aux limites que  $\Theta + \overline{Y_F} = 1$  à chaque instant.

Longueur de diffusion de la chaleur ( $m$ )	$L$	$4.10^{-3}$
Diffusivité thermique ( $m^2.s^{-1}$ )	$a$	$2,26.10^{-5}$
Nombre de Lewis du Fuel	$Le_F$	1
Richesse du mélange	$\phi$	0,8
Pourcentage massique initial du mélange	$Y_M^1$	0,35
Température initiale ( $K$ )	$T_1$	300
Chaleur de réaction / capacité calorifique ( $K$ )	$Q/c_p$	34550
Facteur de fréquence ( $USI$ )	$B_1$	$1.10^7$
Température d'activation ( $K$ )	$T_a$	10055
Taille du domaine ( $m$ )	$x_m$	0,05
Nombre de points de discrétisation	$n$	4001
Intervalle de temps de résolution ( $s$ )	$T$	0,04

TAB. 1 – Paramètres du modèle de combustion, issus de [6]

- Si de plus  $Le_O = 1$  et  $\phi = 1$ , en faisant (15)-(16) on voit de même que  $\overline{Y_F} = \overline{Y_O}$  à tout instant.

Notons aussi qu'il suffit de se donner la fraction molaire du mélange fuel-oxygène  $Y_M^1 = Y_F^1 + Y_O^1$  pour obtenir  $Y_F^1$  et  $Y_O^1$ . En effet, grâce à la définition de  $\phi$  en (13) on obtient que

$$Y_F^1 = \frac{Y_M^1}{1 + \frac{1}{\phi} \frac{\nu'_O W_O}{\nu'_F W_F}} \quad \text{et} \quad Y_O^1 = Y_M^1 - Y_F^1$$

## 2.3 Une solution numérique

La partie (9) explique comment est effectuée la résolution numérique des équations ci-dessus. Donnons-en juste quelques résultats qui justifieront l'emploi de ce modèle. Les données sont celles de la table 2.3, issues de [6]. On en tire en particulier (combustion du méthane :  $\nu'_O = 2$ ,  $W_O = 0,032kg.mol^{-1}$ ,  $\nu'_F = 1$ ,  $W_F = 0,016kg.mol^{-1}$ ) :

$$\begin{aligned} Y_F^1 &= \frac{0,35}{1 + \frac{1}{0,8} \frac{2 \times 0,032}{1 \times 0,016}} = 0,0583 \\ Y_O^1 &= 0,35 - 0,0583 = 0,2917 \\ T_2 &= T_1 + \frac{Q}{c_p} Y_F^1 = 300 + 34550 \times 0,0583 = 2308K \end{aligned}$$

Sur les figures suivantes, on a tracé  $\Theta$ ,  $\overline{Y_F}$  et  $\overline{Y_O}$  en fonction de  $x$  à 8 intervalles de temps régulièrement répartis entre 0 et  $T$ . On constate ainsi que la solution du système est une onde, semblable à l'onde de KPP, de pente extrêmement raide, ce qui correspond aux profils usuels de flamme. La relation  $\Theta(x) + \overline{Y_F}(x) = 1$  est vérifiée car  $Le_F = 1$ .

Le programme calcule une vitesse adimensionnée d'onde  $\frac{dx}{dt} \simeq 73,723$  (à peu près identique

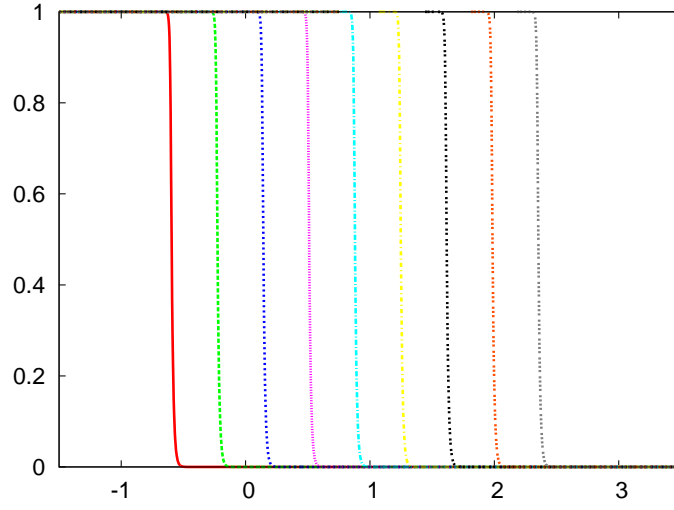


FIG. 3 – Allure de  $\Theta(x)$  à  $0, T/8, 2T/8, \dots, T$

pour  $\Theta$ ,  $\overline{Y_F}$  et  $\overline{Y_O}$ . Puisque

$$\frac{dx}{dt} = \frac{a}{L} \frac{d\bar{x}}{d\bar{t}}$$

compte tenu de l'adimensionnement, on obtient une vitesse réelle

$$\frac{dx}{dt} \simeq \frac{2,26 \cdot 10^{-5}}{4 \cdot 10^{-3}} \times 73,723 \simeq 0,416 \text{ m.s}^{-1} = 41,6 \text{ cm.s}^{-1}$$

ce qui correspond à l'ordre de grandeur d'une vitesse de flamme lors de la combustion du méthane et achève de justifier la validité du modèle utilisé.

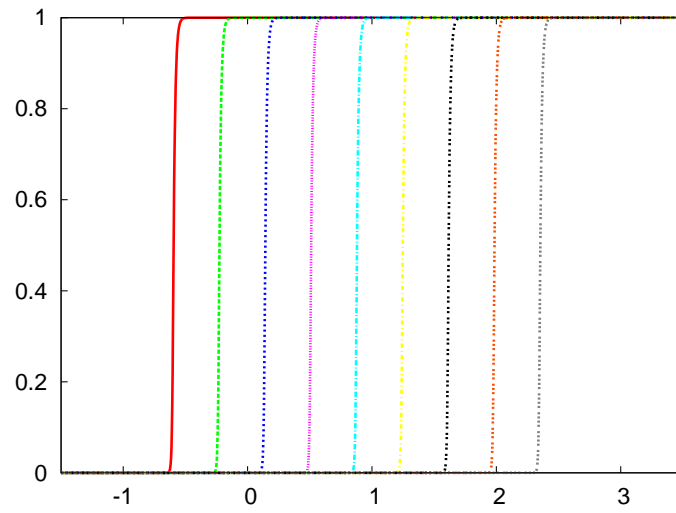


FIG. 4 – Allure de  $\overline{Y_F}(x)$  à  $0, T/8, 2T/8, \dots, T$

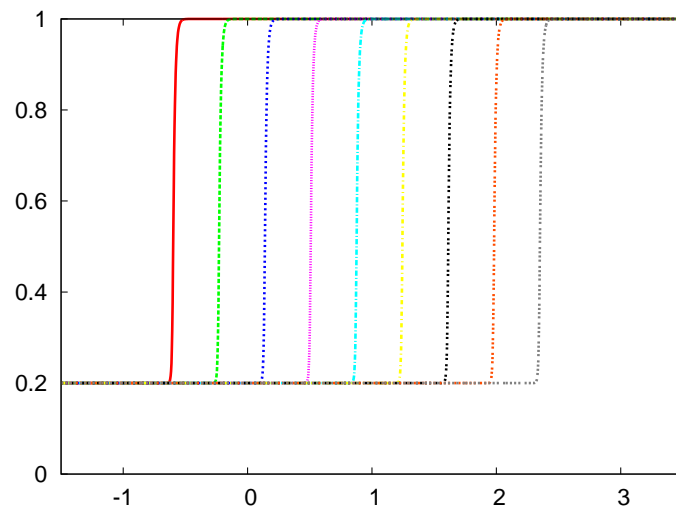


FIG. 5 – Allure de  $\overline{Y_O}(x)$  à  $0, T/8, 2T/8, \dots, T$

### 3 Raideur des systèmes réaction - diffusion

#### 3.1 Notion de raideur

Pour les systèmes différentiels ordinaires, la raideur est une notion difficile à définir. Le terme est utilisé en analyse numérique lorsque les méthodes numériques explicites rencontrent des difficultés à intégrer le système d'EDO : seules des méthodes implicites telles que la méthode BDF donnent des résultats satisfaisants (pour des précisions sur les méthodes numériques, voir l'annexe correspondante à la fin de ce rapport). En général, cela implique qu'une vaste gamme d'échelles de temps intervient dans le problème. En réalité, la raideur est liée à deux aspects :

- Le spectre de la matrice jacobienne associée au système (notamment, la dispersion des valeurs propres)
- La "distance" de la condition initiale à la "variété d'équilibre" du système

Le deuxième point joue un rôle important. Ainsi, considérons le problème de Curtiss & Hirschfelder, cité dans [4].

$$y' = \frac{-y + \cos t}{\varepsilon} \quad 0 < \varepsilon \ll 1$$

dont les solutions, représentées sur la figure (6), sont :

$$y(t) = Ce^{-t/\varepsilon} + \frac{\cos t}{1 + \varepsilon^2} + \frac{\varepsilon \cos t}{1 + \varepsilon^2} \quad C \in \mathbb{R} \quad (17)$$

Si  $1/\varepsilon$  (valeur propre du système) est grand, **et** si la condition initiale est éloignée de la solution de régime permanent  $\frac{\cos t}{1 + \varepsilon^2} + \frac{\varepsilon \cos t}{1 + \varepsilon^2}$ , la méthode d'Euler explicite a beaucoup de difficultés à converger vers la solution. On voit donc que les deux aspects cités ci-dessus ont leur importance.

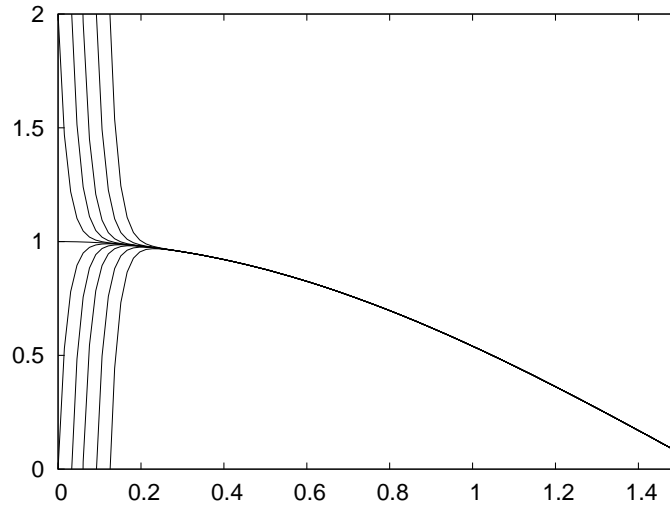


FIG. 6 – Solutions (17) de l'équation de Curtiss & Hirschfelder avec  $\varepsilon = 1/50$  et  $C \in \{\pm 1, \pm 5, \pm 20, \pm 100, \pm 500\}$

Lors de la résolution numérique des systèmes de réaction-diffusion, nous sommes fréquemment confrontés au problème de leur raideur, ce qui nous conduit à utiliser des méthodes de type BDF. En effet, le terme de réaction peut présenter des échelles de temps très courtes et d'autres très rapides, selon les réactions chimiques mises en jeu. Mais la résolution du problème de diffusion seul peut lui aussi poser des problèmes de raideur, à cause des deux aspects déjà évoqués : c'est ce que nous allons détailler maintenant.

### 3.2 Dispersion des valeurs propres du laplacien discrétisé

On considère l'équation de la chaleur :

$$\frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2}, \quad x \in [a, b]$$

que l'on discrétise par méthode des lignes (cf 4). On obtient une équation différentielle ordinaire :

$$\frac{dU}{dt} = D \left( \frac{n}{b-a} \right)^2 AU$$

où, si les  $(x_i)_{0 \leq i \leq n}$  sont les points de discrétisation :

$$U = {}^t(U_0, \dots, U_n) = {}^t(u(x_0), \dots, u(x_n))$$

et  $A \in \mathcal{M}_{n+1}(\mathbb{R})$  est la matrice de discrétisation du laplacien (20).

Les valeurs propres de la matrice  $A$  sont :

$$\lambda_i = -2 \left( 1 + \cos \left( \frac{i\pi}{n+2} \right) \right) = -4 \cos^2 \left( \frac{i\pi}{2(n+2)} \right) \quad i \in [1, n+1]$$

Donc les valeurs propres de la matrice jacobienne du système sont telles que :

$$\forall i \quad -4D \left( \frac{n}{b-a} \right)^2 \cos^2 \left( \frac{\pi}{2(n+2)} \right) \leq \lambda_i \leq -4D \left( \frac{n}{b-a} \right)^2 \cos^2 \left( \frac{n+1}{n+2} \frac{\pi}{2} \right)$$

Lorsque le nombre de points de discrétisation devient grand, elles deviennent donc très dispersées.

### 3.3 Influence de la condition initiale sur la raideur du système

Nous allons maintenant voir pourquoi la condition initiale a une grande influence sur le caractère raide du système. Considérons d'abord le cas d'une onde plane :

$$U(x, t) = \text{Re} \left( e^{i(kx - \omega t)} \right) \quad (18)$$

sur laquelle on fait agir l'équation de la chaleur :  $\frac{\partial U}{\partial t} = D \frac{\partial^2 U}{\partial x^2}$ . On obtient la relation de dispersion :

$$i\omega = Dk^2 \quad \text{soit} \quad \omega = -iDk^2$$

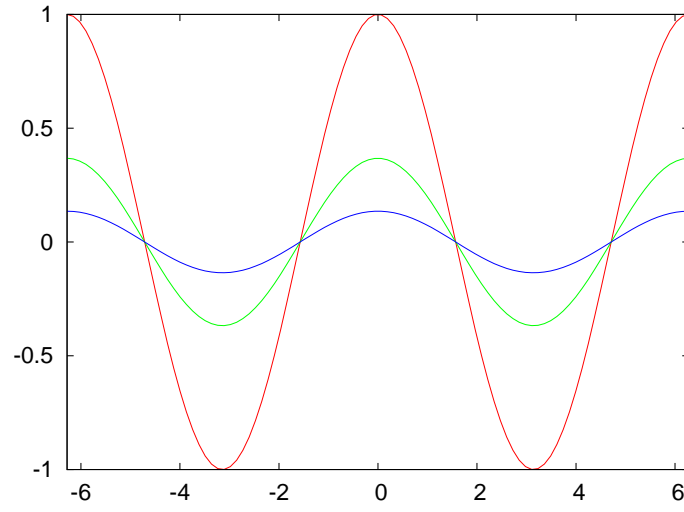


FIG. 7 – Atténuation par l'équation de la chaleur de l'onde plane (18) avec  $k = 1$ , aux temps 0, 1, 2

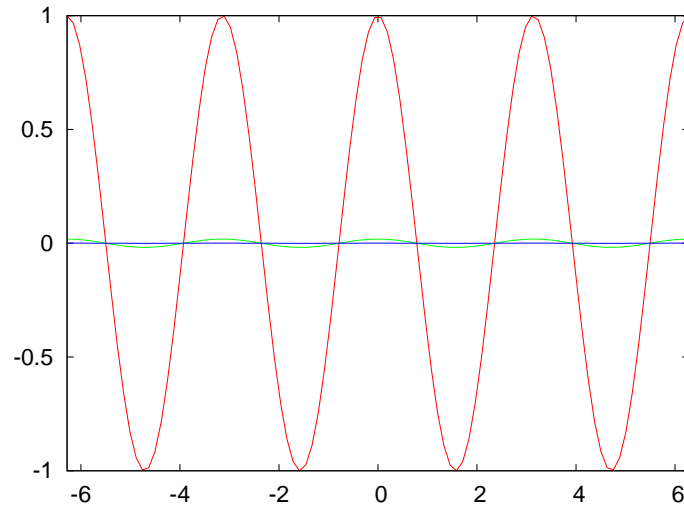


FIG. 8 – Atténuation par l'équation de la chaleur de l'onde plane (18) avec  $k = 2$ , aux temps 0, 1, 2

Ainsi,

$$U(x, t) = \text{Re} \left( e^{i(kx - \omega t)} \right) = \cos(kx) e^{-Dk^2 t}$$

On voit que la sinusoïde relaxe vers la constante 0, l'atténuation se faisant proportionnellement au *carré* de sa fréquence.

Soit maintenant une condition initiale  $u_0(x)$  quelconque : en en prenant la transformée de Fourier on voit d'après ce qui précède que, si les projections sur les exponentielles de nombre d'onde élevé ont des valeurs non négligeables, des échelles de *temps* rapides vont intervenir et perturber la résolution numérique du système. Plus précisément, le système relaxe progressivement vers sa moyenne, les fréquences les plus élevées étant d'abord atténuées.

Il est donc intéressant de regarder la transformée de Fourier en espace de la condition initiale pour savoir si des échelles de temps rapides vont intervenir. On peut d'ailleurs tout aussi bien regarder la transformée de Fourier de sa dérivée, les coefficients étant directement liés. Dans le cas de KPP et celui de la combustion, la dérivée de la condition initiale est proche d'une gaussienne, d'autant moins étalée en espace, donc d'autant plus en fréquence, que le système est "raide". On comprend donc pourquoi de forts gradients spatiaux sur la condition initiale sont susceptibles de créer des problèmes de raideur (au sens de la résolution numérique temporelle). C'est ce phénomène que nous observerons tout au long de notre étude.



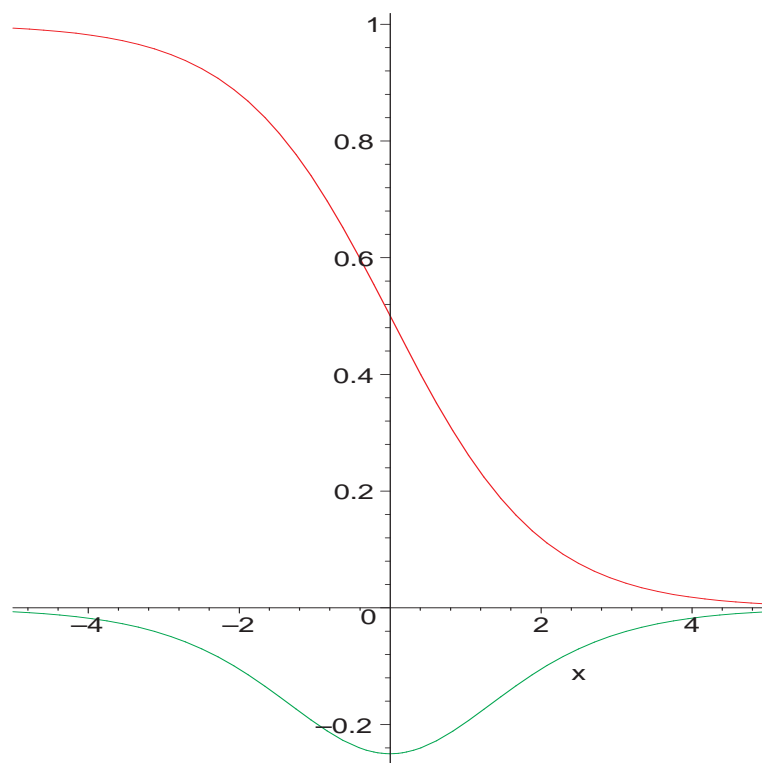


FIG. 9 – Condition initiale de type "KPP non raide" et sa dérivée (très étalée en  $x$ , donc peu en  $k$ )

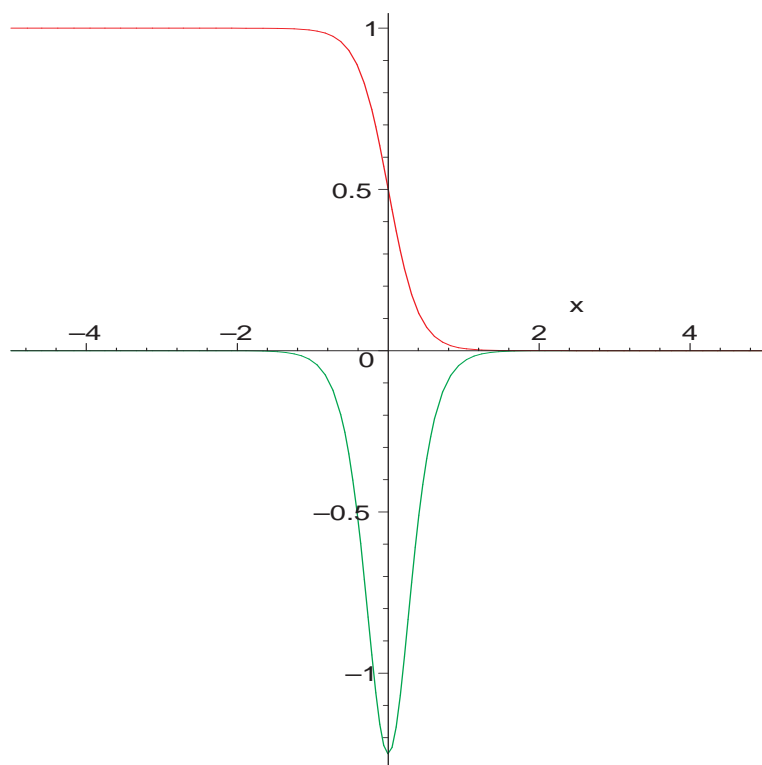


FIG. 10 – Condition initiale de type "KPP raide" et sa dérivée (peu étalée en  $x$ , donc fortement en  $k$ )

## Deuxième partie

# Méthodes numériques de résolution

## 4 La méthode des lignes (MOL)

### 4.1 Principe

La *méthode des lignes* (Method of Lines, ou MOL) est une technique générale de résolution d'équations aux dérivées partielles. Elle consiste à en effectuer une discrétisation spatiale - souvent par différences finies - pour obtenir un système d'équations différentielles ordinaires. Ceux-ci se résolvent alors par l'une des nombreuses méthodes de calcul disponibles (voir l'annexe sur les méthodes numériques). Ces méthodes sont d'ailleurs implémentées dans des codes de calcul qui sont très efficaces, ce qui évite d'avoir à les reprogrammer.

### 4.2 Exemple

Illustrons la méthode des lignes sur un exemple simple en une dimension : on cherche souvent dans le projet à résoudre des équations de type :

$$\begin{cases} \frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2} + f(u) \\ \frac{\partial u}{\partial t}(a, t) = 0 \\ \frac{\partial u}{\partial t}(b, t) = 0 \\ u(x, 0) = u_0(x) \end{cases} \quad x \in [a, b], \quad t \in [0, T] \quad (19)$$

On a pris des conditions aux limites de Neumann homogènes. En accord avec la méthode des lignes, effectuons une discrétisation en espace par différences finies. On fixe nombre  $(n+1) \in \mathbb{N}^*$  de nombre de points de discrétisation et on pose

$$h = \frac{b-a}{n}$$

On définit alors la subdivision régulière du segment  $[a, b]$ ,  $(x_i)_{i \in [0, n]}$ , de pas  $h$  par :

$$\forall i \in [0, n], \quad x_i = a + ih$$

La fonction  $u(x, t)$  sera représentée par le vecteur  $U(t) \in \mathbb{R}^{n+1}$  tel que  $\forall i \in [0, n]$ ,  $U_i(t) = u(t, x_i)$ . On approche la dérivée seconde au point  $x_i, i \in [1, n-1]$  par la formule :

$$\frac{\partial^2 u}{\partial x^2}(x_i) = \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1}))}{h^2} = \left( \frac{n}{b-a} \right)^2 (U_{i+1} - 2U_i + U_{i-1})$$

Au bord, les conditions de Neumann homogènes font qu'on a simplement :

$$\begin{cases} \frac{\partial^2 u}{\partial x^2}(x_0) = \frac{u(x_1) - u(x_0)}{h^2} = \left( \frac{n}{b-a} \right)^2 (U_1 - U_0) \\ \frac{\partial^2 u}{\partial x^2}(x_n) = \frac{-u(x_{n-1}) + u(x_{n-2}))}{h^2} = \left( \frac{n}{b-a} \right)^2 (-U_{n-1} + U_{n-2}) \end{cases}$$

Le vecteur  $U$  est alors la solution d'une équation différentielle ordinaire :

$$\frac{dU}{dt} = D\left(\frac{n}{b-a}\right)^2 A_{nh}U + F(U)$$

où  $A_{nh}$  est la matrice de discrétisation du laplacien avec conditions de Neumann homogènes :

$$A_{nh} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ & & \ddots & \ddots & \ddots & \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \in \mathcal{M}_{n+1}(\mathbb{R})$$

et  $F(U)$  le vecteur  ${}^t(f(U_0), f(U_1), \dots, f(U_n))$ . Remarquons que la matrice  $A_{nh}$  n'est pas inversible, son noyau contenant le vecteur  ${}^t(1, 1, \dots, 1)$ .

C'est ce système que l'on passe en argument à des solveurs d'équations différentielles ordinaires comme LSODE lors du codage de l'algorithme.

Dans le cas général, les conditions aux limites imposent des légères modifications à la matrice :

$$A = \begin{pmatrix} -2 & 1 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ & & \ddots & \ddots & \ddots & \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 1 & -2 \end{pmatrix} \quad (20)$$

## 5 Le solveur LSODE

### 5.1 Introduction

Le solveur LSODE (Livermore Solver for Ordinary Differential Equations) résout numériquement les équations différentielles ordinaires, autrement dit les problèmes de la forme

$$\begin{cases} y'(t) = f(t, y(t)) & t \in I \\ y(t_0) = y_0 \in \mathbb{R}^d \end{cases} \quad (21)$$

L'objectif est de déterminer la fonction solution  $y$  en un ou plusieurs points de l'intervalle de résolution  $I$ . Ici on supposera que  $f$  est localement lipschitzienne sur  $I$  (ce qui assure, on le rappelle, l'existence et l'unicité de la solution de (21)). Les méthodes de résolution du solveur consistent à partir de la condition initiale, pour générer des approximations de la solution exacte en des points eux-même déterminés par le solveur. La distance entre deux points est appelée le pas et est noté  $h_n$ . La principale caractéristique du solveur LSODE est sa capacité à résoudre les problèmes raides (voir le paragraphe 3) grâce à l'utilisation de la méthode des différenciations rétrogrades (Backward differentiation formula ou BDF).

## 5.2 La méthode BDF

Le solveur utilise des schémas de résolution multipas linéaires pour résoudre le système. Ces schémas sont définis par

$$Y_n = \sum_{j=1}^{K_1} \alpha_j Y_{n-j} + h_n \sum_{j=0}^{K_2} \beta_j f_{n-j} \quad (22)$$

Les  $(Y_n)$  sont des valeurs approchées de  $y$  aux points  $(t_n)$ , et les  $f_n$  valent  $f(Y_n)$ . Pour les problèmes raides, la méthode des différenciations rétrogrades (Backward Differentiation Formula, ou BDF) est préférentiellement utilisée. Elle vérifie une relation de la forme (22) avec  $K_1 = q$  et  $K_2 = 0$ , où  $q$  désigne l'ordre de la méthode. Elle consiste à approcher la fonction  $y$  par le polynôme interpolateur  $Q_{n+1}$  de degré inférieur ou égal à  $q$  qui interpole  $Y_{n+1}$  ainsi que les  $q$  valeurs précédentes  $(Y_k)_{k \in [n-q+1, n]}$  aux points  $(t_k)_{k \in [n-q+1, n+1]}$ . Le polynôme  $Q_{n+1}$  est alors défini dans le cas où les pas sont constants égaux à  $h$  par

$$Q_{n+1}(t_n + sh) = \sum_{i=0}^q (-1)^i \binom{-s+1}{i} (\nabla^i Y_{n+1}) \quad s \in \mathbb{R} \quad (23)$$

avec

$$\binom{-s+1}{i} = \frac{1}{i!} \prod_{k=0}^{i-1} ((-s+1) - i)$$

et  $(\nabla^i Y)_{n+1}$  est la différence finie régressive d'ordre  $i$ . On rappelle que les différences finies sont définies par

$$\nabla^0 Y_{n+1} = Y_{n+1}, \quad \nabla^{i+1} Y_{n+1} = \nabla^i Y_{n+1} - \nabla^i Y_n \quad \forall i \geq 1$$

Pour déterminer l'inconnue  $Y_{n+1}$ , on impose  $\dot{Q}(Y_{n+1}) = f_{n+1}$ . Cette relation permet de trouver que

$$\alpha_k = \frac{1}{k} \quad \forall k \geq 1$$

## 5.3 Formulation canonique du problème

On peut réécrire le schéma (22) sous la forme

$$Y_n = \psi_n + h_n = \psi_n + h_n \beta_0 f(Y_n) \quad (24)$$

où  $\psi_n$  contient uniquement des termes connus, calculés aux rangs précédents et est donné pour la méthode BDF par

$$\psi_n = \sum_{j=1}^q \alpha_j Y_{n-j}$$

Comme on l'a vu au paragraphe précédent, dans cette méthode, les coefficients sont déterminés de façon à ce que l'équation (22) soit exacte si la solution du système différentiel est un polynôme de degré au plus  $q$ . Cette méthode est dite implicite car dans (24),  $Y_n$  est solution d'une équation généralement non linéaire. Le code va donc fournir une méthode de résolution de cette équation.

## 5.4 Schémas prédicteur-correcteur

Les schémas prédicteur-correcteur sont utilisés afin de calculer la solution de l'équation non linéaire (24). Ils consistent en la donnée de deux schémas multipas : un schéma explicite défini par  $(\alpha_j^*)$ ,  $1 \leq q$  et  $\beta_1^*$  et un schéma implicite défini par  $(\alpha_j)$ ,  $1 \leq q$  et  $\beta_1$ . On détermine alors pour tout  $n$  une suite récurrente  $(Y_n^{[m]})_m$  qui converge vers la solution de (24). Le premier terme  $Y_n^{[0]}$ , appelé prédicteur, est calculé à partir des  $(Y_k)_{k \in [n-q+1, n]}$  grâce à la méthode explicite. Les termes suivants sont déterminés à partir de (24) (et par conséquent, à partir du schéma explicite) grâce à des méthodes itératives de résolution de systèmes non linéaires. La plus utilisée dans LSODE est la méthode de Newton-Raphson, que nous verrons en détail. Numériquement, on prend  $Y_n = Y_n^{[M]}$  pour  $M$  bien choisi.

## 5.5 Méthode de Newton-Raphson

Soit la suite  $(Y_k)_{k \in [0, n-1]}$  des valeurs approchées de  $y(t_k)$  précédemment calculées. Nous allons réécrire l'équation (24) afin que le problème soit équivalent à la recherche des racines d'une fonction que l'un notera  $R$ . On a alors

$$R(Y_n) = Y_n - \psi_n - h_n \beta_0 f(Y_n) = 0 \quad (25)$$

On va définir la suite  $(Y_n^{[m]})_m$  qui converge vers la solution de (25). On définit le premier terme ou **prédicteur** par

$$Y_n^{[0]} = \sum_{j=1}^q \alpha_j^* Y_{n-j} + h_n \beta_1 f_{n-1}$$

Pour obtenir la valeur  $Y_n^{[m+1]}$ , la méthode de Newton-Raphson suppose que  $Y_n^{[m+1]}$  est proche de  $Y_n^{[m]}$  et que  $R(Y_n^{[m+1]}) \approx 0$  (car c'est le résultat que l'on cherche). En faisant un développement limité au voisinage de  $Y_n^{[m]}$  on obtient

$$\mathbf{P}(Y_n^{[m+1]} - Y_n^{[m]}) = -R(Y_n)$$

où  $\mathbf{P}$  est la matrice jacobienne de  $R$ . On a alors

$$Y_n^{[m+1]} = Y_n^{[m]} - \mathbf{P}^{-1} R(Y_n)$$

La connaissance du jacobien de  $R$  (et plus précisément de son inverse) nous permet donc de déterminer  $Y_n^{[m+1]}$ . On a  $\mathbf{P} = \mathbf{I} - h_n \beta_0 \mathbf{J}$  où  $\mathbf{J}$  est la matrice jacobienne de  $f$ . Cette dernière matrice n'est certes pas connue, mais dans la mesure où elle n'apparaît pas explicitement dans le système différentiel initial, elle n'a pas besoin d'être calculée avec une très grande précision. On peut le calculer de la manière suivante :

$$J_{ij} = \frac{f_i(Y_j + \Delta Y_j) - f_i(Y_j)}{\Delta Y_j}$$

La méthode numérique d'inversion de cette matrice n'est pas étudiée ici. Notons simplement qu'une méthode, la méthode de Jacobi-Newton calcule le jacobien en négligeant tous les termes non diagonaux du jacobien. On trouve alors pour la matrice du jacobien

$$J_{ij} = \delta_{ij} \frac{f_i(Y_j + \Delta Y_j) - f_i(Y_j)}{\Delta Y_j}$$

L'avantage de cette méthode est que l'inversion est simplifiée, l'inconvénient est que la convergence est moins rapide qu'avec un calcul explicite de  $\mathbf{P}^{-1}$ .

## 5.6 Résumé de la méthode de Newton-Raphson

L'objectif est de calculer  $Y_n$  à partir des  $(Y_k)_{k \in [0, n-1]}$ , valeurs approchées de  $y(t_k)$  précédemment calculées.

$$\begin{cases} Y_n^{[0]} = \sum_{j=1}^q \alpha_j^* Y_{n-j} + h_n \beta_1 f_{n-1} \\ Y_n^{[m+1]} = Y_n^{[m]} - \mathbf{P}^{-1} R(Y_n) \\ Y_n = Y_n^{[M]} \end{cases} \quad (26)$$

## 5.7 Formulation matricielle

Dans un souci d'implémentation informatique des méthodes multipas, on introduit la matrice  $Z_n$  qui contient toute l'information calculé au rang  $n$

$$Z_n = (Y_n, h_{n+1} \dot{Y}_n, \frac{h_{n+1}^2}{2!} \ddot{Y}_n, \dots, \frac{h_{n+1}^q}{q!} Y_n^{(q)})$$

où on définit les  $Y_n^i$  à l'aide du polynôme (23) par

$$Y_n^{(i)} = Q_{n-1}^{(i)}(\xi_n)$$

On pose également

$$Z_n^{[0]} = Z_{n-1} A$$

où  $A$  est la matrice définie par

$$A_{ij} = \begin{cases} 0 & \text{si } i < j \\ C_i^j & \text{si } i \geq j \end{cases}$$

On a également la relation

$$Z_n^{[m+1]} = Z_n^{[m]} + \mathbf{P}^{-1} g(Y_n^{[m]}) L$$

où  $Z_n^{[m]}$  est la matrice de Nordsieck à la  $m$ -ième itération, :

$$Z_n^{[m]} = (Y_n^{[m]}, h_n \dot{Y}_n^{[m]}, \frac{h_n^2}{2!} \ddot{Y}_n^{[m]}, \dots, \frac{h_n^q}{q!} Y_n^{[m](q)})$$

$L$  est une matrice qui dépend de  $q$  et qui est tabulée pour la méthode BDF. Finalement, en introduisant le vecteur  $e_n^{[m]}$  défini par

$$e_n^{[m]} = \sum_{j=0}^m P^{-1} g(Y_n^{[j]})$$

## 5.8 Résumé de la formulation matricielle

On peut résumer simplement la méthode BDF par le schéma suivant :

$$\begin{cases} Z_n^{[0]} = Z_{n-1}A \\ e_n^{[0]} = 0 \\ g(Y_n^{[m]}) = h_n f(Y_n^{[m]}) \\ e_n^{[m+1]} = e_n^{[m]} + \mathbf{P}^{-1}g(Y_n^{[m]}) \\ Y_n^{[m+1]} = Y_n^{[0]} + L_0 e_n^{[m+1]} \\ e_n = e_n^{[M]} \\ Z_n = Z_n^{[0]} + e_n L \end{cases} \quad (27)$$

## 5.9 Estimation et contrôle de l'erreur locale

L'erreur locale est définie pour la méthode BDF par

$$d_n = \sum_{j=0}^q \left( \frac{\alpha_j}{\beta_0} \right) y(t_{n-j}) + h_n \dot{y}(t_n)$$

En développant en série de Taylor au voisinage de  $t_n$  cette expression (en supposant que la solution  $y$  est infiniment dérivable) on obtient

$$d_n = \sum_{k=0}^{\infty} C_k h_n^k y^{(k)}(t_n)$$

La méthode est dite d'ordre  $q$  si  $C_0 = C_1 = \dots = C_q = 0$  et  $C_{q+1} \neq 0$ . Pour contrôler l'erreur, on définit le vecteur du poids de l'erreur par

$$EWT_{i,n} = RTOL_i |Y_{i,n-1}| + ATOL_i \quad (28)$$

où RTOL et ATOL sont des tolérances données par l'utilisateur. Le schéma est considéré comme suffisamment précis si

$$\|d_n\| \leq 1$$

avec

$$\|d_n\| = \sqrt{\frac{1}{N} \sum_{i=1}^N \left( \frac{d_{i,n}}{EWT_{i,n}} \right)^2}$$

Concrètement, on peut dire le test d'erreur locale est satisfait si pour chaque composante de la solution, ou bien l'erreur relative est inférieure à RTOL, ou bien l'erreur absolue est inférieure à ATOL.



## 6 Les méthodes de splitting

### 6.1 Motivations

Lors de la présentation des systèmes de réaction diffusion dans la première partie, nous avons vu que le second membre de l'équation aux dérivées partielles (1) se compose d'un terme de réaction et d'un terme de diffusion. Par ailleurs, la solution de cette équation est un vecteur  $U : \Omega \times \mathbb{R}^+ \rightarrow \mathbb{R}^n$  dont chaque composante représente une grandeur physico-chimique (concentration d'une espèce, température...)

Ces deux termes pris séparément ont un comportement différent : le terme de diffusion (dans le cas où la matrice  $M$  est diagonale) ne couple pas les différentes espèces chimiques mais fait intervenir les variations de concentration d'une espèce sur tout le domaine spatial ; le terme de réaction quant à lui couple les variables localement (dans la mesure où il modélise les réactions chimiques), sans faire intervenir les gradients spatiaux des solutions.

Le système complet est donc complexe à résoudre car il mêle couplages d'inconnues et de différentiations spatiales. Les méthodes de splitting étudient la possibilité de résoudre séparément les équations de réaction et de diffusion et d'en déduire la solution du système global, afin d'éviter la résolution simultanée de problèmes couplant différents points de l'espace et de problèmes couplant toutes les inconnues. L'intérêt est d'autant plus grand que des méthodes numériques très efficaces existent pour résoudre chacun des termes séparément.

### 6.2 Première approche

Soit  $n \in \mathbb{N}$ ,  $(A, B) \in M_n(\mathbb{R})^2$ . On considère le système différentiel

$$\begin{cases} \frac{dy}{dt} = Ay + By \\ y(0) = y_0 \end{cases} \quad (29)$$

Le flot de cette équation différentielle s'écrit

$$\begin{aligned} \Phi : \mathbb{R}^n \times \mathbb{R} &\mapsto \mathbb{R}^n \\ \Phi(y_0, t) &= \exp(t(A + B))y_0 \end{aligned} \quad (30)$$

On notera dans la suite l'application  $\Phi_t$  définie à partir du flot de la manière suivante

$$\forall t \in \mathbb{R} \quad \Phi_t(y_0) = \Phi(y_0, t) \quad (31)$$

On considère les systèmes différentiels qui découplent les opérateurs de 6.2 ainsi que leurs flots

$$\frac{dy}{dt} = Ay \quad (32)$$

$$\frac{dy}{dt} = By \quad (33)$$

On notera  $\Phi_t^A$  (resp.  $\Phi_t^B$ ) le flot de (32) (resp. (33)). On a alors

$$\Phi_t^A(y_0) = \Phi^A(y_0, t) = \exp(tA)y_0 \quad (34)$$

$$\Phi_t^B(y_0) = \Phi^B(y_0, t) = \exp(tB)y_0 \quad (35)$$

On remarque que si  $[A, B] = 0$ , on a  $\exp(t(A+B)) = \exp(tA)\exp(tB)$  et donc le flot de 6.2 est exactement le composé des flots des équations découplés, soit en raisonnant plus généralement en terme de flot (ce qui sera fait lorsque les opérateurs ne seront plus de simples matrices, mais des opérateurs différentiels)

$$\Phi_t = \Phi_t^A \circ \Phi_t^B$$

Notons que l'égalité est fausse si  $[A, B] \neq 0$ . Mais on verra dans le paragraphe suivant que l'erreur commise en remplaçant  $\Phi_t$  par  $\Phi_t^A \circ \Phi_t^B$  est en  $O(t^2)$ .

Les méthodes de splitting consistent à approcher le flot du système différentiel initial en composant les flots des systèmes différentiels obtenus en séparant les opérateurs, comme dans l'exemple ci-dessus. Elles sont d'autant meilleures que l'ordre de la différence est petit. La formule de Baker-Campbell-Hausdorff (dites de BCH) permet d'affirmer que les coefficients du développement en puissance de  $t$  de la différence entre le flot exact et le flot approché font intervenir les crochets de Lie itérés des opérateurs. Notons que nous utilisons ces méthodes sur deux matrices de dimension finie mais qu'elles ont vocation à être généralisées (dans un cadre théorique tout du moins) à des systèmes dans lesquels les deux matrices sont remplacées par  $n$  opérateurs aux dérivées partielles et les crochets de Lie des matrices par des crochets de Lie d'opérateurs. Cela pose des problèmes car il est nécessaire de considérer des séries formelles d'opérateurs (en particulier pour définir l'exponentiel d'un opérateur différentiel) qui ne convergent généralement pas.

### 6.2.1 Méthodes de Lie

- La méthode de Lie  $A - B$  consiste à approcher  $e^{t(A+B)}$  par  $e^{tB}e^{tA}$ . On l'appelle ainsi car on résout d'abord  $dy/dt = Ay$ , puis  $dy/dt = By$ . Remarquons que cette convention change selon les auteurs.

En terme de flux différentiel, avec les notations précédentes, on approche  $\Phi_t^{A+B}$  par  $\Phi_t^B \circ \Phi_t^A$ . Or :

$$e^{t(A+B)} - e^{tB}e^{tA} = \frac{t^2}{2}[A, B] + O(t^3)$$

Cette méthode est donc d'ordre local 2 en  $t$ . On verra en (6.5) que ceci implique qu'elle est d'ordre global 1.

- La méthode de Lie  $B - A$  consiste de même à approcher  $e^{t(A+B)}$  par  $e^{tA}e^{tB}$ . Cette méthode est aussi d'ordre local 2 puisque :

$$e^{t(A+B)} - e^{tA}e^{tB} = \frac{t^2}{2}[B, A] + O(t^3)$$

### 6.2.2 Méthodes de Strang

- La méthode de Strang  $A - B - A$  consiste à approcher  $e^{t(A+B)}$  par  $e^{\frac{t}{2}A}e^{tB}e^{\frac{t}{2}A}$  (notons qu'il n'y a pas d'ambiguïté possible dans le nom ici). En terme de flux différentiel, avec

les notations précédentes, on approche  $\Phi_t^{A+B}$  par  $\Phi_{\frac{t}{2}}^A \circ \Phi_t^B \circ \Phi_{\frac{t}{2}}^A$ . On remarque qu'elle est d'ordre 3 local en  $t$ , donc d'ordre 2 global. On a en effet

$$e^{t(A+B)} - e^{\frac{t}{2}A} e^{tB} e^{\frac{t}{2}A} = -t^3 \left( -\frac{1}{24}[A, [A, B]] + \frac{1}{12}[B, [B, A]] \right) + O(t^4)$$

- De la même manière, la méthode de Strang  $B - A - B$  consiste à approcher  $e^{t(A+B)}$  par  $e^{\frac{t}{2}B} e^{tA} e^{\frac{t}{2}B}$  et elle est aussi d'ordre local 3.

### 6.3 Splitting et calcul numérique

L'intérêt des méthodes de splitting réside dans leur utilisation pour la résolution numérique des équations aux dérivées partielles. Etudions l'algorithme de résolution dans le cas de la méthode de Lie A-B :

- Choix du pas de temps  $\Delta t$  et de la condition initiale :  $y := y_0$
- Pour chaque instant  $t$  entre 0 et  $T$  :
  - Calculer  $y^* := \Phi_{\Delta t}^A(y)$  à l'aide du solveur d'équation différentiel que l'on souhaite (LSODE par exemple).
  - Calculer  $y^{**} := \Phi_{\Delta t}^B(y^*)$ , éventuellement à l'aide d'un autre solveur plus adapté à l'opérateur  $B$ .
  - Faire  $t := t + \Delta t$  et  $y := y^{**}$

### 6.4 Splitting et perte d'ordre

#### 6.4.1 Expression intégrale de l'erreur locale

Le problème de ces méthodes de splitting pour le calcul numérique est qu'on ne sait pas mesurer avec précision le terme en  $O(t^n)$  car il s'écrit sous la forme d'une série infinie. On va donc chercher à écrire la différence entre l'approximation splittée et le flot exact sous la forme d'une fonction finie. On trouve que

$$e^{tA} e^{tB} - e^{t(A+B)} = \int_0^t \int_0^s e^{(t-s)(A+B)} e^{(s-r)A} [A, B] e^{rA} e^{sB} dr ds \quad (36)$$

*Démonstration.* Soit  $L : t \mapsto e^{tA} e^{tB}$ . En dérivant cette expression, on trouve que  $L$  vérifie l'équation différentielle ordinaire

$$\frac{dL}{dt} = (A + B)L + R_1(t)$$

où  $R_1(t) = R_2(t)e^{tB}$  avec  $R_2(t) = [e^{tA}, B]$ . En résolvant cette équation différentielle, on écrit  $L$  sous la forme

$$L(t) = e^{t(A+B)} + \int_0^t e^{(t-s)(A+B)} R_1(s) ds$$

Par ailleurs,  $R_2$  vérifie l'équation différentielle

$$\frac{dR_2}{dt} = AR_2(t) + [A, B] e^{tA}$$

et se réécrit donc sous la forme

$$R_2(t) = \int_0^t e^{t-s} [A, B] e^{sA} ds$$

Finalement on obtient l'expression de  $L$

$$L(t) = e^{t(A+B)} + \int_0^t e^{(t-s)(A+B)} \left( \int_0^s e^{(s-r)A} [A, B] e^{rA} dr \right) e^{tB} ds$$

D'où le résultat.  $\square$

#### 6.4.2 Evaluation de la perte d'ordre quand la condition initiale présente de forts gradients

On va étudier le phénomène de perte d'ordre sur un exemple. On s'intéresse à la résolution du système

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + V(x)u \\ u(a, t) = u(b, t) = 0 \\ u(x, 0) = u_0(x) \\ x \in [a, b] \end{cases} \quad (a, b) \in \mathbb{R} \quad (37)$$

où  $a, b \in \mathbb{R}$ . On cherche à résoudre le problème par la méthode des lignes, comme au paragraphe 4 : soit  $n \in \mathbb{N}^*$ , on pose donc

$$h = \frac{b-a}{n+1}$$

et on définit la subdivision régulière du segment  $[a, b]$ ,  $(x_i)_{i \in [0, n+1]}$ , de pas  $h$  :

$$\forall i \in [0, n+1], x_i = a + ih$$

On introduit le vecteur  $U$  tel que la  $i$ -ème composante de  $U$ , que l'on notera  $U_i$ , constitue une approximation de  $u(t, x_i)$ , solution exacte estimée au point  $(t, x_i)$ . On cherche  $U_i$  pour  $1 \leq i \leq n$ , car les valeurs de  $U_i$  sont connues pour  $i = 0$  et  $i = n+1$ . On note  $A$  la matrice de discrétisation du laplacien (20). De même, le potentiel  $V$  sera représenté par la matrice  $B$  telle que

$$B = \begin{pmatrix} V(x_1) & 0 & \dots & 0 & 0 \\ 0 & V(x_2) & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & V(x_{n-1}) & 0 \\ 0 & 0 & \dots & 0 & V(x_n) \end{pmatrix}$$

Pour simplifier, on notera  $V(x_i) = V_i$  pour  $1 \leq i \leq n$ . On pose  $(U_0)_i = u_0(x_i)$  la condition initiale. L'équation vérifiée par  $U$  est donc de même forme que celle de (6.2).

On voit sur l'expression (36) que l'expression de l'erreur de splitting dépend du commutateur  $[A, B]$ . Dans la suite, on va en donner une expression faisant ressortir les termes intéressants.

On a

$$\begin{aligned}(AB)_{ij} &= a_{ij}V_j \\ (BA)_{ij} &= a_{ij}V_i\end{aligned}$$

Soit  $W \in \mathbb{R}^n$ , on a alors

$$\begin{aligned}(ABW)_i - (BAW)_i &= \sum_{k=1}^n (a_{ik}V_kW_k - a_{ik}V_iW_k) \\ &= \frac{1}{h^2} (V_iW_{i+1} - V_{i-1}W_{i-1} - V_{i+1}W_{i+1} + V_iW_{i+1})\end{aligned}$$

avec la convention  $W_0 = W_{n+1} = 0$ . En conclusion, après quelques manipulations, on trouve

$$(ABW)_i - (BAW)_i = \frac{1}{h^2} (2V_i - V_{i+1} - V_{i-1})W_i + \frac{V_{i-1} - V_i}{h} \frac{W_i - W_{i-1}}{h} + \frac{V_{i+1} - V_i}{h} \frac{W_i - W_{i+1}}{h}$$

Soit

$$[A, B]W_i = -V_i''W_i - V_i'(DW)_i - V_i'(D^*W)_i + O(h^2)W_i + O(h) \frac{W_{i+1} - W_{i-1}}{h}$$

Donc

$$[A, B]W = -B''W - B'(D - D^*)W + O(h)W + O(h)(D - D^*)W \quad (38)$$

où on a défini la matrice  $D$  par

$$D = \frac{1}{h} \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -1 & 1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix}$$

Le crochet de Lie de  $A$  et  $B$  se compose donc de deux termes. Cependant, dans le cas d'un système où  $u_0(x)$  possède un fort gradient, le second terme est prépondérant par rapport au premier. En effet, si on suppose que  $0 \leq u_0 \leq 1$  et que  $V''$  est de "taille raisonnable" (on entend par taille raisonnable que  $\|V''\|$  est de l'ordre de grandeur de l'unité), alors  $B''U_0$  est de taille raisonnable, tandis que les composantes du vecteur  $DU_0$  valent en réalité  $(DU_0)_i = u'(x_i) + O(h)$ , et le gradient de  $u_0$  peut être très important. Le terme dominant de l'erreur peut donc donner asymptotiquement un ordre 2 mais avec une très grosse constante. On va

donc chercher à obtenir une estimation qui ne fait pas intervenir la matrice  $D$ . On utilisera la norme matricielle subordonnée

$$\|A\|_2^2 = \rho(AA^*)$$

où  $\rho(A)$  désigne la plus grande valeur propre de  $A$ . En reprenant l'expression (36) et en notant

$$E(t) = \left( e^{t(A+B)} - e^{tA} e^{tB} \right) U_0$$

on a

$$\|E(t)\|_2 \leq \int_0^t \int_0^s \|e^{(t-s)(A+B)}\|_2 \|e^{(s-r)A}\|_2 \|[A, B]e^{rA}\|_2 \|e^{sB}\|_2 dr ds \quad (39)$$

L'objectif de ce qui va suivre est de montrer le résultat suivant :

**Théorème 6.1** (Perte d'ordre). *L'erreur de splitting vérifie la majoration*

$$\|E(t)\|_2 \leq \frac{4}{3}(1 + O(h))t\sqrt{t} + O(t^2)$$

où le terme en  $O(h)$  s'écrit

$$O(h) = hB'' + O(h^2)$$

*Démonstration.* Les termes contenant les exponentielles de matrice se majorent aisément compte tenu du fait que les matrices  $A$  et  $B$  sont symétriques positives. En effet, Si pour une matrice  $V$  symétrique positive on note  $Diag_V$  la matrice diagonale contenant les valeurs propres de  $V$  rangées dans l'ordre croissant, on a

$$\|e^{-tV}\|_2^2 = \rho(e^{-2tV}) = \rho(e^{-2tDiag_V}) = e^{-2tV_n} \leq 1$$

Pour  $t \geq 0$  On en déduit que

$$\|e^{-(t-s)(A+B)}\|_2 \leq 1, \|e^{-(s-r)A}\|_2 \leq 1, \|e^{-sB}\|_2 \leq 1$$

car  $0 \leq r \leq s \leq t$ . D'où

$$\|E(t)\|_2 \leq \int_0^t \int_0^s \|[A, B]e^{-rA}\|_2 dr ds \quad (40)$$

On en déduit, d'après (38) que

$$\|E(t)\|_2 \leq \int_0^t \int_0^s \left\| \left( -B'' - B'(D - D^*) + O(h) + O(h)(D - D^*) \right) e^{-rA} \right\|_2 dr ds \quad (41)$$

Le premier terme de l'intégrale se majore aisément. En effet

$$\int_0^t \int_0^s \| -B'' e^{-rA} \|_2 dr ds \leq \|B''\|_2 \frac{t^2}{2}$$

De même

$$\int_0^t \int_0^s \|O(h)e^{-rA}\|_2 dr ds \leq O(h) \frac{t^2}{2}$$

On obtient donc des termes d'ordre deux avec des constantes dont l'ordre de grandeur est 1. Il ne nous reste plus qu'à majorer le terme

$$\int_0^t \int_0^s \|(D - D^*)(1 + O(h))e^{-rA}\|_2 dr ds$$

On a

$$\|(D - D^*)e^{-rA}\| \leq 2 \|De^{-rA}\|_2$$

L'idée pour étudier ce terme consiste à voir  $D$  comme la représentation discrétisée de l'opérateur de dérivation. Il apparaît alors que  $D$  peut être approché par la matrice  $\sqrt{A}$  qui doit également être une représentation discrète de l'opérateur de dérivation car  $A$  est la discrétisation du laplacien et dans le cas continue on a bien  $\Delta = \partial_x^2$ . On en déduit que le produit  $D\sqrt{A}^{-1}$  doit être "petit". Pour voir cela, on introduit le terme  $\sqrt{A}$  dans l'inégalité précédente. On a alors

$$\|De^{-rA}\|_2 = \|D\sqrt{A}^{-1}\sqrt{A}e^{-rA}\|_2 \leq \|D\sqrt{A}^{-1}\|_2 \|\sqrt{A}e^{-rA}\|_2$$

Le premier terme se majore aisément en diagonalisant les matrices  $A$  et  $\sqrt{A}$ .

$$\|\sqrt{A}e^{-rA}\|_2^2 = \max_{\lambda \in Sp(A)} \lambda e^{-2r\lambda} \leq \frac{1}{2r}$$

Enfin, le terme

$$\|D\sqrt{A}^{-1}\|_2 = \rho(D(\sqrt{A})^{-1}(\sqrt{A})^{-1}D^*) = \rho(DD^*A^{-1})$$

On remarque alors que  $DD^* = A - \frac{R}{h^2}$  où  $R$  est la matrice définie par  $R_{NN} = 1$  et  $R_{i,j} = 0$  pour  $(i,j) \neq (N,N)$ . On en déduit que

$$DD^*A^{-1} = Id - R\frac{A^{-1}}{h^2}$$

Cette matrice est triangulaire inférieure avec comme valeurs propres 1 et  $\frac{1}{N+1}$ . On en déduit que

$$\rho(DD^*\frac{A^{-1}}{h^2}) = 1$$

Et par suite

$$\|D\sqrt{A}^{-1}\|_2 = 1$$

On a alors

$$\int_0^t \int_0^s \|(D - D^*)e^{-rA}\|_2 dr ds \leq 2 \int_0^t \int_0^s \frac{1}{\sqrt{r}} dr ds = \frac{4}{3}t\sqrt{t}$$

Et finalement

$$\|E(t)\|_2 \leq \frac{4}{3}(1 + O(h))t\sqrt{t} + O(t^2)$$

□

On obtient donc une deuxième majoration de l'erreur. D'après les hypothèses faites, le terme en  $O(h)$  est négligeable devant 1 car  $\|B''\| \simeq 1$ . On a donc réussi à trouver une majoration locale de l'erreur qui ne dépend pas de la matrice de dérivation, mais ce au prix d'une diminution de l'ordre de cette erreur.

Pour donner une idée de l'intérêt de ce résultat, on peut dire que l'erreur de splitting vérifie deux majorations dans le cas raide :

$$E(t) \leq C_1 t^2 + o(t^2)$$

$$E(t) \leq C_2 t^\alpha + o(t^\alpha)$$

et donc

$$E(t) \leq \min(C_1 t^2, C_2 t^\alpha)$$

avec  $C_1 \geq C_2$ . La meilleure approximation dépend donc de la valeur de  $t$ . En effet, si on pose

$$K = \left(\frac{C_2}{C_1}\right)^{\frac{1}{2-\alpha}}$$

pour  $t \geq K$ , on a  $C_2 t^\alpha \leq C_1 t^2$  et donc la meilleure approximation de  $E(t)$  sera d'ordre  $\alpha$ . Au contraire, si  $t \leq K$ ,  $E(t)$  sera d'ordre 2. On voit avec cette approche simplifiée du problème qu'à partir d'un pas de temps donné, il peut y avoir une dégénérescence de l'ordre, selon la raideur de la condition initiale.

## 6.5 Passage de l'erreur locale à l'erreur globale

### 6.5.1 Introduction

Nous avons évoqué le fait que, lorsqu'on obtient une erreur locale d'ordre  $\alpha$ , l'erreur globale est alors d'ordre  $\alpha - 1$ . Le but de ce paragraphe est de démontrer cette propriété sur l'exemple du splitting de Lie dans le cas linéaire.

On a vu que dans ce cas, on majorer l'erreur locale de splitting de la manière suivante :

$$\left| \left( e^{t(A+B)} - e^{tA} e^{tB} \right) U_0 \right| \leq C_\alpha |U_0| t^\alpha \quad (42)$$

On rappelle que la solution exacte du système d'équation est

$$u(t) = e^{t(A+B)} u_0$$

On se donne  $N \in \mathbb{N}$  et considère la subdivision régulière de l'intervalle  $[t_0, T]$  de pas  $h = \frac{T-t_0}{N}$ . On considère la suite des points de cette subdivision  $(t_k)_{k \in [0, N]}$  vérifiant

$$\forall t \in [0, N-1], t_{k+1} = t_k + h$$

On remarque (ce sera utile pour la suite) que :

$$u(t_1 + t_2) = e^{t_1(A+B)} e^{t_2(A+B)} u_0 \quad (43)$$



### 6.5.2 Schéma de splitting

On définit le schéma de splitting suivant :

$$U_{k+1} = e^{hA}e^{hB}U_k \quad (44)$$

On rappelle que les  $U_k$  ont vocation à donner une valeur approchée des  $u(t_k)$  et que le choix du schéma est motivé par le fait que

$$u(t_{k+1}) = e^{h(A+B)}u(t_k) \quad (45)$$

ce que l'on montre à l'aide de (43). On peut réécrire (44) sous la forme

$$U_{k+1} = U_k + h \left( \frac{e^{hA}e^{hB} - Id}{h} \right) U_k = U_k + hF(t_k, U_k, h) \quad (46)$$

L'erreur de consistance de ce schéma est

$$\varepsilon_k = u(t_{k+1}) - u(t_k) - hF(h, u(t_k)) = u(t_{k+1}) - e^{hA}e^{hB}u(t_k)$$

D'après (45), l'erreur de consistance s'écrit

$$\varepsilon_n = \left( e^{h(A+B)} - e^{hA}e^{hB} \right) u(t_k) \quad (47)$$

### 6.5.3 Convergence du schéma de splitting

Pour que les erreurs créées par le schéma puissent se cumuler sans que la solution calculée ne s'éloigne trop de la solution du problème, il faut que le schéma de splitting soit stable. Le théorème suivant répond à cette question.

**Théorème 6.2** (Stabilité du schéma de splitting). *Le schéma (44) est stable.*

*Démonstration.* D'après (46), le schéma (44) s'écrit  $U_{k+1} = U_k + hF(t_k, U_k, h)$  avec

$$F(t, u, h) = \frac{e^{hA}e^{hB} - Id}{h}u$$

Soit la fonction suivante :

$$\omega : h \mapsto \frac{e^{hA}e^{hB} - Id}{h}$$

Elle est définie sur  $\mathbb{R}^*$ , prolongeable par continuité en 0, et le prolongement obtenu est de classe  $C^\infty$ . Soit  $\Lambda > |\omega(0)|$  un nombre réel strictement positif. D'après la continuité de  $\omega$ , il existe  $h^* > 0$  tel que

$$\forall h \in [0, h^*], |\omega(h) - \omega(0)| \leq \Lambda - \omega(0)$$

On en déduit que

$$\forall h \in [0, h^*], |\omega(h)| \leq \Lambda$$

D'où

$$\forall u, v \in \mathbb{R}^{N+1}, \forall h \in [0, h^*],$$

$$|F(t, u, h) - F(t, v, h)| = |\omega(h)||u - v| \leq \Lambda|u - v|$$

On conclut d'après le théorème (2.3) de l'annexe sur les méthodes numériques.  $\square$

Le schéma défini avec la méthode de splitting est donc stable, et consistant d'après l'inégalité vue en (42). Il est donc convergent.

#### 6.5.4 Cumul des erreurs locales

La stabilité nous permet d'écrire l'égalité suivante :

$$\forall k \leq N, |U_k - u(t_k)| \leq M \left( \sum_{0 \leq j \leq k-1} |\varepsilon_j| \right)$$

*Démonstration.* D'après la définition de la stabilité dans l'annexe sur les méthodes numériques, en prenant pour  $(U_k)$  le schéma défini ci-dessus, pour  $(V_k)$  les  $u(t_k)$  et comme condition initiale la même pour le système différentiel et pour le schéma ( $U_0 = u(t_0)$ ), on obtient le résultat directement.  $\square$

On en déduit très simplement

$$\forall k \leq N, |U_k - u(t_k)| \leq M \left( \sum_{0 \leq j \leq N-1} |\varepsilon_j| \right)$$

D'après (47) et (42), on a alors

$$\forall k \leq N, |U_k - u(t_k)| \leq \left( MCN \sup_{k \in [0, N-1]} |u(t_k)| \right) h^\alpha$$

En remarquant que  $N = \frac{T-t_0}{h}$ , on trouve que l'erreur globale est en  $h^{\alpha-1}$ .

## 7 Description des algorithmes

Le programme que nous utilisons pour nos calculs sur KPP se base sur un code écrit en Fortran 77 que nous avons reçu de notre encadrant, Marc Massot, et auquel nous avons apporté quelques modifications pour notre étude.

### 7.1 Déclaration des variables

Nous allons présenter les variables utilisées pour le code

**Les entiers importants :**

*np* : nombres de points de la subdivision.

*ntemp* : nombre de subdivisions temporelles de l'intervalle de temps d'intégration.

*nvar* : nombre de fonctions inconnues.

*jlect* : lorsque la variable *jlect* est initialisée à 0, la condition initiale est une fonction proposée par l'utilisateur tandis que dans le cas où elle vaut 1, le code récupère comme condition initiale la dernière fonction calculée.

**Les nombres réels importants :**

*inttps* : largeur de l'intervalle de temps.

*intspac* : largeur de l'intervalle d'espace.

*errmax* : norme infinie de la différence entre la solution exacte et la solution quasi-exacte.

*t* : variable de temps.

*evalvit* : vitesse de l'onde calculée par la solveur.

*vitreelle* : vitesse exacte de l'onde.

*normel2* : variable utilisée pour le calcul des normes  $L_2$ .

*normelinf* : variable utilisée pour le calcul des normes  $L_\infty$ .

**Les tableaux importants :**

*b* : solution calculée par le solveur. *b(jy)* représente la valeur de la solution  $\beta$  au jy-ième point de la subdivision.

*bexacte* : solution exacte calculée grâce à l'expression analytique. *phi* : tableau intermédiaire du solveur LSODE dans lequel figure la solution calculée au pas de temps précédent

**Notation :** La correspondance entre l'indice spatial de boucle *jy* et la variable spatiale, que l'on notera dans ce paragraphe  $z_{jy}$ , se fait par la relation suivante

$$z_{jy} = \text{intspac} \left( \frac{jy}{np - 1} \right)$$

Dans tout ce qui va suivre, on écrira souvent par abus de notation *b(jy)* au lieu de  $\beta(z_{jy})$  pour désigner la valeur de la solution au jy-ième points de la subdivision, afin de rester en conformité avec la syntaxe Fortran, et pour rendre l'exposé plus clair.

## 7.2 Intégration quasi-exacte du problème discrétisé avec le solveur LSODE

On réalise l'intégration sur un pas de temps *deltat* Le solveur LSODE va résoudre l'équation différentielle sur le segment  $[sss0, sss]$ , où  $sss = sss0 + \text{deltat}$ .

```

do 400 while (t.le.inttps)
print *, 'avant lsode'
sss0 = t
sss = sss0 + deltat
IELWRK(1) = nvar
IELWRK(2) = nvar
CALL LSODE(ondeeq, ncompo, phi, sss0, sss, ITOL, RTOL,
& ATOL, ITASK, ISTATE,
& IOPT, ELWRK, LENWK, IELWRK, LENIWK,
& JOM, MF)
print *, 'apres lsode', ISTATE
t = t + deltat
nt = nt + 1

```

Il se sert pour cela de la sous-routine "ondeeq" qui elle même appelle la sous-routine "londeeq".

```
subroutine londeeq (ncalls,nvar,np,temps,phi,dphidt)
```

```
implicit none
```

```
ncalls = ncalls+1
```

```
carrey = ((np-1)/intspac)*((np-1)/intspac)
```

```
do jy=1,np
```

```
bj(jy) = phi(jy)
```

```
enddo
```

```
jy = 1
```

```
dphidt(jy) = carrey*D*(bj(jy+1) - bj(jy)) +
```

```
& k1*bj(jy)*bj(jy)*(1 - bj(jy))
```

```
do 550 jy= 2, np-1
```

```
dphidt(jy) = carrey*D*(bj(jy+1) -2.d0*bj(jy)+ bj(jy-1)) +
```

```
& k1*bj(jy)*bj(jy)*(1 - bj(jy))
```

```
550 enddo
```

```
jy = np
```

```
dphidt(jy) = carrey*D*(-bj(jy) + bj(jy-1)) +
```

```
& k1*bj(jy)*bj(jy)*(1 - bj(jy))
```

```
return
```

```
end
```

Cette sous-routine calcule le second membre du système. On obtient grâce à ce programme le tableau  $b$  indicé par les  $(jy)_{jy \in [1,np]}$ , tel que

$$b_{jy} = \beta(z_{jy})$$

### 7.3 Intégration par des méthodes de splitting du problème discrétisé avec le solveur LSODE

Dans le cas de l'algorithme splitté, on n'a plus les routines *ondeeq* et *londeeq* qui calculaient le second membre du système (7), mais deux sous-routines *reaceq* et *lreaceq* qui calculent le terme de réaction, auxquelles s'ajoutent deux sous-routines *diffeq* et *ldiffeq* qui calculent le terme de diffusion.

```
subroutine reaceq (ni,t,phi,dphidt)
```

```
implicit none
```

```
[...]
```

```

call lreaceq(ncalls,nvar,np,t, phi, dphidt)

return end

c----- subroutine lreaceq (ncalls,nvar,np,temps,phi,dphidt)
implicit none

[...]
ncalls = ncalls+1
carrey = ((np-1)/intspac)*((np-1)/intspac)

do jy=1,np bj(jy) = phi(jy) enddo

c equation dans la premiere cellule
jy = 1 dphidt(jy) = k1*bj(jy)*bj(jy)*(1 - bj(jy))

c Cellule generique
do 550 jy= 2, np-1 dphidt(jy) = k1*bj(jy)*bj(jy)*(1 - bj(jy))
550 enddo
c equation dans la derniere cellule
jy = np
dphidt(jy) = k1*bj(jy)*bj(jy)*(1 - bj(jy))

return
end

c----- subroutine diffeq (ni,t,phi,dphidt)

implicit none

[...]
call ldiffeq(ncalls,nvar,np,t, phi, dphidt)

return
end

c----- subroutine ldiffeq (ncalls,nvar,np,temps,phi,dphidt)

implicit none

[...]
ncalls = ncalls+1

```

```

carrey = ((np-1)/intspac)*((np-1)/intspac)

do jy=1,np
bj(jy) = phi(jy)
enddo

c equation dans la premiere cellule
jy = 1
dphidt(jy) = carrey*D*(bj(jy+1) - bj(jy))

c Cellule generique
do 550 jy= 2, np-1
dphidt(jy) = carrey*D*(bj(jy+1) -2.d0*bj(jy)+ bj(jy-1))
550 enddo

c equation dans la derniere cellule
jy = np
dphidt(jy) = carrey*D*(-bj(jy) + bj(jy-1))

return
end

```

Ensuite, on intègre l'intervalle de temps  $[sss0, sss]$  différemment suivant la méthode de splitting choisie :

- dans le cas de la méthode de Lie RD, on résout d'abord l'équation différentielle avec comme second membre le terme de réaction en prenant comme condition initiale la solution calculée au pas de temps précédent et un pas de temps égal à *deltat*, puis on résout l'équation de diffusion avec comme condition initiale le terme calculé ci-dessus et un pas de temps égal à *deltat*.
- dans le cas de la méthode de Lie DR, on résout d'abord l'équation différentielle avec comme second membre le terme de diffusion en prenant comme condition initiale la solution calculée au pas de temps précédent et un pas de temps égal à *deltat*, puis on résout l'équation avec second membre de réaction avec comme condition initiale le terme calculé ci-dessus et un pas de temps égal à *deltat*.
- dans le cas de la méthode de Strang DRD, on résout d'abord l'équation différentielle avec comme second membre le terme de diffusion en prenant comme condition initiale la solution calculée au temps  $t$  précédent et un pas de temps égal à *deltat*/2, puis on résout l'équation de réaction avec comme condition initiale le vecteur calculé ci-dessus et un pas de temps égal à *deltat*, et enfin on résout l'équation de diffusion avec comme condition initiale le deuxième vecteur calculé ci-dessus et un pas de temps égal à *deltat*/2.
- dans le cas de la méthode de Strang RDR, on résout d'abord l'équation différentielle avec comme second membre le terme de réaction en prenant comme condition initiale le vecteur solution calculée au temps  $t$  précédent et un pas de temps égal à *deltat*/2, puis on résout l'équation de réaction avec comme condition initiale le vecteur calculé ci-dessus et un pas de temps égal à *deltat*, et enfin on résout l'équation de diffusion avec comme condition initiale le

deuxième vecteur calculé ci-dessus et un pas de temps égal à  $deltat/2$ .  
On peut voir l'exemple du code pour la méthode de Lie DR ci-dessous :

```
do 400 while (t.le.inttps)

c Pas de temps de diffusion
sss0diff = t
sssdiff = sss0diff + deltat

IELWRKDIFF(1) = nvar
IELWRKDIFF(2) = nvar

ISTATEDIFF = 1

CALL LSODE(diffeq,ncompo,phi,sss0diff,sssdiff,
& ITOLDIFF,RTOLDIFF,
& ATOLDIFF,ITASKDIFF,ISTATEDIFF,
& IOPTDIFF,ELWRKDIFF,LENWKDIFF,IELWRKDIFF,LENIWKDIFF,
& JOMDIFF,MFDIFF)

c Pas de temps de reaction
sss0reac = t
sssreac = sss0reac + deltat

IELWRKREAC(1) = 1 IELWRKREAC(2) = 1
ISTATEREAC = 1
CALL LSODE(reaceq,ncompo,phi,sss0reac,sssreac,
& ITOLREAC,RTOLREAC,
& ATOLREAC,ITASKREAC,ISTATEREAC,
& IOPTREAC,ELWRKREAC,LENWKREAC,IELWRKREAC,LENIWKREAC,
& JOMREAC,MFREAC)

t = t + deltat
nt = nt + 1

c On repasse ici du vecteur d'integration phi au variables
c de depart b et c

do 420 jy = 1,np
b(nt,jy) = phi(jy)
420 enddo
enddo
```

## 7.4 Calcul de la vitesse

### 7.4.1 Principe de l'algorithme

L'algorithme de calcul de la vitesse consiste à déterminer pour une ordonnée fixée au cours du temps (que l'on prendra égale à 0,5 dans le code) le taux d'accroissement de l'abscisse associée en fonction du temps. Le code de résolution du système différentiel sort 9 fichiers qu'on indice par  $k \in [0, 8]$  et qui représentent les solutions aux temps  $(t_k)_{k \in [0, 8]}$  avec

$$t_k = t_0 + k \frac{t_n - t_0}{8}, \quad 0 \leq k \leq 8$$

Dans chaque fichier donnant la solution quasi-exacte ou splittée on calcule l'abscisse que l'on notera  $x_k$  telle que  $b(x_k) = 0,5$ . Pour cela, l'algorithme de la vitesse détermine la valeur de l'indice  $jy$  telle que  $b(jy) \leq 0,5 \leq b(jy+1)$  en se servant de la décroissance de  $b$  (boucle while). On a alors

$$z_{jy} \leq x_k \leq z_{jy+1}$$

On détermine  $x_k$  en réalisant une interpolation linéaire de la manière suivante. On a

$$0,5 = \alpha b(jy) + (1 - \alpha)b(jy + 1)$$

avec

$$\alpha = \frac{0,5 - b(jy + 1)}{b(jy) - b(jy + 1)}$$

On prend alors comme valeur approchée de  $x_k$

$$x_k \cong \alpha z_{jy} + (1 - \alpha)z_{jy+1}$$

Soit

$$x_k \cong z_{jy} + (1 - \alpha)(z_{jy+1} - z_{jy})$$

On définit alors la vitesse  $v_k$  pour chaque temps  $t_k$  par

$$v_k = \frac{x_k - x_{k-1}}{t_k - t_{k-1}}, \quad k \in [1, 8]$$

On définit alors la vitesse globale par

$$v = \frac{1}{8} \sum_{k=1}^8 v_k, \quad k \in [1, 8]$$

### 7.4.2 Algorithme complet

```
write(6,*) 'Calcul de la vitesse...'
do kk=0,8
preminf = .true.
do jy = 1,np
```



```

read ((20+kk), 7050) zz(jy), b(jy)
if (b(jy).le.0.5d0.and.preminf) then
xx(kk) = zz(jy-1) +
& (b(jy-1)-0.5d0)/(b(jy-1)-b(jy))*(zz(jy)-zz(jy-1)) preminf = .false.
print*, 'test', jy, xx(kk)
endif
enddo
enddo

vit = 0.d0

do k = 1,8
vit = vit + (xx(k) - xx(k-1))/inttps
enddo

```

## 7.5 Calcul des erreurs

### 7.5.1 Norme infinie

On définit la norme infinie (spatiale) de la solution par la formule discrète

$$\|b\|_{\infty} = \max_{jy \in [1, np]} |b(jy)|$$

Elle se calcule dans une boucle for en initialisant la variable *normelinf* à 0, puis en prenant à chaque itération le max de *normelinf* et de *abs(b(jy))*.

### 7.5.2 Norme $L_2$

On définit la norme  $L_2$  (spatiale) de la solution par la formule discrète

$$\|b\|_2^2 = \sum_{jy=2}^{np} b(jy)^2 (z_{jy} - z_{jy-1})$$

Elle se calcule dans une boucle for en initialisant la variable *normel2* à 0, puis en ajoutant  $b(jy)^2(z_{jy} - z_{jy-1})$  à *normel2* pour chaque itération.

### 7.5.3 Algorithme du calcul des erreurs

```

do jy = 1, np
[...
& exacte(1,jy)-b(nt,jy)
c Mise a jour norme linf
if (dabs(exacte(1,jy)-b(nt,jy)).gt.normelinf) then
normelinf = dabs(exacte(1,jy)-b(nt,jy))

```

```

endif
c Mise a jour norme l2
normel2 = normel2 + (exacte(1,jy)-b(nt,jy)) *
& (exacte(1,jy)-b(nt,jy))
enddo
normel2=dsqrt(normel2/np)

```

## 7.6 Etude dans le plan de phase

L'objectif de l'algorithme est de déterminer les diagrammes de phase quasi-exact et splitté, puis d'en faire la différence. Comme vu dans la partie ??, cela revient à déterminer la fonction

$$\beta \mapsto \frac{\partial \beta}{\partial x}$$

Cela pose quelques difficultés dans la mesure où la détermination des dérivées spatiales ne peut se faire que de manière approchée et que les abscisses dont on dispose pour le diagramme de phase quasi-exact ne correspondent pas aux abscisses du diagramme de phase splitté (ces abscisses, qui correspondent aux  $b(jy)$  sont calculées pour les mêmes  $z_{jy}$ , et donc diffèrent selon que l'on utilise l'algorithme quasi-exact ou splitté pour résoudre le système initial).

### 7.6.1 Les variables utilisées

La solution quasi-exacte est désignée par *bexact*, les dérivées spatiales associées par *pexacte* tandis que la solution splittée est désignée par *b*, et ses dérivées spatiales par *pente*. *bexact*, *pexacte* et *b* sont des tableaux à  $np$  composantes, tandis que *pente* est un simple réel (on ne stocke pas en mémoire toutes les pentes splittées).

### 7.6.2 Calcul des dérivées

Le calcul des dérivées se fait en remarquant que

$$f(t + h_1) = f(t) + h_1 f'(t) + \frac{h_1^2}{2} f''(t) + O(h_1^3) \quad (48)$$

$$f(t - h_2) = f(t) - h_2 f'(t) + \frac{h_2^2}{2} f''(t) + O(h_2^3) \quad (49)$$

En faisant la différence entre (48) et (49) on obtient

$$\frac{f(t + h_1) - f(t - h_2)}{h_1 + h_2} = f'(t) + \frac{h_1^2 - h_2^2}{2} f''(t) + O(h_1^3)$$

Comme dans notre cas, on a  $h_1 \approx h_2$ , on en déduit que

$$\frac{f(t + h_1) - f(t - h_2)}{h_1 + h_2} = f'(t) + O(h_1^2 + h_2^2) \quad (50)$$

Ce calcul de la dérivée est donc d'ordre deux en  $h_1$  et est donc plus précis qu'un calcul classique de la forme

$$\frac{f(t + h_1) - f(t)}{h_1} = f'(t) + O(h_1^2)$$

Pour un fichier  $k$ , on calcule la dérivée par

$$\frac{b(jy + 1) - b(jy - 1)}{x(jy + 1) - x(jy - 1)} \quad (51)$$

### 7.6.3 Calcul des différences entre les diagrammes de phase

Le calcul des différences entre les diagrammes de phase se fait en plusieurs étapes :

- Premièrement, le programme calcule l'intégralité du diagramme de phase quasi-exact (pour tous les  $bexact(jy)$  déjà déterminé par le code quasi-exact, il détermine  $pexact(jy)$  par la formule vue dans le paragraphe précédent).
- Ensuite, pour chaque  $b(jy)$  de la solution splittée, on calcule *pente* avec (51). En effet, si  $b(jy)$  représente  $\beta_{sp}$ , pente représente  $\frac{\partial \beta_{sp}}{\partial x} |_{\beta_{sp}}$ . On détermine ainsi la valeur du diagramme de phase splitté en  $b(jy)$ .
- On cherche maintenant à calculer la valeur du diagramme de phase quasi-exact en  $b(jy)$ . Autrement dit, on cherche à calculer  $\frac{\partial bexact}{\partial x} |_{\beta_{sp}}$ . Dans la mesure où les  $(b(jy))$  ne sont pas égaux aux  $(bexact(jy))$ , une interpolation est de nouveau nécessaire. On procède de la manière suivante :
- On recherche l'indice  $l$  tel que  $bexact(jy + l)$  est le plus proche possible de  $b(jy)$ . Pour cela, on se sert de la décroissance des  $bexact(jy)$ . On réalise alors une première boucle while qui incrémente  $l$  jusqu'à ce que  $bexact(jy + l)$  soit plus petit que  $b(jy)$ , et une seconde qui décrémente  $l$  jusqu'à ce que  $bexact(jy + l)$  soit plus grand que  $b(jy)$ . On laisse le soin au lecteur de prouver qu'à l'issue de cette double boucle on a  $bexact(jy + l + 1) \leq b(jy) \leq bexact(jy + l)$ .
- En réalisant une interpolation linéaire similaire à celle faite pour la vitesse, on en déduit l'expression de la pente exacte en  $b(jy)$ , que l'on note *ilpente*

$$ilpente = pexacte(jy + l + 1) + (1 - \alpha)(pexacte(jy + l) - pexacte(jy + l + 1))$$

avec

$$\alpha = \frac{b(jy) - bexact(jy + l + 1)}{bexact(jy + l) - bexact(jy + l + 1)}$$

- Enfin, on calcule la différence  $pente - ilpente$  qui mesure la différence entre le diagramme de phase exact et le diagramme de phase splitté en  $b(jy)$ .

### 7.6.4 Algorithme complet

```
do kk=0,8
erreur=0.d0
pexacte(1)=0.d0
```

```

do cpt=2,(np-1)
read ((40+kk), 7050) bexact(cpt), pexacte(cpt)
c dans les fichiers (40+kk) figurent les diagrammes de phases exacte
c déjà calculés
enddo

do jy = 1,np
read ((20+kk), 7050) x(jy), b(jy)
c dans les fichiers (20+kk) figurent les solutions splittées
c déjà calculées
enddo

do jy = 2, (np-1)
pente = -(b(jy+1) - b(jy-1))/(x(jy+1)-x(jy-1))
write((30+kk), 7050) b(jy), pente

l=0
do while (bexact(jy+1).gt.b(jy))
l=l+1
enddo

do while (bexact(jy+1).lt.b(jy))
l=l-1 enddo

if(dabs(bexact(jy+1+1)-bexact(jy+1)).ge.1e-9) then
alpha=(b(jy)-bexact(jy+1))/(bexact(jy+1+1)-bexact(jy+1))
else
alpha=0.d0
endif

ilpente=pexacte(jy+1)+alpha*
& (pexacte(jy+1+1)-pexacte(jy+1)) diffmin=pente-ilpente

write((50+kk), 7050) b(jy), diffmin

if (dabs(diffmin).gt.erreur) then
erreur = dabs(diffmin)
endif
enddo

```

## Troisième partie

# Resultats

## 8 Modèle KPP

### 8.1 Présentation du programme

Le but du programme est d'effectuer l'intégration numérique de l'équation du modèle KPP :

$$\frac{\partial \beta}{\partial t} = D \frac{\partial^2 \beta}{\partial x^2} + k\beta^2(1 - \beta) \quad (52)$$

L'étude théorique (partie 1 et [3]) a donné une solution analytique de cette équation, qui est une onde se propageant à vitesse constante  $c = \frac{1}{\sqrt{2}}(kD)^{1/2}$ . De plus, Kolmogorov, Petrovskii et Piskunov ont démontré dans leur article historique ([5]) que n'importe quel état initial suffisamment régulier converge vers cette solution analytique. Une première étape consistera à vérifier numériquement ce résultat.

Le programme effectue une discrétisation par méthode des lignes (voir 4). Sauf indication contraire, la discrétisation spatiale est faite avec 5000 points. Ce choix permet d'avoir une précision dans les calculs qui soit suffisante pour être sûr d'estimer correctement l'erreur effectuée lors du splitting, tout consommant un temps de calcul et un espace mémoire raisonnables, comme on le montre dans la partie 8.3.

Au bord, on prend des conditions aux limites de Neumann : en effet, on s'arrange pour que le front d'onde reste suffisamment éloigné des bords pour que le gradient de la solution y soit nul à la précision machine.

Le solveur utilisé est LSODE (partie 5). Toutes les tolérances sont fixées à  $10^{-14}$ , ce qui permet de s'assurer que les erreurs dues à la méthode numérique d'intégration temporelle (en l'occurrence la méthode BDF) soient négligeables devant toutes les autres erreurs que l'on va mesurer, et donc de pouvoir considérer que la solution donnée par LSODE est la solution "exacte" du système, discrétisé spatialement, qui lui est passé en argument.

La solution est sauvée, de façon à pouvoir être visualisée, à 8 instants régulièrement répartis : ceci permet d'avoir un bon aperçu du phénomène sur le temps considéré.

### 8.2 Convergence vers la solution analytique

Dans un premier temps, nous regardons l'évolution d'un état initialement discontinu. Les paramètres sont  $D = 1$ ,  $k = 1$ , 5000 points de discrétisation, intervalle en  $x$  :  $[0, 80]$ , intervalle de temps  $[0, 50]$ . Graphiquement, on observe que la solution numérique prend la forme de la solution analytique, en accord avec le résultat démontré par KPP [5] : on peut en effet considérer que la fonction est suffisamment régulière pour que le résultat lui soit applicable, il suffit de voir la donnée initiale comme la discrétisation d'une fonction  $\mathcal{C}^\infty$  valant 1 en  $-\infty$  et 0 en  $+\infty$  et ayant une pente très forte localisée sur la zone de "discontinuité".

La vitesse calculée : 0,6804, n'est pas trop éloignée de la vitesse de la solution analytique :  $\frac{1}{\sqrt{2}} = 0,7071$ , ce qui confirme cette première observation. Cependant, nous n'étudierons pas

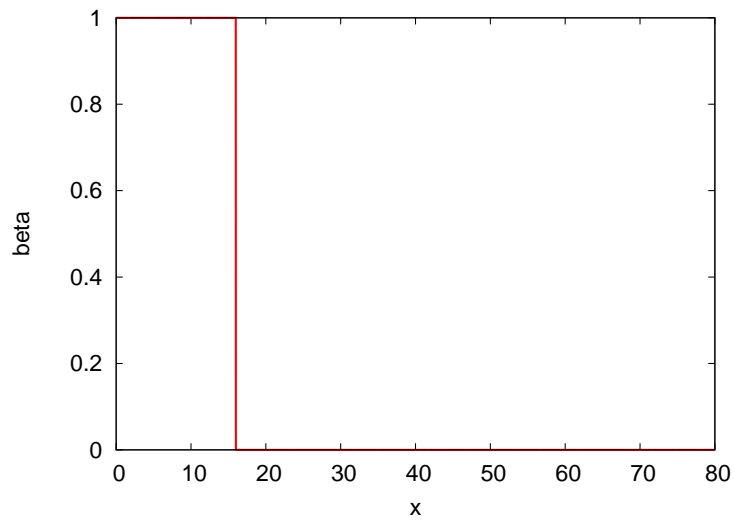


FIG. 11 – Donnée initiale discontinue

plus en détail le phénomène : la discontinuité introduite par l'état initial entraîne une perturbation importante du système, et un calcul d'"erreur" n'aurait pas de sens ici.

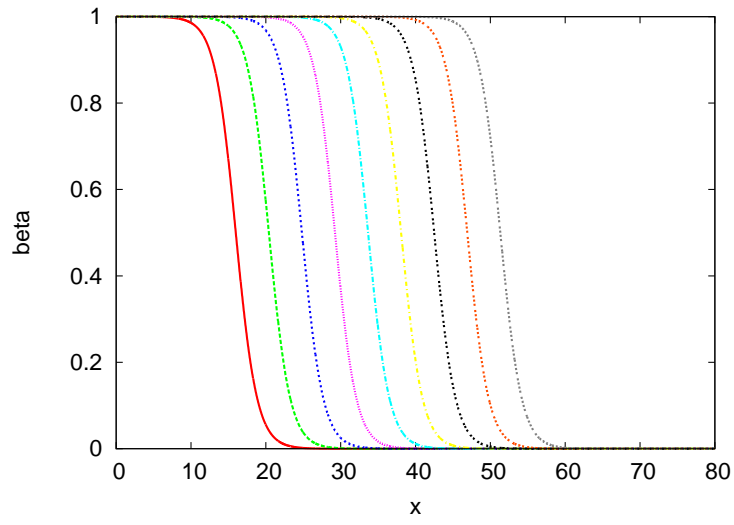


FIG. 12 – Solution exacte du système

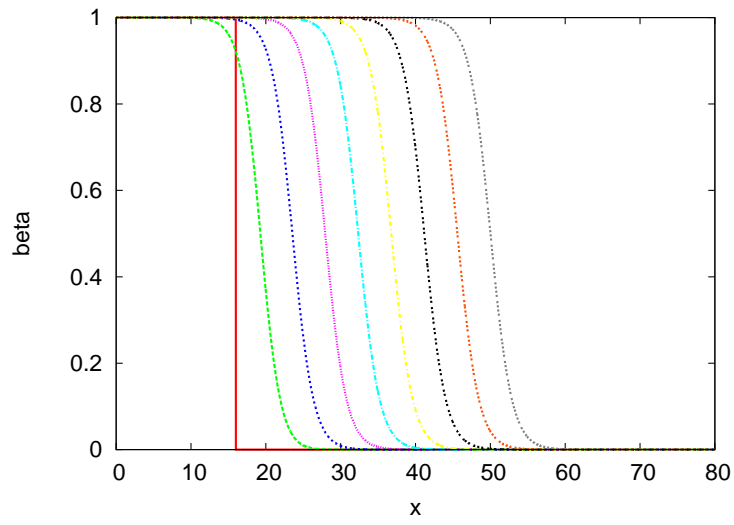


FIG. 13 – Solution numérique calculée à partir de l'état discontinu

### 8.3 Influence de la discrétisation par la méthode des lignes

Nous cherchons maintenant à estimer l'influence de la discrétisation par la méthode des lignes sur la solution de l'équation aux dérivées partielles. Nous sommes ici dans un cas particulièrement favorable pour effectuer cette étude, puisque nous connaissons la solution exacte de l'EDP (52). L'idée est donc de prendre comme condition initiale cette solution exacte, et de visualiser l'évolution de la différence entre l'onde exacte et l'onde calculée. Les tolérances du solveur étant fixées à un très faible niveau, nous pouvons considérer que nous observons ainsi la différence entre la solution exacte de (52) et la solution exacte du système discrétisé. On se place dans le cas  $D = 1$  et  $k = 1$  et on effectue l'étude selon différentes discrétisations spatiales.

- Avec 5000 points de discrétisation spatiale :

Les nouveaux paramètres étant : intervalle en  $x$   $[0, 140]$ , intervalle de temps  $[0, 30]$ .

La vitesse calculée est de 0,70696, vitesse légèrement inférieure à la vitesse de l'onde exacte, ce qui explique en partie pourquoi l'erreur augmente au cours du temps. L'erreur maximale obtenue, de l'ordre de  $10^{-5}$ , est relativement importante. En effet, on verra que lorsque l'on effectue du splitting, la différence entre la solution calculée en splitting et l'onde calculée sans splitting peut être de l'ordre de  $10^{-6}$ . Ceci justifie de prendre comme référence dans ces calculs cette dernière, que l'on appellera *onde quasi-exacte*, et non la solution analytique.



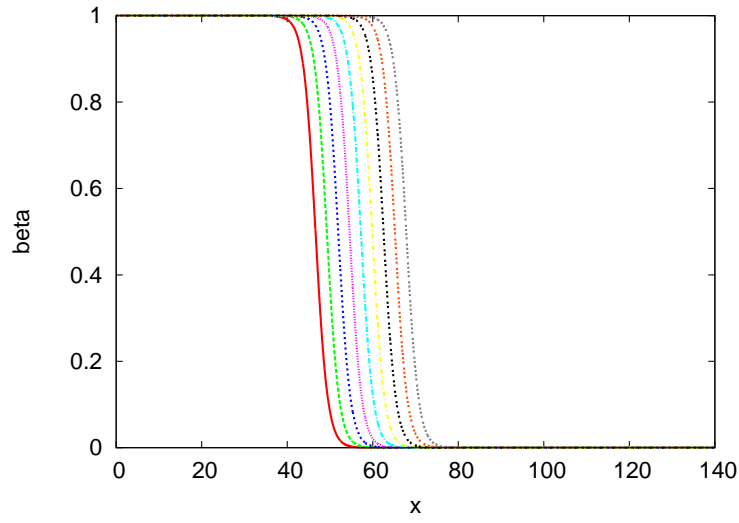


FIG. 14 – Evolution du système à partir de l'onde analytique

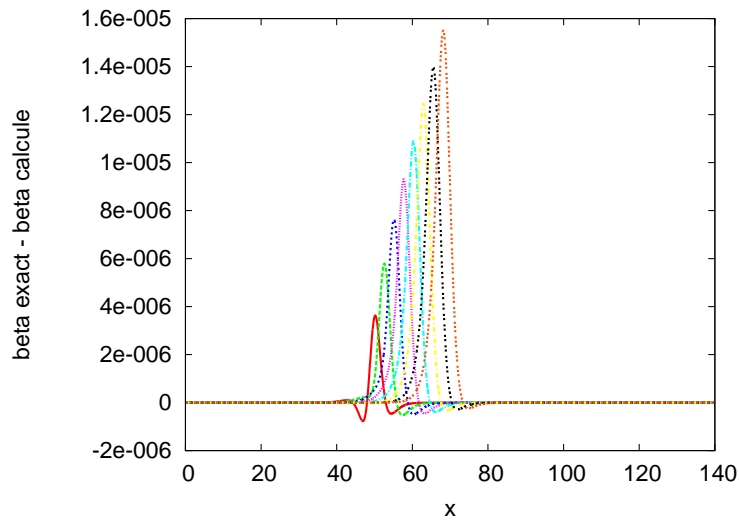


FIG. 15 – Différence entre l'onde analytique et l'onde calculée

– Autres discrétisations :

Les résultats trouvés sont résumés dans le tableau 2. On observe qu'il y a bien convergence vers la solution analytique lorsqu'on augmente la discrétisation. Cependant, un calcul avec 30000 points, par exemple, demande un temps de calcul dix fois supérieur au calcul avec 5000 points, pour obtenir un niveau de précision qui n'est pas nécessaire pour pouvoir observer des résultats intéressants lors de l'étude de splitting. Le choix de 5000 points constitue un bon compromis.

Discrétisation (points)	Vitesse de l'onde	Erreur maximale	Temps de calcul (s)
Solution exacte	0,70711	-	-
40000	0,70709	$2,42.10^{-7}$	42,5
30000	0,70708	$4,30.10^{-7}$	31,1
20000	0,70707	$9,68.10^{-7}$	18,5
10000	0,70704	$3,87.10^{-6}$	8,29
5000	0,70696	$1,55.10^{-5}$	3,71
2000	0,70673	$9,69.10^{-5}$	1,07
1000	0,70633	$3,87.10^{-4}$	0,51
500	0,70541	$1,55.10^{-3}$	0,26
200	0,70186	$9,66.10^{-3}$	0,08

TAB. 2 – Comparaison des résultats obtenus selon la discrétisation choisie

## 8.4 Etude des résultats obtenus en résolvant par splitting (cas non raide)

Nous cherchons maintenant à vérifier numériquement que l'erreur commise en résolvant l'équation différentielle par splitting est cohérente avec ce que donnent les résultats théoriques : ceux-ci prévoient (partie 6) un ordre global de 1 pour les formules de Lie et de 2 pour la formule de Strang, et nous allons voir que l'on retrouve parfaitement ces résultats dans le cas "non raide"  $k = 1$   $D = 1$ .

### 8.4.1 Splitting de Lie RD

On s'intéresse d'abord au splitting de Lie en effectuant alternativement un pas de temps de réaction, puis un pas de temps de diffusion (schéma "R-D" comme défini en 6.2.1). On étudie la norme  $L2$  de la différence entre l'onde ainsi obtenue  $\beta_{sp}$  et l'onde quasi-exacte  $\beta_{qe}$  pour des pas de temps que l'on choisit tels que leurs logarithmes soient régulièrement espacés :  $\Delta t = \frac{30}{2048}, \frac{30}{1024}, \frac{30}{512}, \dots, \frac{30}{16}$ . Conformément à la théorie, on obtient ainsi un ordre très proche de 1.

Pour effectuer une observation plus en détail de ce qu'il se passe, on trace la différence entre l'onde quasi-exacte et l'onde splittée pour trois pas de temps différents, ainsi que la vitesse de l'onde en fonction du logarithme du pas de temps. Il apparaît clairement que, pour des pas de temps trop grands, la vitesse chute de manière quasi-exponentielle, bien que la structure de l'onde semble être conservée (en particulier, la différence a le même profil pour les trois pas de temps considérés, nous étudierons plus en détail cette remarque en IV). Pour confirmer l'observation d'une chute exponentielle de vitesse, on trace  $\ln(v_{qe} - v_{sp})$  en fonction de  $\ln(\Delta t)$  (où  $v_{sp}$  désigne la vitesse de l'onde obtenue par splitting pour le pas de temps considéré et  $v_{qe}$  la vitesse de l'onde quasi-exacte), et on observe une pente qui vaut 1 pour les pas de temps suffisamment petits.

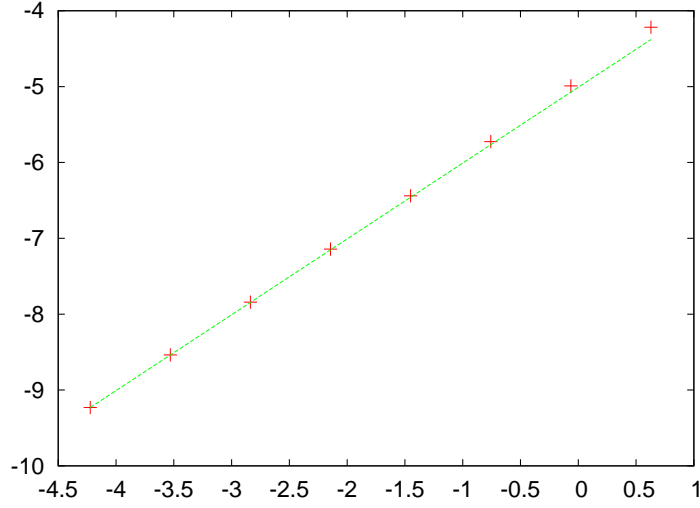


FIG. 16 – Erreur en splitting de Lie RD. En abscisse :  $\ln(\Delta t)$ . En ordonnée :  $\ln(\|\beta_{qe} - \beta_{sp}\|_2)$ . En vert : droite de pente 1

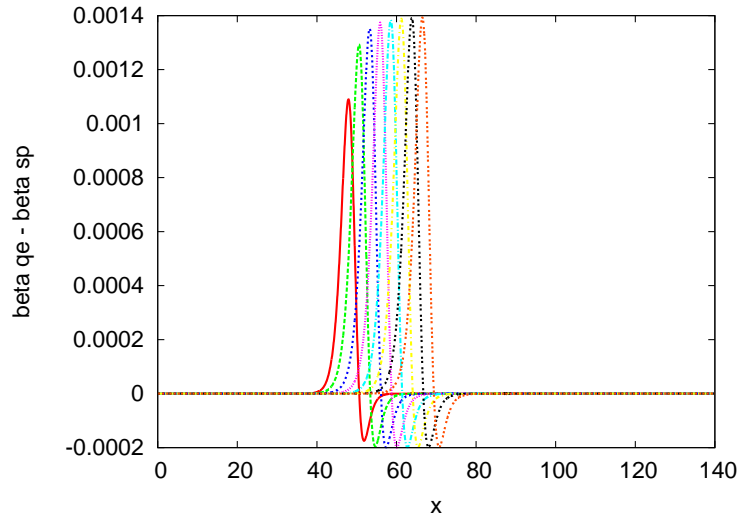


FIG. 17 – Splitting RD : différence  $\beta_{qe} - \beta_{sp}$  en fonction de  $x$ . Pas de temps : 30/1024

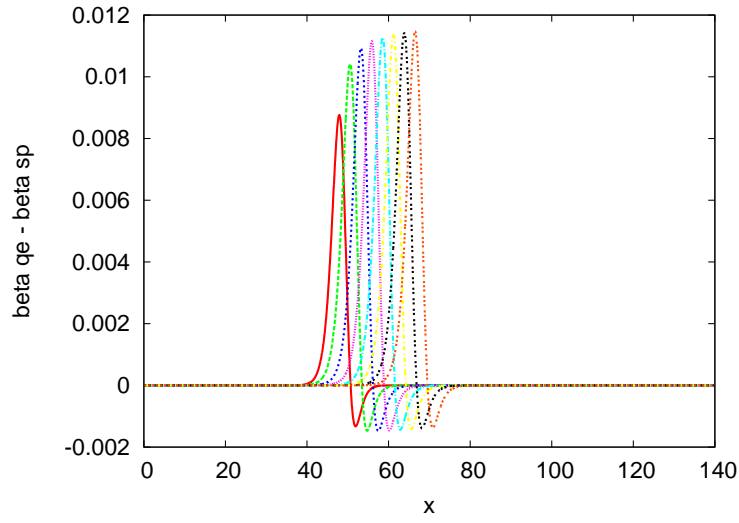


FIG. 18 – Splitting RD : différence  $\beta_{qe} - \beta_{sp}$  en fonction de  $x$ . Pas de temps : 30/128

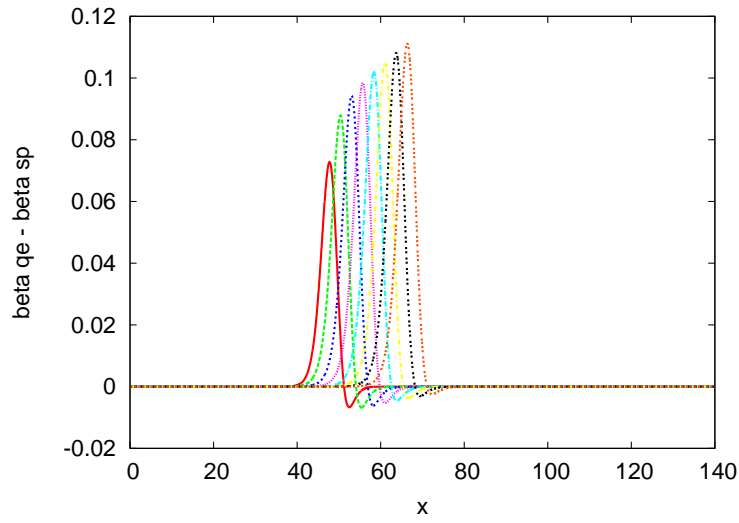


FIG. 19 – Splitting RD : différence  $\beta_{qe} - \beta_{sp}$  en fonction de  $x$ . Pas de temps : 30/16

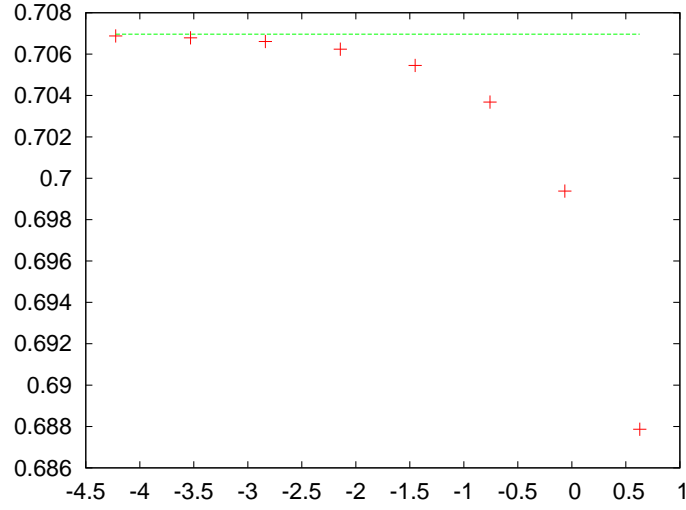


FIG. 20 – Splitting RD : vitesse de l'onde en fonction de  $\ln(\Delta t)$ . En vert : vitesse de l'onde quasi-exacte.

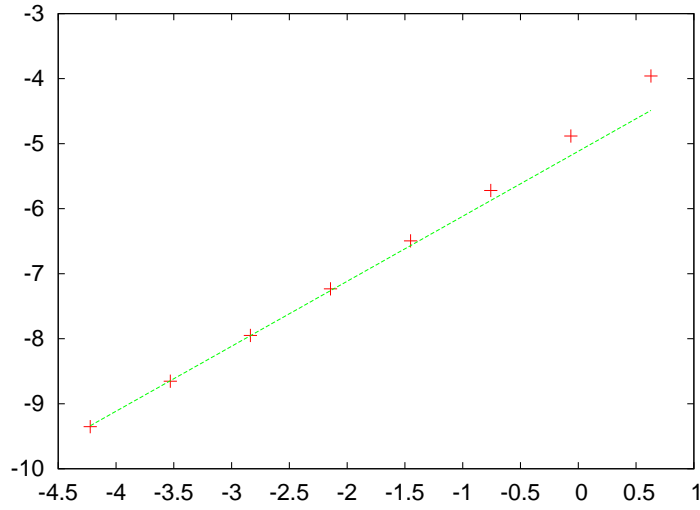


FIG. 21 – Splitting RD :  $\ln(v_{qe} - v_{sp})$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 1.

### 8.4.2 Splitting de Lie DR

Conformément aux attentes, on observe que l'ordre reste proche de 1 lorsqu'on effectue d'abord un pas de temps de diffusion, puis un pas de temps de réaction (DR).

Il est intéressant de comparer les profils de différences d'ondes avec le cas RD (ici pour les mêmes pas de temps 30/1024, 30/128) : les courbes semblent être symétriques par rapport à l'axe des abscisses. Ceci s'explique parfaitement : le coefficient de  $t^2$  dans l'étude théorique (cf partie 6.2.1) est  $\frac{1}{2}[A, B]$  pour le splitting A-B et  $\frac{1}{2}[B, A] = -\frac{1}{2}[A, B]$  pour le splitting B-A.

On observe cette fois-ci une *croissance* exponentielle de la vitesse, ce que confirme l'étude de son logarithme. La pente est à nouveau de 1 pour les petits pas de temps.

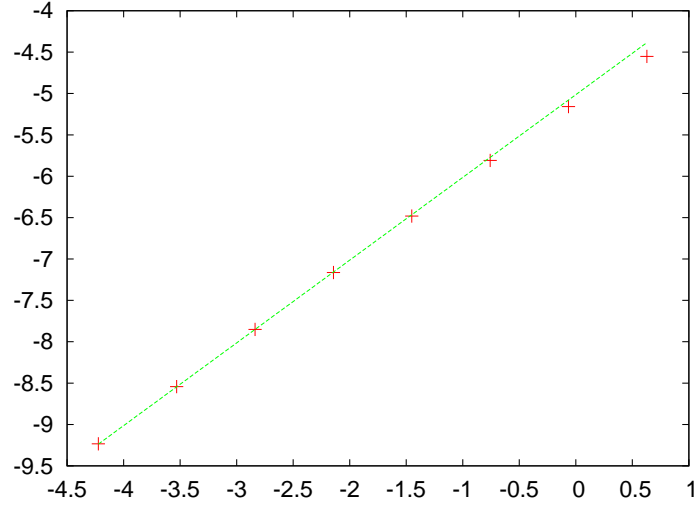


FIG. 22 – Erreur en splitting de Lie DR. En abscisse :  $\ln(\Delta t)$ . En ordonnée :  $\ln(\|\beta_{qe} - \beta_{sp}\|_2)$ . En vert : droite de pente 1

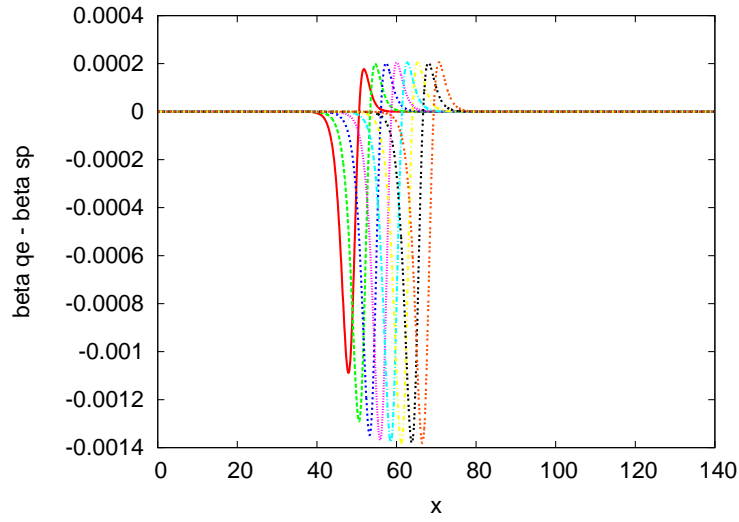


FIG. 23 – Splitting DR : différence  $\beta_{qe} - \beta_{sp}$  en fonction de  $x$ . Pas de temps : 30/1024



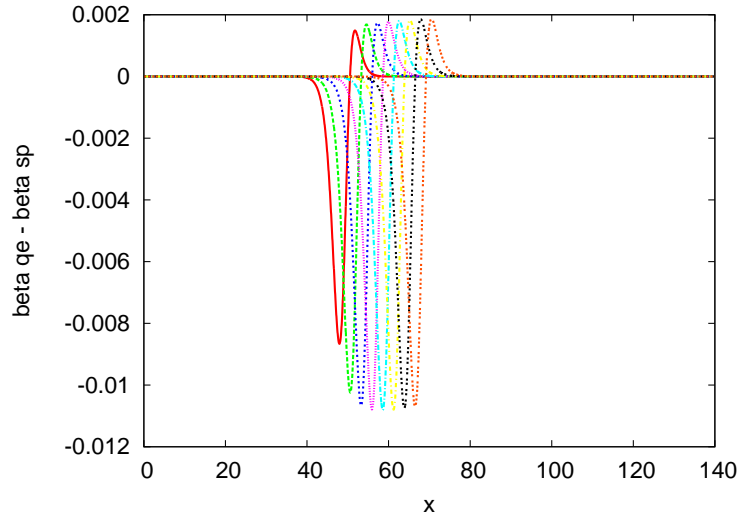


FIG. 24 – Splitting DR : différence  $\beta_{qe} - \beta_{sp}$  en fonction de  $x$ . Pas de temps : 30/128

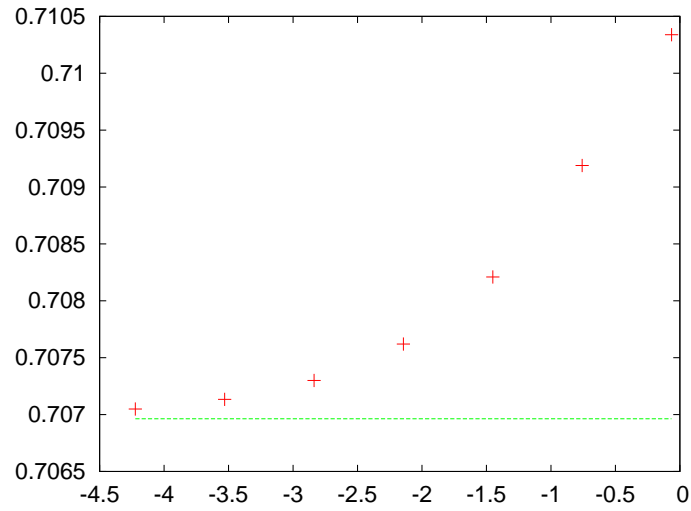


FIG. 25 – Splitting DR : vitesse de l'onde en fonction de  $\ln(\Delta t)$ . En vert : vitesse de l'onde quasi-exacte.

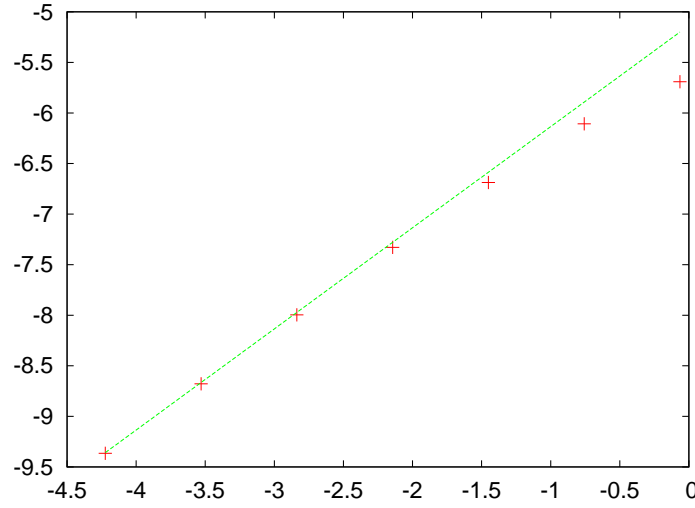


FIG. 26 – Splitting DR :  $\ln(v_{sp} - v_{qe})$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 1.

#### 8.4.3 Splitting de Strang RDR

On effectue la même opération mais cette fois-ci la méthode de splitting est celle de Strang : un demi-pas de temps de réaction, un pas de temps de diffusion et à nouveau un demi-pas de temps de réaction. Dans le cas  $k = 1$   $D = 1$ , on obtient parfaitement l'ordre 2.

On donne à nouveau la différence l'onde quasi-exacte et l'onde splitée est donnée pour des pas de temps de 30/1024, 30/64 : on peut observer que l'ordre de grandeur de la norme  $L^\infty$  de l'erreur chute beaucoup par rapport aux méthodes d'ordre 1 : ici  $10^{-6}$  à comparer à  $10^{-3}$  pour un pas de temps 30/1024. La vitesse a la même allure, cette fois ci parfaitement exponentielle de coefficient exactement égal à 2 comme on peut l'observer.

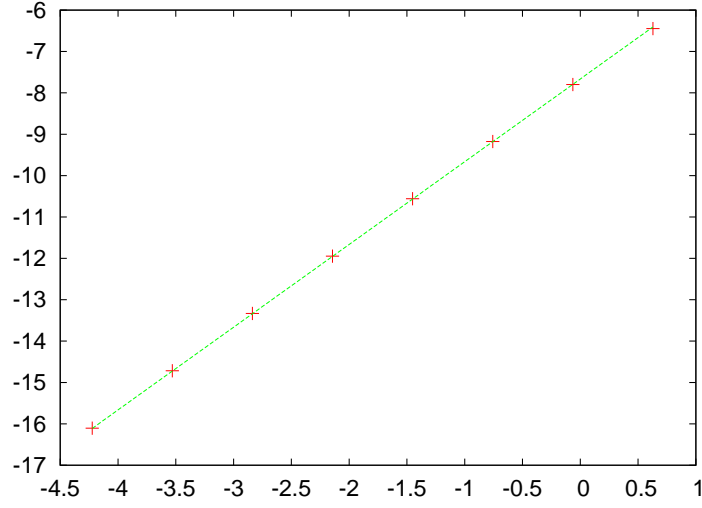


FIG. 27 – Erreur en splitting de Strang RDR. En abscisse :  $\ln(\Delta t)$ . En ordonnée :  $\ln(\|\beta_{qe} - \beta_{sp}\|_2)$

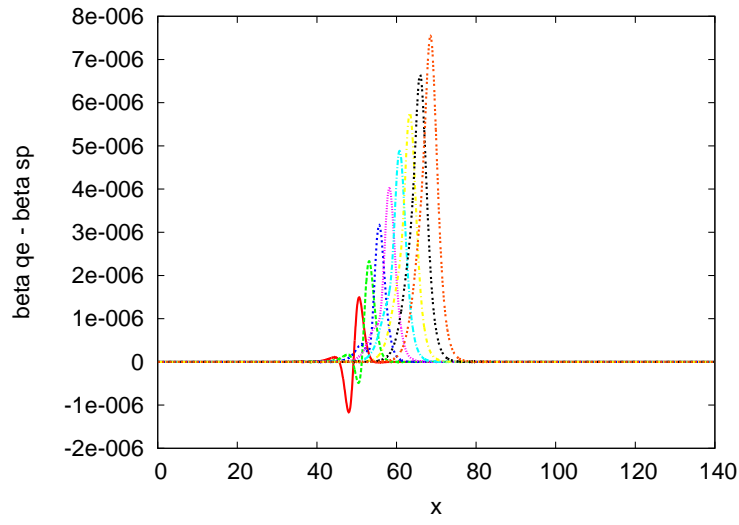


FIG. 28 – Splitting RDR : différence  $\beta_{qe} - \beta_{sp}$  en fonction de  $x$ . Pas de temps : 30/1024

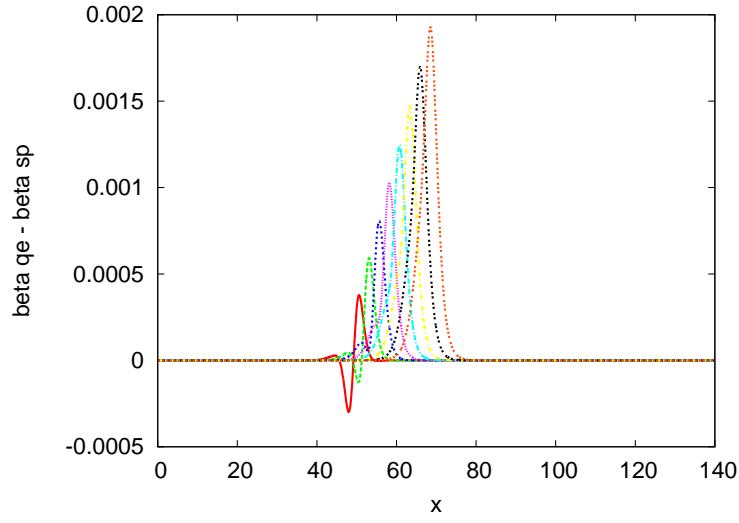


FIG. 29 – Splitting RDR : différence  $\beta_{qe} - \beta_{sp}$  en fonction de  $x$ . Pas de temps : 30/64

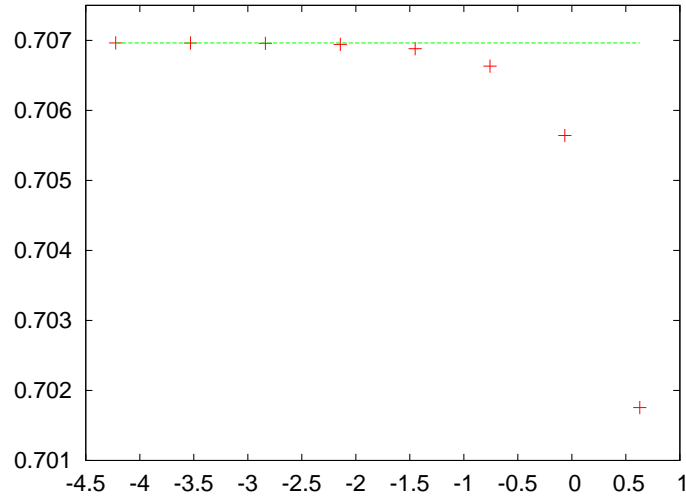


FIG. 30 – Splitting RDR : vitesse de l'onde en fonction de  $\ln(\Delta t)$ . En vert : vitesse de l'onde quasi-exacte.

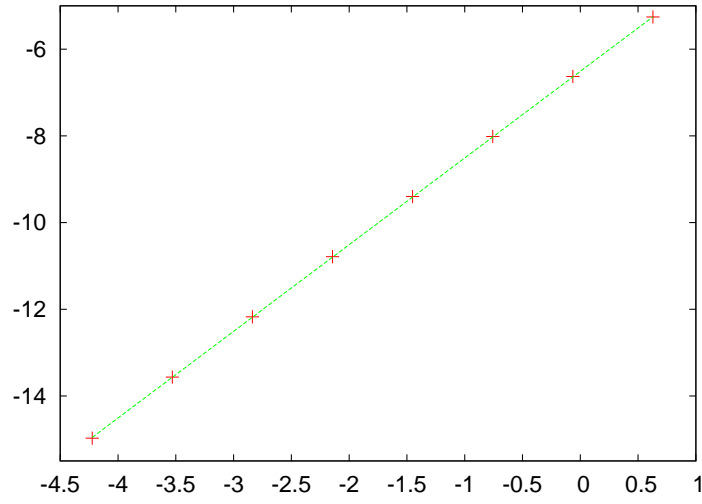


FIG. 31 – Splitting RDR :  $\ln(v_{qe} - v_{sp})$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 2.

#### 8.4.4 Splitting de Strang DRD

On retrouve aussi l'ordre 2 lorsqu'on effectue le splitting dans l'ordre DRD. La vitesse a à nouveau une décroissance exponentielle de coefficient 2.

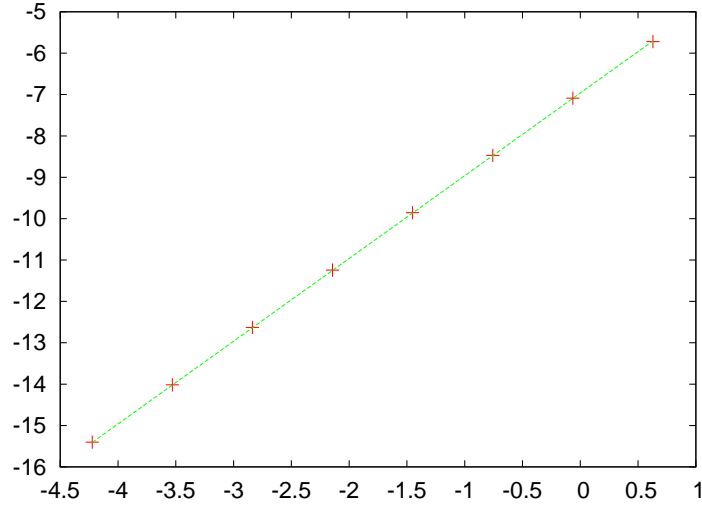


FIG. 32 – Erreur en splitting de Strang DRD. En abscisse :  $\ln(\Delta t)$ . En ordonnée :  $\ln(\|\beta_{qe} - \beta_{sp}\|_2)$

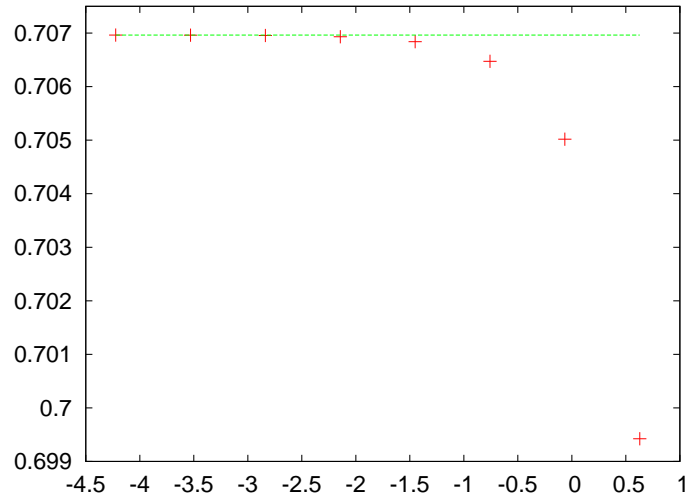


FIG. 33 – Splitting DRD : vitesse de l'onde en fonction de  $\ln(\Delta t)$ . En vert : vitesse de l'onde quasi-exacte.

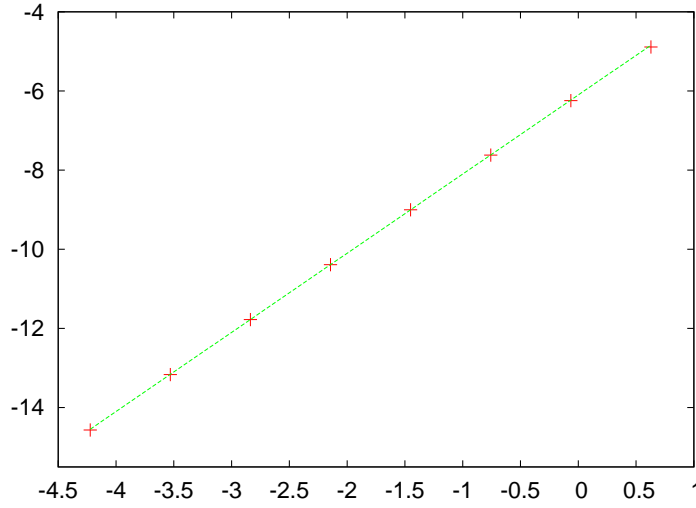


FIG. 34 – Splitting DRD :  $\ln(v_{ge} - v_{sp})$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 2.

#### 8.4.5 Conclusion

Il est frappant de voir que, dans ce cas non raide, l'erreur globale reproduit exactement l'erreur locale : on perd exactement un ordre, comme dans le cas linéaire (cf 6.5). Ce phénomène n'est plus très bien observé dans le cas raide.

### 8.5 Introduction de raideur dans le système : pertes d'ordre en splitting

Rappelons que l'analyse théorique montre que l'onde exacte se propage à la vitesse  $c = \frac{1}{\sqrt{2}}(kD)^{1/2}$  et que son profil admet pour pente maximale  $p = -\frac{1}{\sqrt{32}}(\frac{k}{D})^{1/2}$ . En faisant varier  $k$  et  $D$ , on fait donc varier ces deux paramètres.

Introduisons de la raideur dans le système en augmentant  $p$  et en gardant la même vitesse :  $k = 10$   $D = 0,1$ . Nous avons vu lors de l'étude théorique qu'une condition initiale dont la dérivée prend des valeurs importantes, comme c'est le cas ici, peut être à l'origine de pertes d'ordres : c'est ce résultat que nous allons retrouver ici.

La première chose que l'on remarque est que l'onde quasi-exacte approche de manière beaucoup moins précise l'onde exacte que dans le cas  $k = 1$   $D = 1$ . Avec 5000 points de discrétisation l'erreur maximale devient  $2,47 \cdot 10^{-2}$  alors qu'elle était de  $1,55 \cdot 10^{-5}$  dans le cas non raide. En augmentant la discrétisation à 30000 points, on trouve  $6,88 \cdot 10^{-4}$  au lieu de  $4,30 \cdot 10^{-7}$  (perte de trois ordres de grandeur).

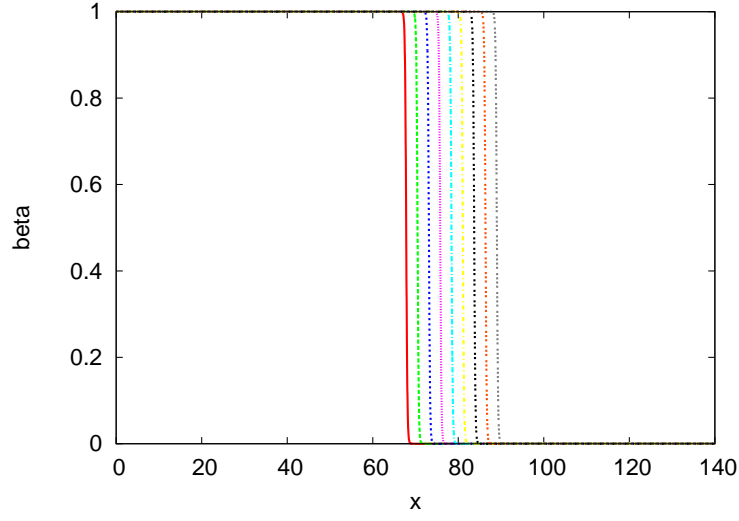


FIG. 35 – Solution quasi-exacte du système raide  $k = 10$ ,  $D = 0, 1$

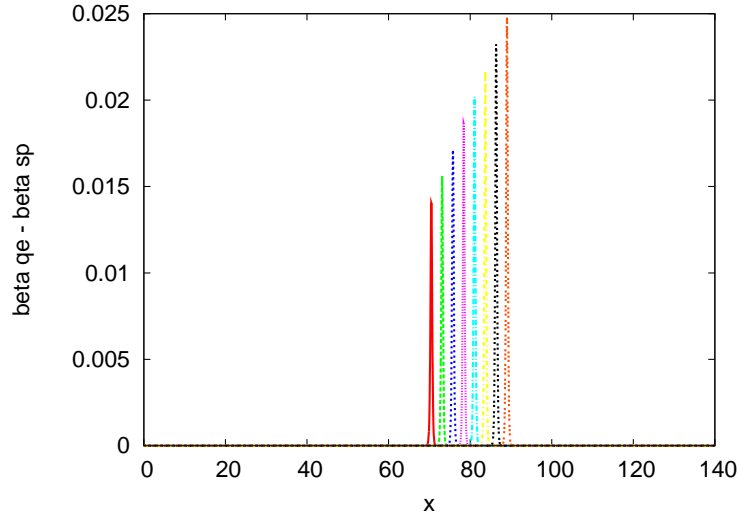


FIG. 36 – Différence entre l'onde analytique et l'onde quasi-exacte - cas raide  $k = 10$ ,  $D = 0, 1$  - 5000 points de discrétisation.



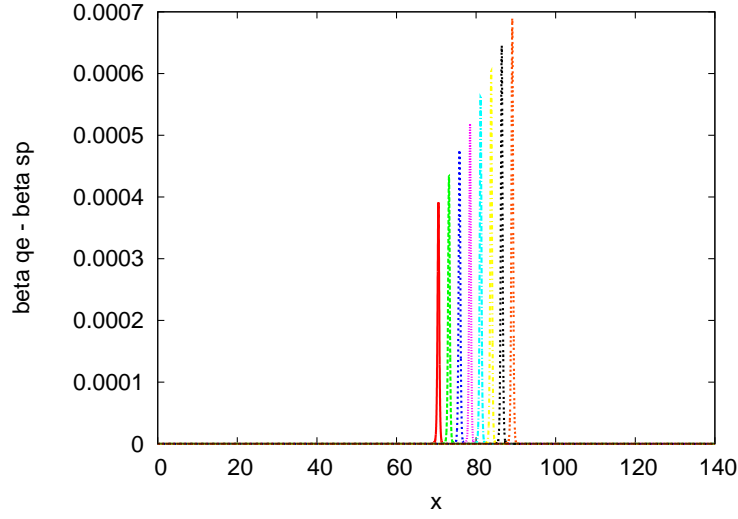


FIG. 37 – Différence entre l'onde analytique et l'onde quasi-exacte - cas raide  $k = 10$ ,  $D = 0, 1$  - 30000 points de discrétisation.

### 8.5.1 Splitting de Lie

**Splitting RD** En splitting de Lie RD, avec la raideur introduite, on observe une perte d'ordre dès que les pas de temps deviennent plus grands que  $\frac{30}{2048}$  : il faut donc chercher des pas de temps bien plus petits que dans le cas non raide pour obtenir l'ordre 1. Ceci est confirmé par l'étude du logarithme de la différence de vitesse entre l'onde splittée et l'onde quasi-exacte, qui montre aussi une dégénérescence par rapport à l'ordre 1.

Pour comprendre ce qu'il se passe, on peut visualiser les écarts de l'onde splittée à l'onde quasi-exacte pour plusieurs pas de temps. On se rend compte que la vitesse devient extrêmement faible pour les petits pas de temps : en effet pour un pas de temps  $30/32$  par exemple l'onde a perdu plus de 50% de sa vitesse, comme on peut le voir sur le graphique.

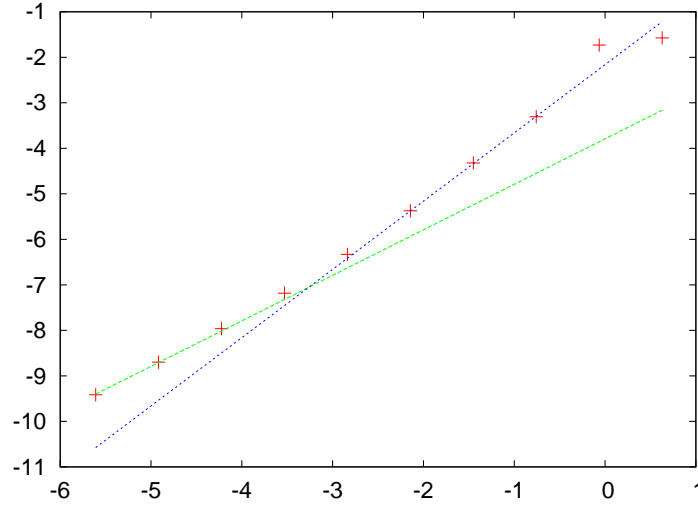


FIG. 38 – Erreur en splitting de Lie RD dans le cas raide  $k = 10$ ,  $D = 0,1$ . En abscisse :  $\ln(\Delta t)$ . En ordonnée :  $\ln(\|\beta_{qe} - \beta_{sp}\|_2)$  En vert : droite de pente 1 ; en bleu : droite de pente 1,5

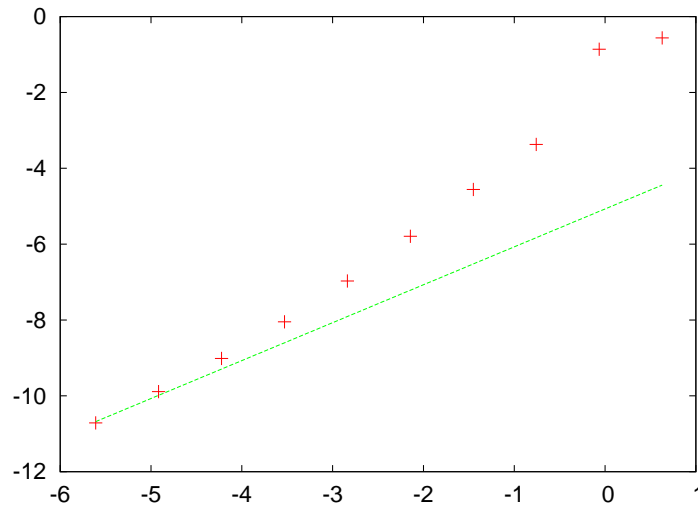


FIG. 39 – Splitting RD raide :  $\ln(v_{qe} - v_{sp})$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 1.

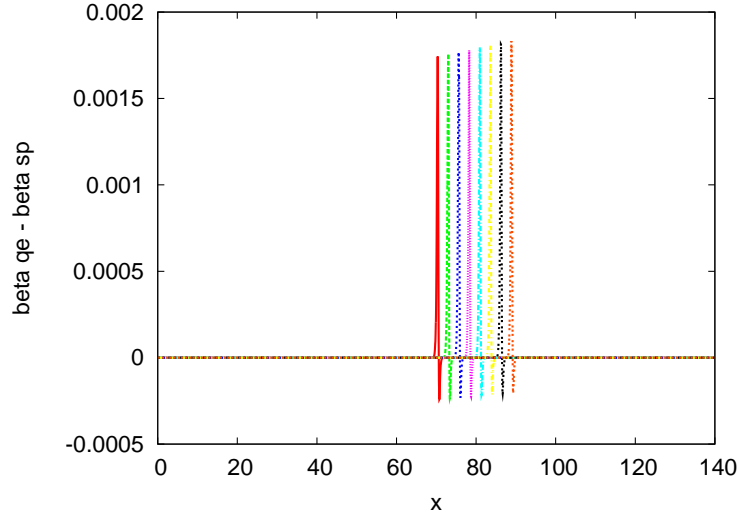


FIG. 40 – Splitting RD raide : différence  $\beta_{qe} - \beta_{sp}$  en fonction de  $x$ . Pas de temps : 30/8192

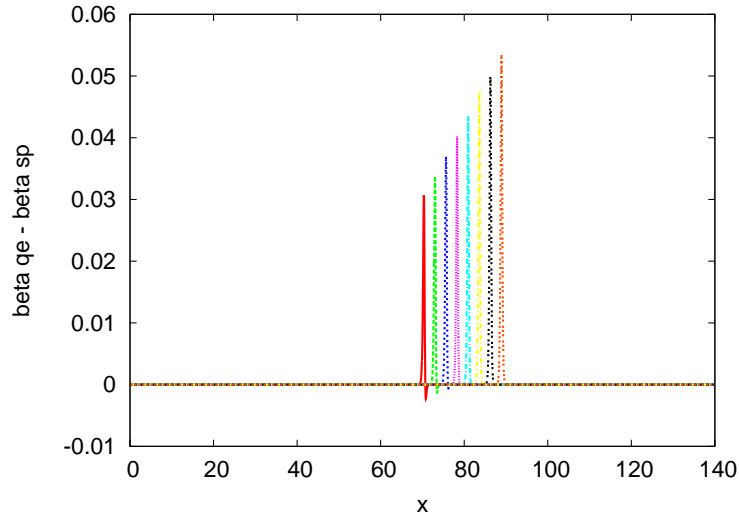
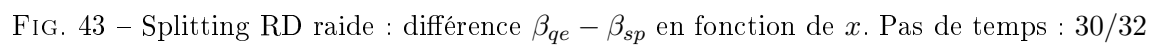
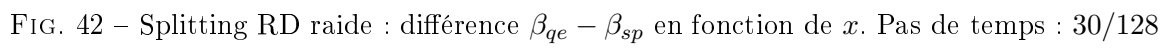


FIG. 41 – Splitting RD raide : différence  $\beta_{qe} - \beta_{sp}$  en fonction de  $x$ . Pas de temps : 30/512



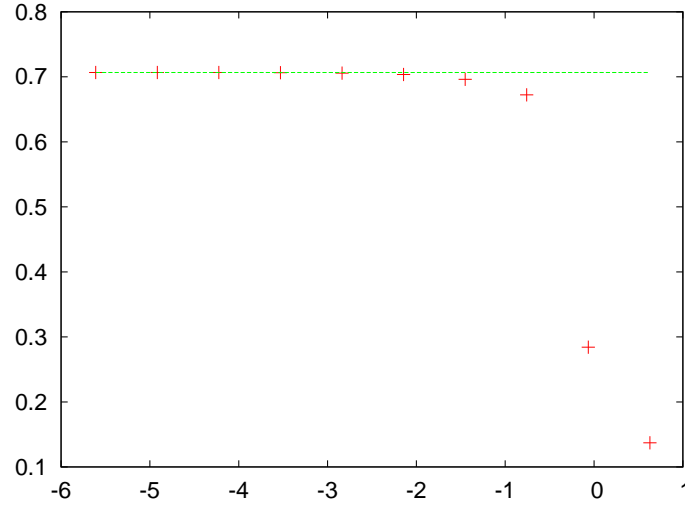


FIG. 44 – Splitting RD raide : vitesse de l’onde en fonction de  $\ln(\Delta t)$ . En vert : vitesse de l’onde quasi-exacte.

**Splitting DR** En splitting DR, le comportement est encore surprenant, en particulier en ce qui concerne la vitesse : celle-ci prend des valeurs beaucoup plus importantes que celles prédites par l’ordre 1. On voit très bien ici que cette approximation, qui fonctionnait dans le cas non raide, n’est plus du tout valable. Notons de plus que la vitesse décroît quand le pas de temps augmente, ce qui est l’inverse de ce qu’il se passait au cas DR non raide. Ce résultat plutôt étonnant est peut-être dû à une erreur de code ou de manipulation...

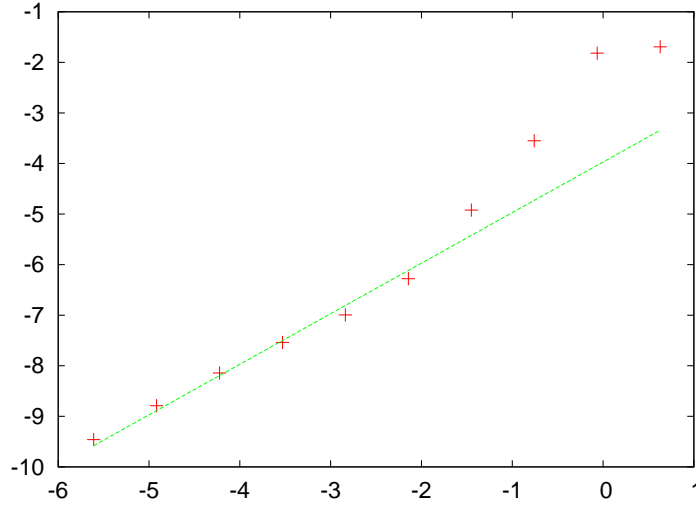


FIG. 45 – Erreur en splitting de Lie DR dans le cas raide  $k = 10$ ,  $D = 0,1$ . En abscisse :  $\ln(\Delta t)$ . En ordonnée :  $\ln(\|\beta_{qe} - \beta_{sp}\|_2)$  En vert : droite de pente 1

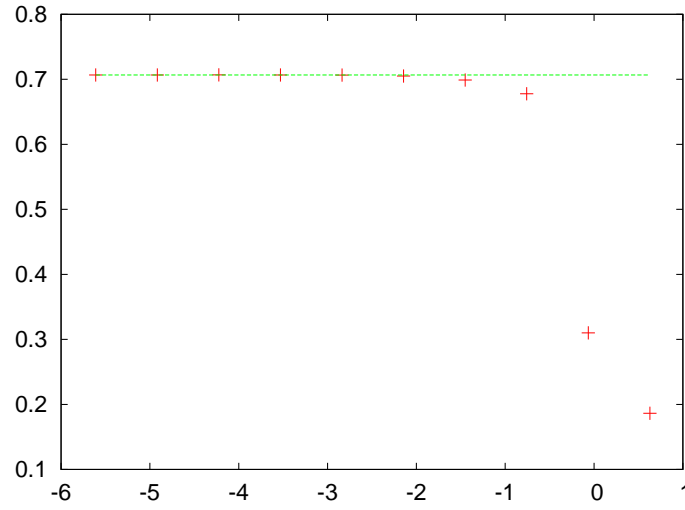


FIG. 46 – Splitting DR raide : vitesse de l'onde en fonction de  $\ln(\Delta t)$ . En vert : vitesse de l'onde quasi-exacte.

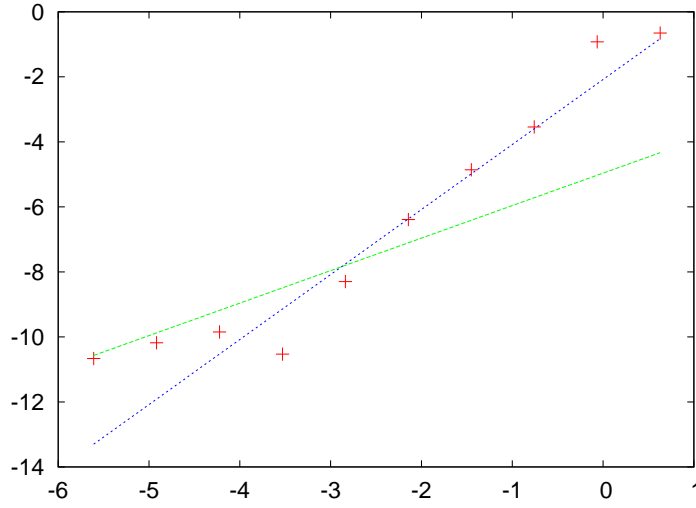


FIG. 47 – Splitting DR raide :  $\ln(v_{qe} - v_{sp})$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 1. En bleu : droite de pente 2

### 8.5.2 Splitting de Strang

**Splitting DRD** Le splitting de Strang DRD permet de conserver l'ordre 2 pour des pas de temps allant jusqu'à  $\frac{30}{128}$ . Au delà, nous pouvons observer une perte d'ordre : on obtient presque un ordre 1,5. Il se passe le même processus de perte de vitesse que dans le cas de Lie, mais seulement pour des pas de temps plus petits et dans de moindres proportions.

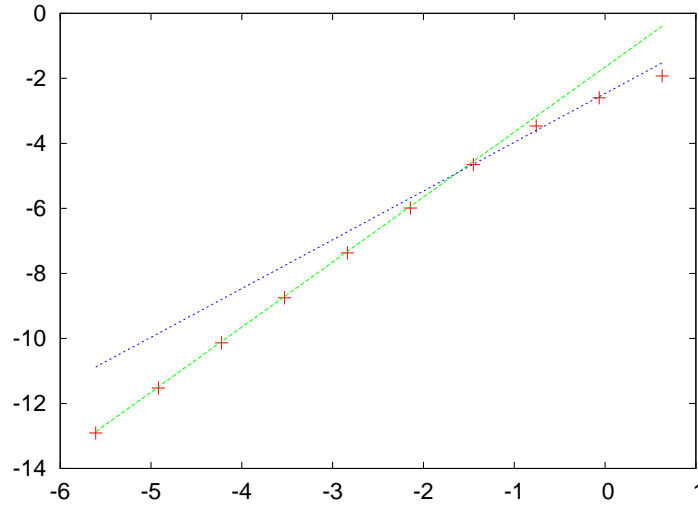


FIG. 48 – Erreur en splitting de Strang DRD dans le cas raide  $k = 10$  et  $D = 0, 1$ . En abscisse :  $\ln(\Delta t)$ . En ordonnée :  $\ln(\|\beta_{qe} - \beta_{sp}\|_2)$ . En vert : droite de pente 2, en bleu : droite de pente 1.5

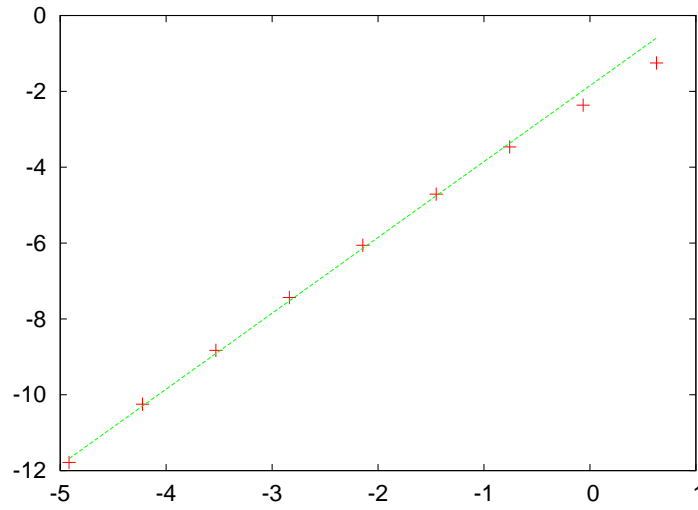


FIG. 49 – Splitting DRD raide :  $\ln(v_{qe} - v_{sp})$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 2.



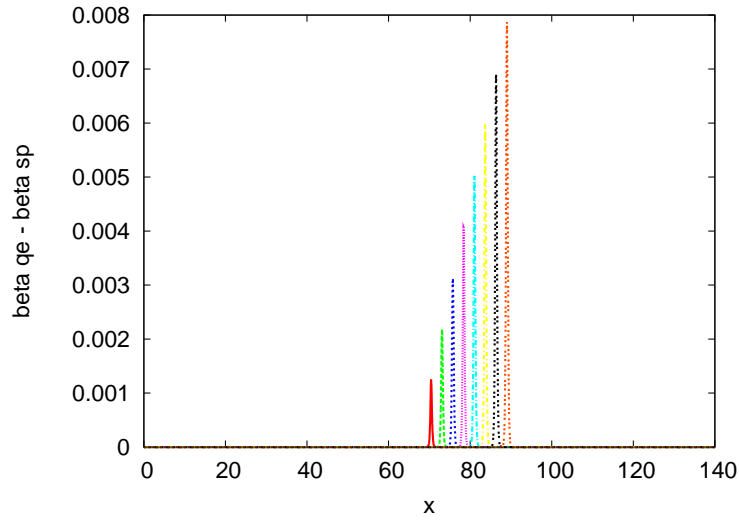


FIG. 50 – Splitting DRD raide : différence  $\beta_{qe} - \beta_{sp}$  en fonction de  $x$ . Pas de temps : 30/1024

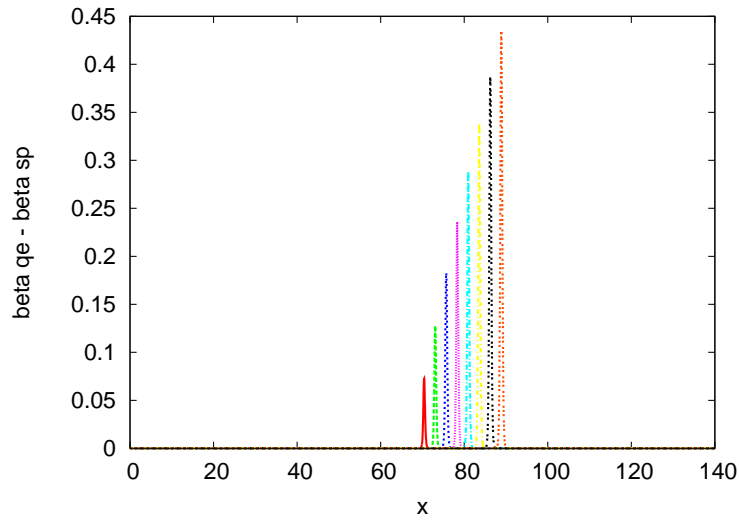


FIG. 51 – Splitting DRD raide : différence  $\beta_{qe} - \beta_{sp}$  en fonction de  $x$ . Pas de temps : 30/128

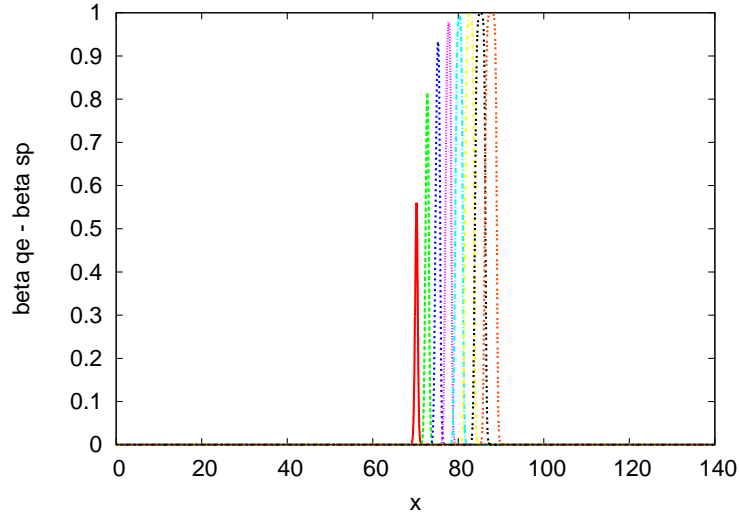


FIG. 52 – Splitting DRD raide : différence  $\beta_{qe} - \beta_{sp}$  en fonction de  $x$ . Pas de temps : 30/32

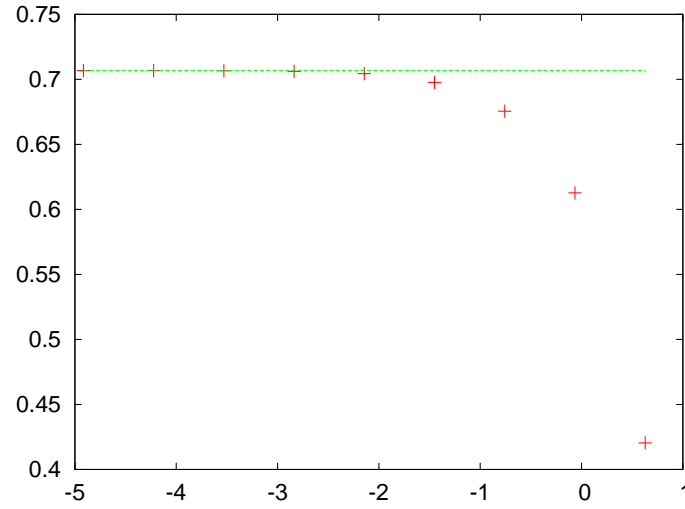


FIG. 53 – Splitting DRD raide : vitesse de l'onde en fonction de  $\ln(\Delta t)$ . En vert : vitesse de l'onde quasi-exacte.

**Splitting RDR** Le splitting RDR donne des résultats similaires au splitting DRD : on retrouve bien l'ordre 2 avec une légère perte d'ordre pour les pas de temps trop élevés. Cependant, la chute de vitesse pour les faibles pas de temps est beaucoup plus importante : elle est, en fait, tout à fait comparable à celle des cas de splitting de Lie.

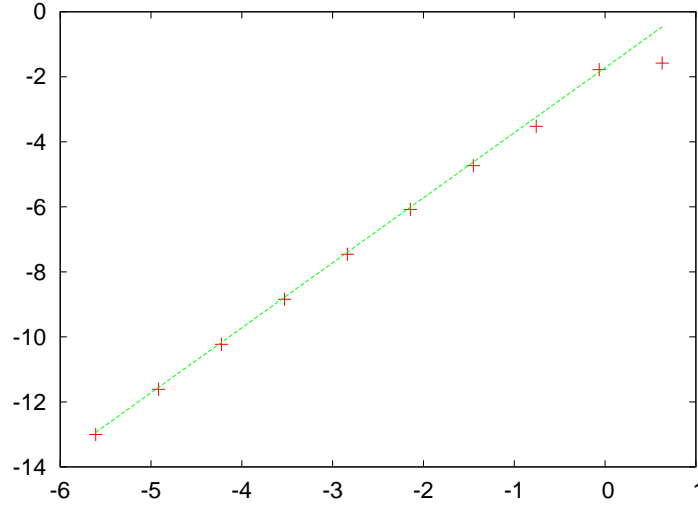


FIG. 54 – Erreur en splitting de Lie RDR dans le cas raide  $k = 10$ ,  $D = 0,1$ . En abscisse :  $\ln(\Delta t)$ . En ordonnée :  $\ln(\|\beta_{qe} - \beta_{sp}\|_2)$  En vert : droite de pente 2

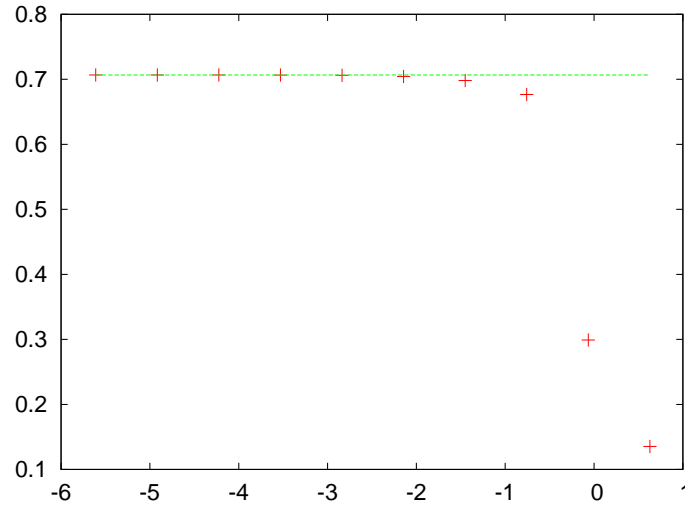


FIG. 55 – Splitting RDR raide : vitesse de l'onde en fonction de  $\ln(\Delta t)$ . En vert : vitesse de l'onde quasi-exacte.

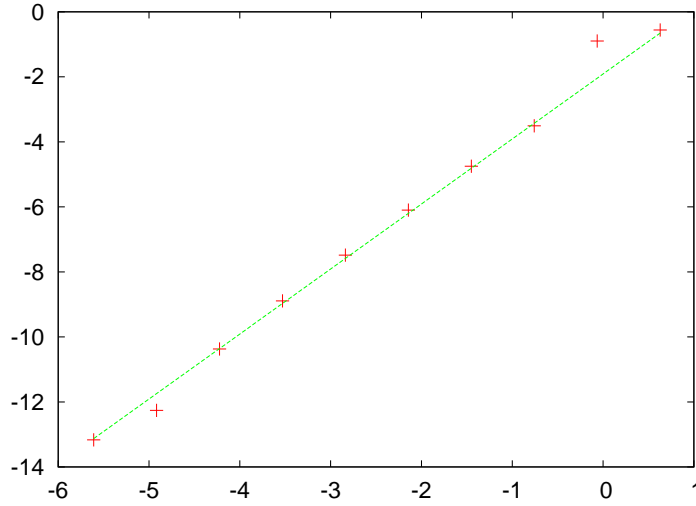


FIG. 56 – Splitting RDR raide :  $\ln(v_{ge} - v_{sp})$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 2

### 8.5.3 Conclusion

L'introduction de raideur dans le système permet de mettre en évidence des pertes d'ordre pour des pas de temps trop faibles. Il reste cependant à en déterminer la cause précise. En effet, il est intéressant de séparer, dans l'erreur  $\|\beta_{ge} - \beta_{sp}\|_2$ , la partie qui est due à l'accroissement de la différence de vitesse de celle qui est due à la perte de la forme de l'onde. Cette synthèse sera faite en regardant les diagrammes de phase, dans la partie IV.

## 9 Modèle de combustion

Dans cette partie, nous étudions en détail la résolution numérique du modèle de la partie 2. Nous expliquons d'abord comment est obtenue la solution numérique présentée en (2.3), puis nous faisons l'étude des résultats donnés par le splitting, en regardant l'erreur que cette méthode donne par rapport à la solution numérique du système complet (la *solution quasi-exacte*). Nous mettrons en évidence des pertes d'ordres qui imposent une certaine prudence lors de l'utilisation de ces méthodes de splitting.

### 9.1 Procédure

Le programme résout le système présenté en 2, avec les paramètres du tableau 2.3, en utilisant à nouveau une méthode des lignes (partie (4)). Comme indiqué dans le tableau, cette discrétisation est effectuée avec 4001 points (le système est unidimensionnel).

Le système d'équation différentielles ordinaires qui en résulte est alors d'abord résolu entièrement, en 2048 pas de temps, grâce à LSODE. La solution obtenue est dite *quasi-exacte*.

On effectue ensuite la résolution de ce système discrétisé par une méthode de splitting RD (Diffusion, puis Réaction) en prenant pour pas de temps :  $\frac{0,04}{1024}, \frac{0,04}{2048}, \dots, \frac{0,04}{65526}$  (pour cette méthode, il faut aller chercher de très petits pas de temps pour voir quelque chose) et par une méthode RDR pour des pas de temps allant de  $\frac{0,04}{512}$  à  $\frac{0,04}{32768}$ .

## 9.2 Etude de l'erreur de splitting - cas RDR

L'étape suivante consiste, comme dans le cas de KPP, à comparer les solutions splittées à la solution quasi-exacte. Regardons d'abord ce qu'il se passe pour le splitting RDR. Si l'on trace (pour  $\Theta$  par exemple, le résultat est exactement le même pour  $\overline{Y_F}$  car  $\Theta + \overline{Y_F} = 1$  et tout à fait similaire pour  $\overline{Y_O}$ ) le logarithme de la différence entre les deux en norme  $L^2$  en fonction du logarithme du pas de temps, on observe que l'on garde un ordre 2 jusqu'au pas de temps  $\frac{0,04}{2048}$ .

Si l'on observe plus en détail l'évolution de la différence pour un pas de temps avant la perte d'ordre (0,04/8192), et un après (0,04/1024), on se rend compte que l'erreur devient extrêmement importante (proche de 1) pour la deuxième catégorie. A ce niveau, on ne peut plus parler d'"ordre" de l'erreur !

L'observation suivante porte sur la vitesse. On observe sur les courbes (vitesse, puis logarithme de la différence de vitesse), que l'on garde cette fois-ci l'ordre 2 presque parfaitement pour tous les pas de temps. La chute de vitesse à  $\Delta t = 0,04/1024$  est tout de même non négligeable : plus de 5% !

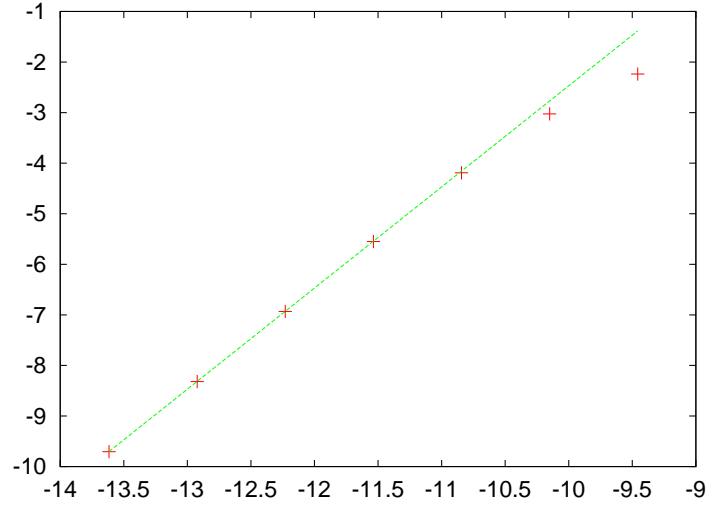


FIG. 57 – Modèle de combustion : erreur en splitting RDR. En abscisse :  $\ln(\Delta t)$ . En ordonnée :  $\ln(\|\Theta_{qe} - \Theta_{sp}\|_2)$

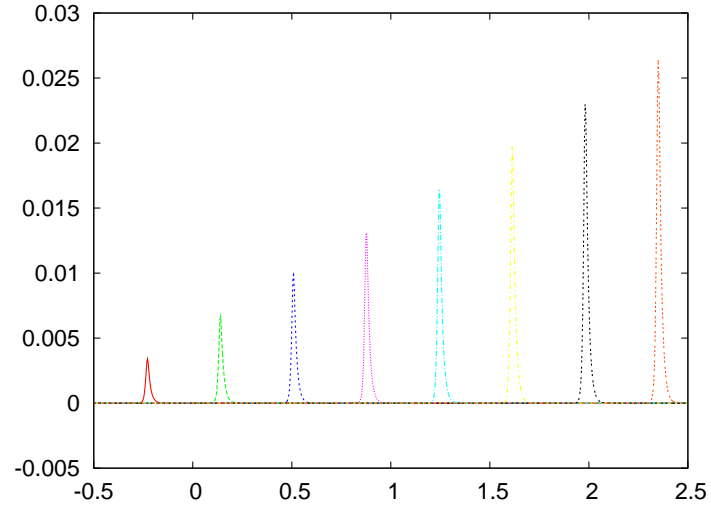


FIG. 58 – Modèle de combustion : Splitting RDR : différence  $\Theta_{qe} - \Theta_{sp}$  en fonction de  $x$ . Pas de temps : 0,04/8192

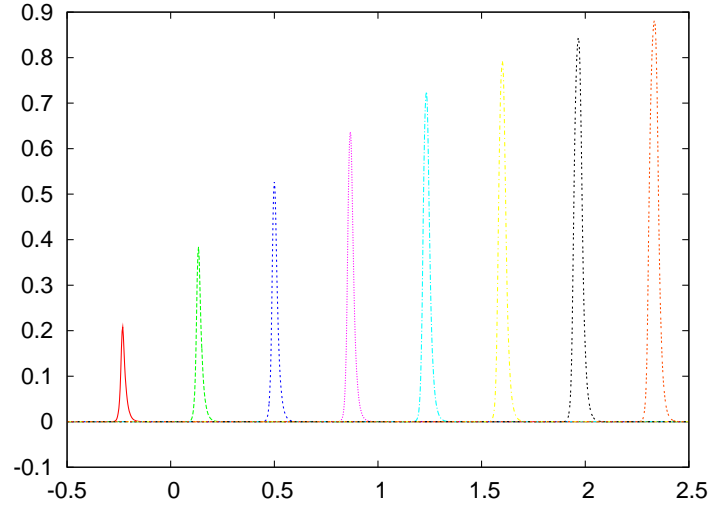


FIG. 59 – Modèle de combustion : Splitting RDR : différence  $\Theta_{qe} - \Theta_{sp}$  en fonction de  $x$ . Pas de temps : 0,04/1024

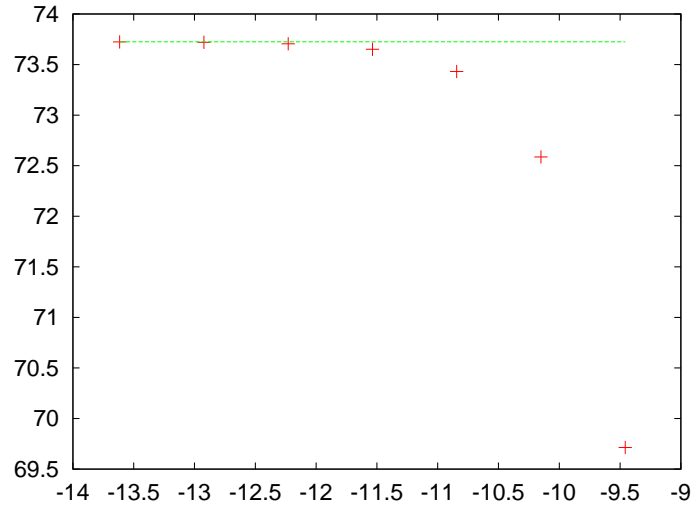


FIG. 60 – Modèle de combustion : Splitting RDR : vitesse de l'onde en fonction de  $\ln(\Delta t)$ . En vert : vitesse de l'onde quasi-exacte.



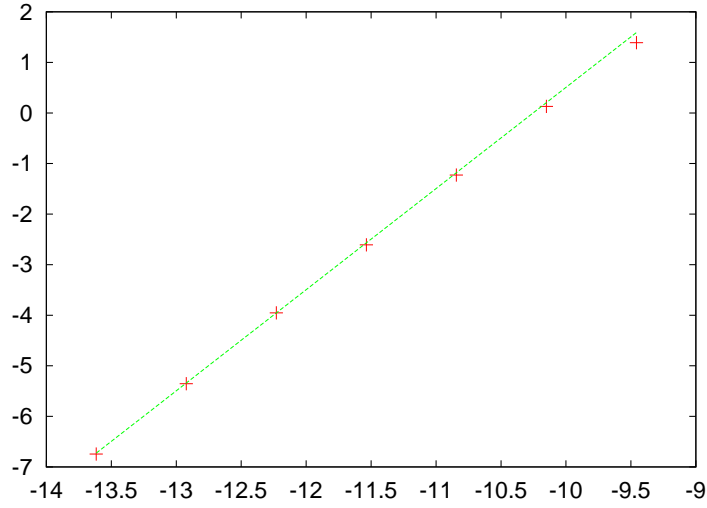


FIG. 61 – Modèle de combustion : Splitting RDR :  $\ln(v_{qe} - v_{sp})$  en fonction de  $\ln(\Delta t)$ . En vert : vitesse de l'onde quasi-exacte.

### 9.3 Etude de l'erreur de splitting - cas RD

Effectuons maintenant la même étude dans le cas du splitting RD. Cette fois-ci, on ne trouve pas l'ordre 1 comme prévu par la théorie mais plutôt un ordre qui s'approche de 1.5 (droite rose) au début pour s'approcher de 2 (droite bleue) ensuite. Pour la vitesse, le phénomène est encore plus marqué.

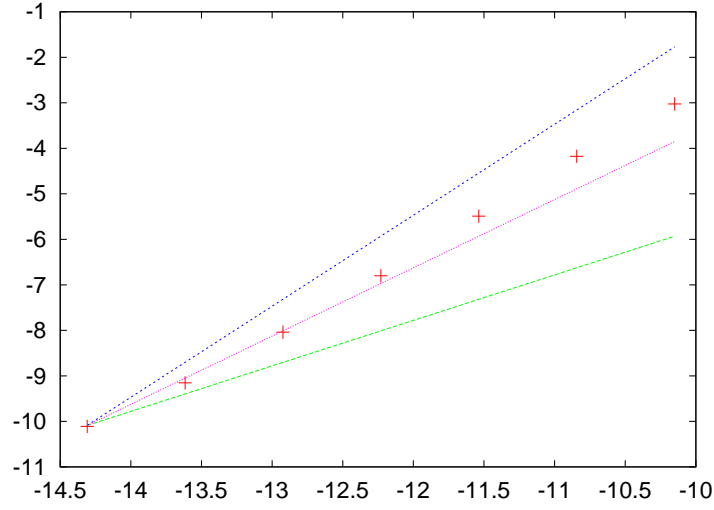


FIG. 62 – Modèle de combustion : erreur en splitting RD. En abscisse :  $\ln(\Delta t)$ . En ordonnée :  $\ln(\|\Theta_{qe} - \Theta_{sp}\|_2)$

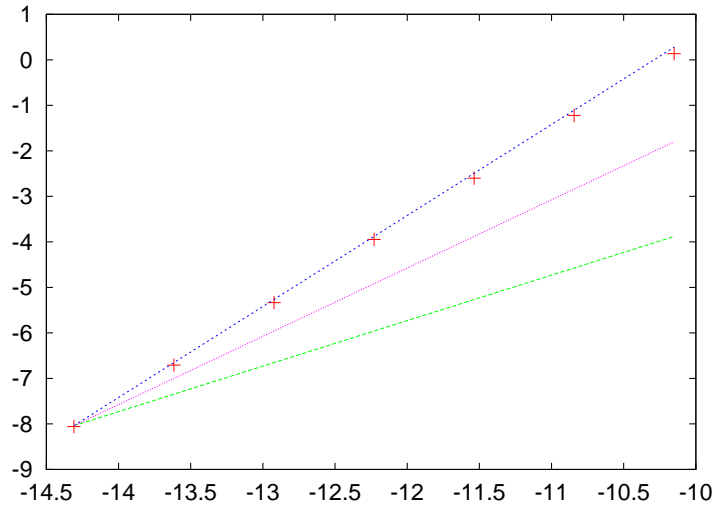


FIG. 63 – Modèle de combustion : Splitting RD :  $\ln(v_{qe} - v_{sp})$  en fonction de  $\ln(\Delta t)$ . En vert : vitesse de l'onde quasi-exacte.

Essayons de regarder ce qu'il se passe plus en détail

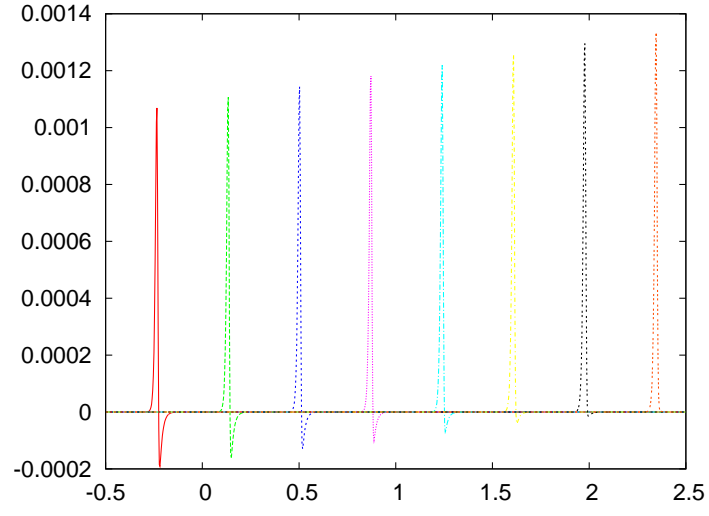


FIG. 64 – Modèle de combustion : Splitting RD : différence  $\Theta_{qe} - \Theta_{sp}$  en fonction de  $x$ . Pas de temps : 0,04/65536

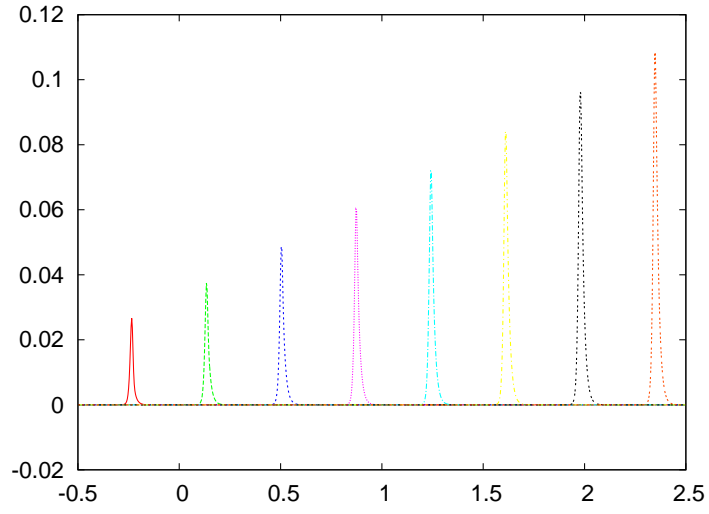


FIG. 65 – Modèle de combustion : Splitting RD : différence  $\Theta_{qe} - \Theta_{sp}$  en fonction de  $x$ . Pas de temps : 0,04/4096

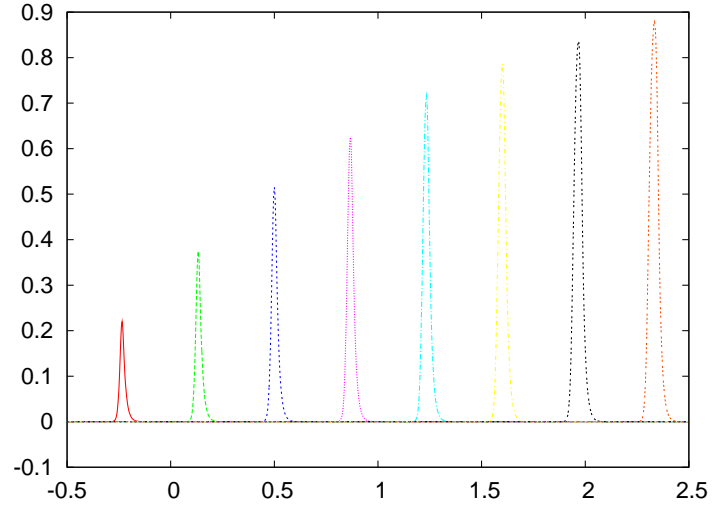


FIG. 66 – Modèle de combustion : Splitting RD : différence  $\Theta_{qe} - \Theta_{sp}$  en fonction de  $x$ . Pas de temps : 0,04/1024

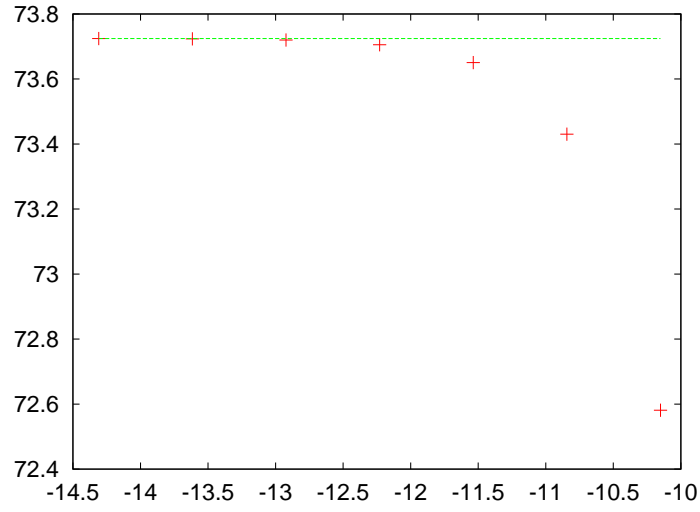


FIG. 67 – Modèle de combustion : Splitting RD : vitesse de l'onde en fonction de  $\ln(\Delta t)$ . En vert : vitesse de l'onde quasi-exacte.

## 9.4 Conclusion

L'onde solution du système réaction-diffusion présenté en 2 est comparable à celle d'un KPP très raide. Ceci explique qu'apparaissent des pertes d'ordre visibles dans le cas RDR. Cependant le fait que l'ordre théorique ne soit pas du tout reproduit dans le cas RD est assez étonnant, et nous n'en avons pas trouvé d'explication.

## Quatrième partie

# Synthèse

## 10 Etude du plan de phase

### 10.1 Synthèse des résultats acquis

Dans la partie précédente, nous avons observé, sur 2 modèles différents, des différences de type  $\|\beta_{qe}(x, \tau) - \beta_{sp}(x, \tau)\|_2$  à  $\tau$  fixé, et leur évolution en fonction du pas de temps de splitting  $\Delta t$  ayant servi au calcul de  $\beta_{sp}$ . La théorie que nous avons présenté dans la partie 6, nous permettait d'obtenir des majorations dans le cas linéaire de type :

$$\|\beta_{qe} - \beta_{sp}\|_2 \leq C(\Delta t)^\delta$$

où  $\delta$  est l'ordre de la méthode :  $\delta = 1$  pour la méthode de Lie,  $\delta = 2$  pour la méthode de Strang. Ainsi :

$$\ln(\|\beta_{qe} - \beta_{sp}\|_2) \leq \delta \ln(\Delta t) + \ln(C)$$

Nous avons très bien retrouvé ce résultat dans le cas de KPP non raide, puisque le tracé de  $\ln(\|\beta_{qe} - \beta_{sp}\|_2)$  en fonction de  $\ln(\Delta t)$  nous donnait parfaitement une droite ayant la pente voulue (il y avait donc même égalité dans l'expression ci-dessus, ce qui montrait bien que les termes d'ordre supérieur étaient négligeables). En revanche, lorsque la solution présentait de forts gradients spatiaux, le résultat était mis en défaut pour des pas de temps trop importants : il apparaissait des "pertes d'ordre". La partie 6.4 propose, dans le cas linéaire et pour le splitting de Lie, une explication de la perte d'un-demi ordre qu'on retrouve à peu près pour DRD, sur la figure 48. En revanche nous ne retrouvons pas ce résultat pour certains cas inexplicables, notamment la combustion RD (des erreurs de code ou de manipulation sont envisageables). En tous les cas, il apparait nettement que lorsque l'onde solution présente de forts gradients spatiaux, les résultats théoriques sont pris en défaut.

Par ailleurs, nous avons observé que, pour des pas de temps trop importants, la vitesse chutait de manière exponentielle. Plus précisément, la forme des courbes tracées suggère une expression du type :

$$v_{sp} = v_{qe} + \alpha_v(\Delta t)^\delta \quad (53)$$

puisque le tracé de  $\ln(v_{qe} - v_{sp})$  en fonction de  $\ln(\Delta t)$  donne une droite de pente  $\delta$ . Notons qu'on a presque toujours  $\alpha_v < 0$  (dans tous les cas que l'on a observés, à part KPP DR non raide, la vitesse de l'onde splittée était inférieure à la vitesse de l'onde quasi-exacte). Il est clair que ce défaut de vitesse intervient directement dans la composition de l'erreur : en effet, si l'on suppose par exemple que les deux ondes ont exactement le même profil  $\beta^0$  :

$$\beta^0(x - (v_{qe} + \alpha_v(\Delta t)^\delta)t) - \beta^0(x - v_{qe}t) = -\alpha_v t \frac{d\beta^0}{dx}(\Delta t)^\delta$$

et donc l'erreur  $L^2$  sur  $\beta$  a la même structure que l'erreur sur la vitesse (ordre  $\delta$ ).

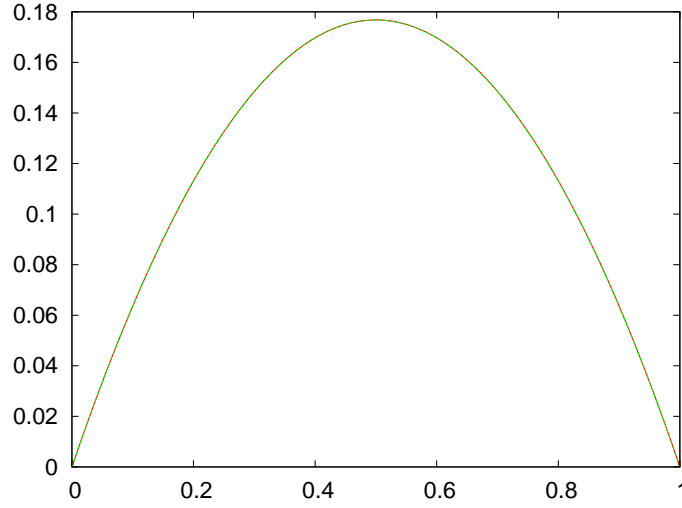


FIG. 68 – KPP non raide : diagramme de phase quasi-exact (en rouge) superposé à son expression analytique (en vert)

Bien entendu, le fait que le profil soit différent a aussi une influence sur la composition de l'erreur. Pour évaluer la variation du profil de l'onde uniquement, nous avons besoin de nous affranchir de la composante spatiale. L'étude dans le plan de phase  $(\beta, \frac{d\beta}{dx})$  est alors tout à fait adaptée à ce problème.

## 10.2 Etude du plan de phase de KPP non raide

Pour commencer, nous présentons à la figure 68 le diagramme de phase de l'onde quasi-exacte de KPP (cf partie 1.4, mais ici on représente en fait toujours le symétrisé du diagramme par rapport à l'axe des abscisses), en superposant la courbe calculée à son expression analytique : on voit que les courbes se superposent parfaitement.

### 10.2.1 Méthode RDR

Nous cherchons à tracer la différence  $\beta'_{sp} - \beta'_{qe}$  en fonction de  $\beta$  (le ' désigne la dérivation spatiale qui se fait pour un instant  $\tau$  fixé). La section 7.6 explique comment aboutir à ce résultat. Superposons les courbes obtenues dans le cas RDR pour des trois pas de temps de splitting différents : le résultat est présenté à la figure 69

Il apparaît de manière frappante que les courbes sont dilatées d'un facteur qui n'est autre que le rapport des pas de temps au carré. Ceci est confirmé par la figure 70 qui rapporte les ordonnées à ce facteur. Plus précisément, on peut chercher à tracer le logarithme du maximum de  $\beta'_{sp} - \beta'_{qe}$  en fonction du logarithme de  $\Delta t$ . On obtient une droite de pente 2.



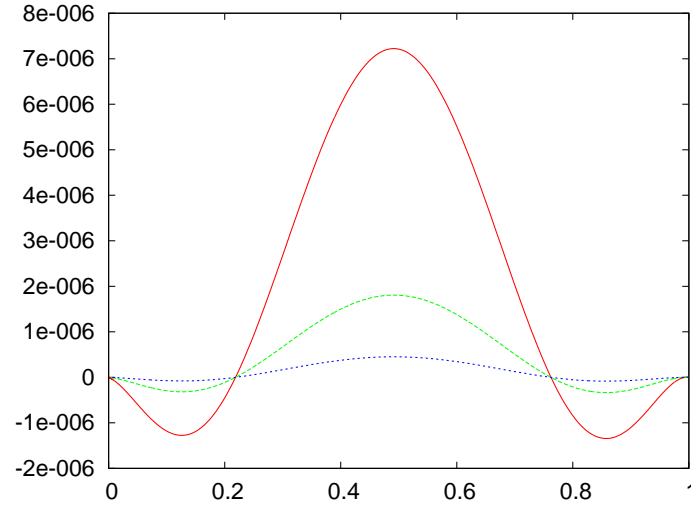


FIG. 69 – Différence des diagrammes de phase :  $\beta'_{sp} - \beta'_{qe}$  en fonction de  $\beta$ . Pas de temps  $\frac{30}{512}$  (rouge),  $\frac{30}{1024}$  (vert),  $\frac{30}{2048}$  (bleu)

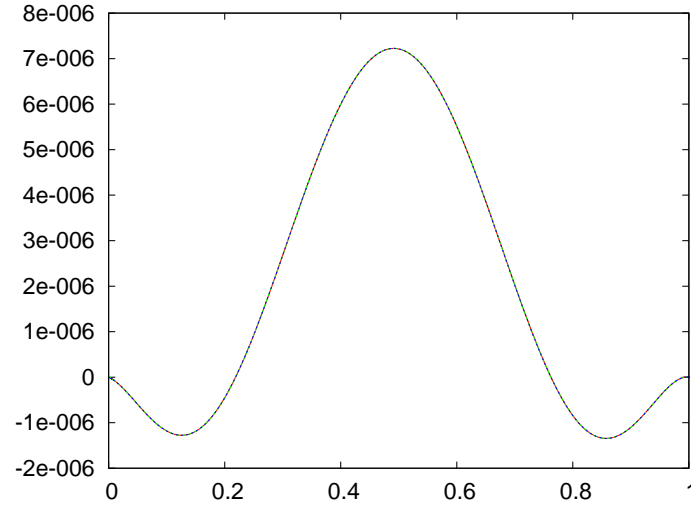


FIG. 70 – Tracé de  $\beta'_{sp} - \beta'_{qe}$  pour RDR 512 (rouge),  $4(\beta'_{sp} - \beta'_{qe})$  pour RDR 1024 (vert), et  $16(\beta'_{sp} - \beta'_{qe})$  pour RDR 1024 (vert) en fonction de  $\beta$ .

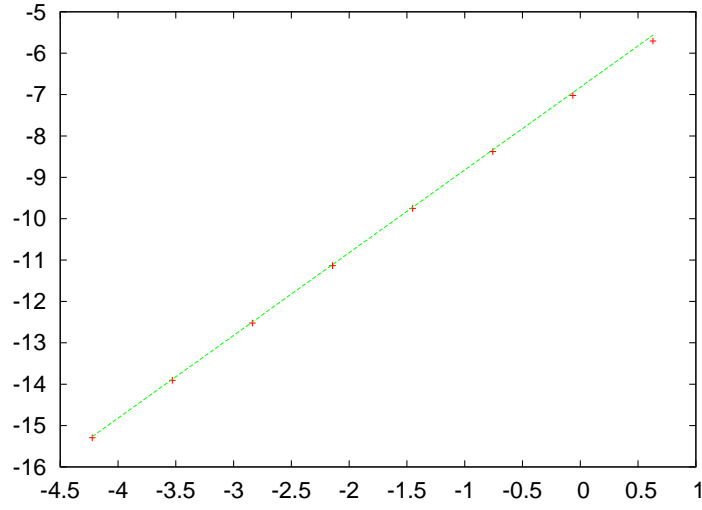


FIG. 71 – Diagramme de phase RDR non raide :  $\ln(\|\beta'_{sp} - \beta'_{qe}\|_2)$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 2.

### 10.2.2 Autres méthodes de splitting

Les autres méthodes de splitting donnent des résultats tout à fait similaires. Pour RD et DR, le facteur d'ajustement est 1. Nous donnons la fonction de forme et le tracé en logarithme pour chacun des cas DRD, RD et DR.

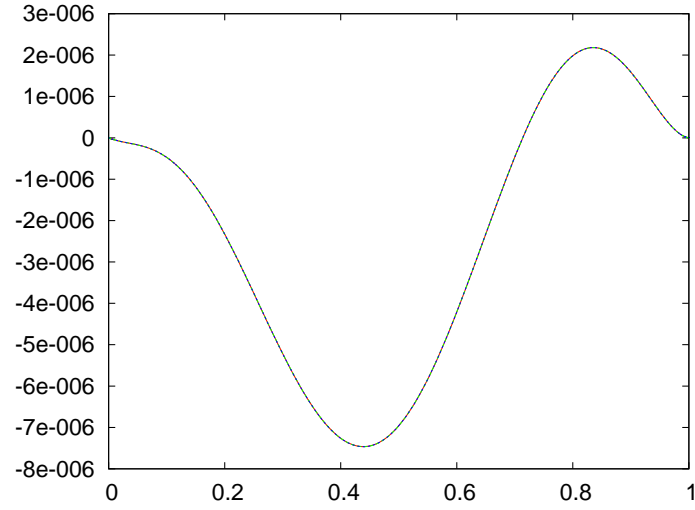


FIG. 72 – Fonction de forme pour KPP - cas DRD non raide

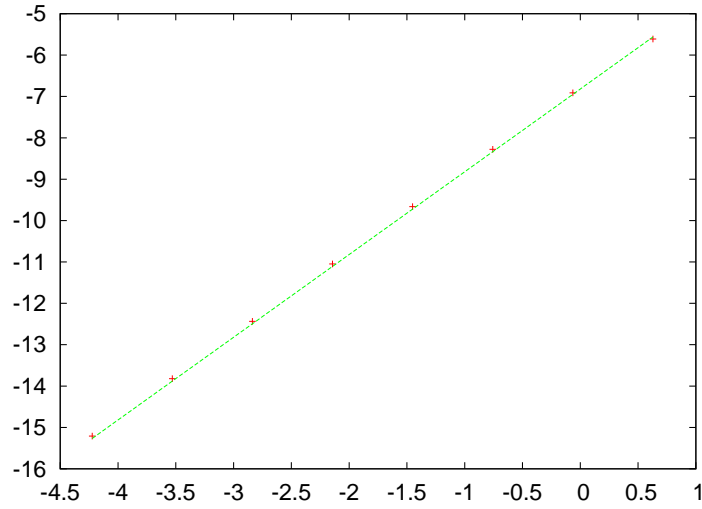


FIG. 73 – Diagramme de phase DRD non raide :  $\ln(\|\beta'_{sp} - \beta'_{qe}\|_2)$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 2.

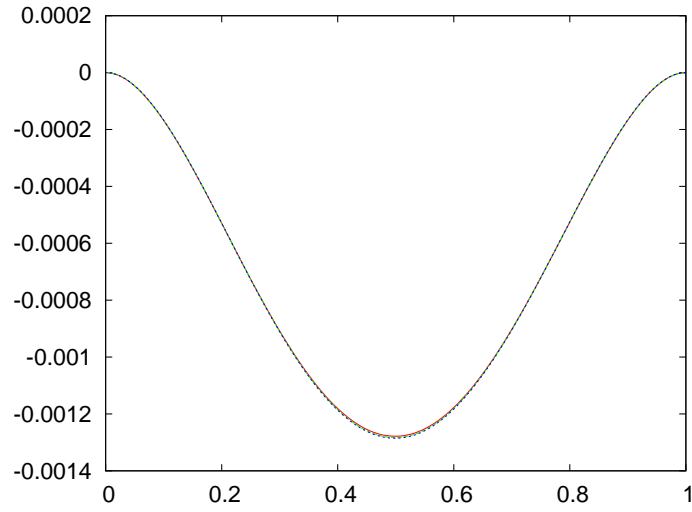


FIG. 74 – Fonction de forme pour KPP - cas RD non raide

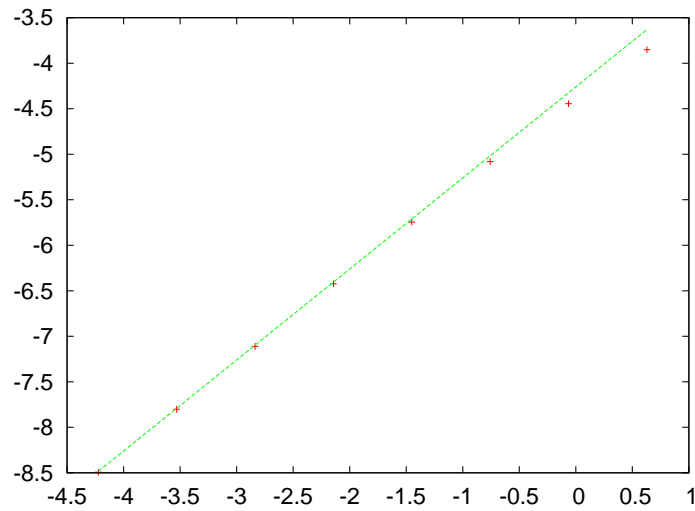


FIG. 75 – Diagramme de phase RD non raide :  $\ln(\|\beta'_{sp} - \beta'_{qe}\|_2)$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 1.

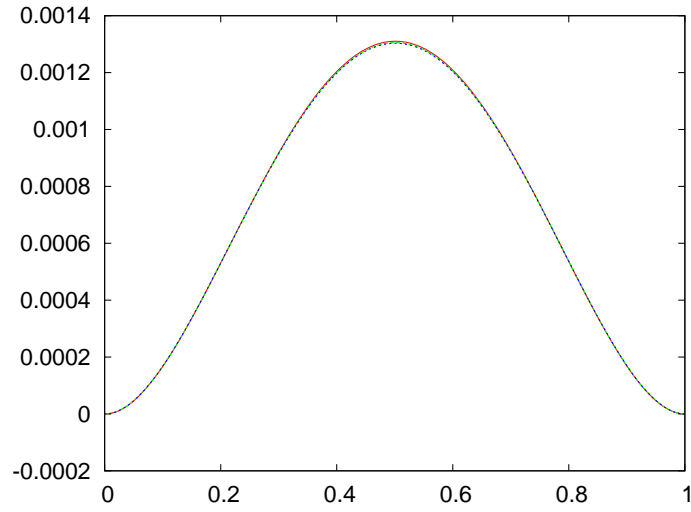


FIG. 76 – Fonction de forme pour KPP - cas DR non raide

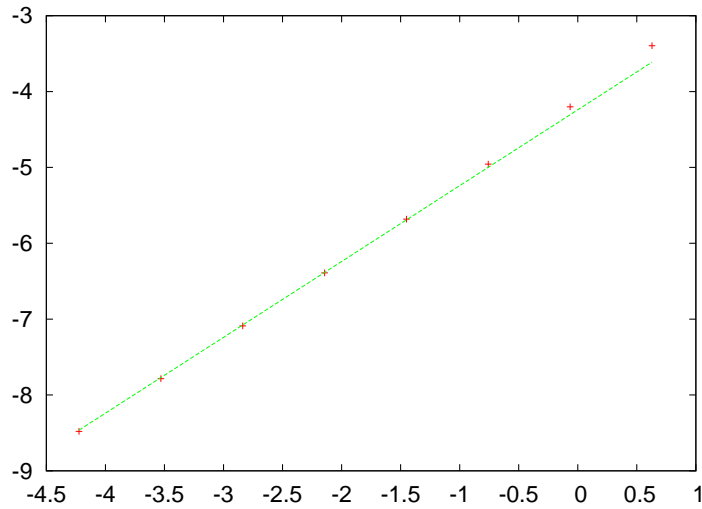


FIG. 77 – Diagramme de phase DR non raide :  $\ln(\|\beta'_{sp} - \beta'_{qe}\|_2)$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 1.

### 10.3 Etude du plan de phase de KPP raide

Effectuons exactement la même étude que précédemment pour KPP raide. Cette fois-ci les courbes ne se superposent pas parfaitement, même si on retrouve l'allure de la fonction de forme associée à chacun des schémas de splitting. L'erreur provient probablement de l'interpolation linéaire, qui n'est pas assez précise. Une interpolation par splines cubiques donnerait probablement des résultats plus satisfaisants. Nous avons essayé de l'implémenter mais le programme produisait des erreurs que nous n'avons pas eu le temps de corriger.

Dans ce cas raide, il apparaît assez nettement une perte d'ordre pour les pas de temps trop élevés. On voit même apparaître une pente 1,5 pour les cas RDR et DRD, ce qui est assez intéressant.

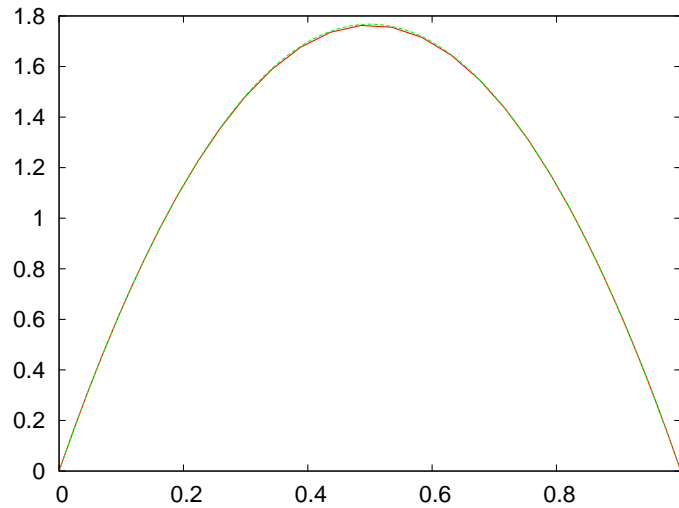


FIG. 78 – KPP raide : diagramme de phase quasi-exact (en rouge) superposé à son expression analytique (en vert)

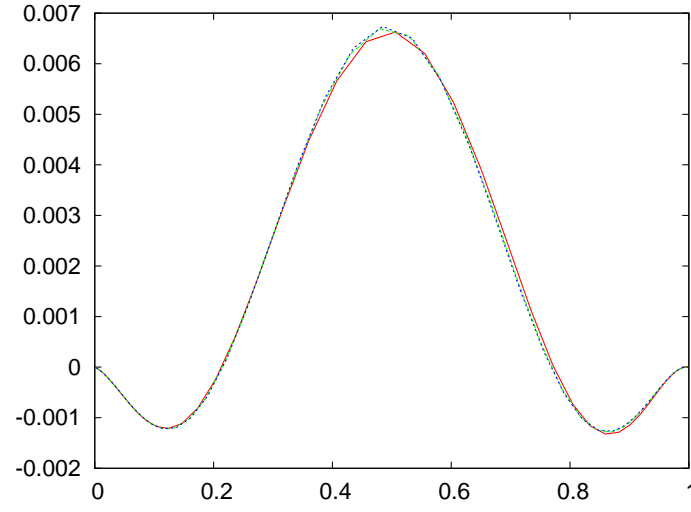


FIG. 79 – KPP raide : tracé de  $\beta'_{sp} - \beta'_{qe}$  pour RDR 512 (rouge),  $4(\beta'_{sp} - \beta'_{qe})$  pour RDR 1024 (vert), et  $16(\beta'_{sp} - \beta'_{qe})$  pour RDR 1024 bleu) en fonction de  $\beta$ .

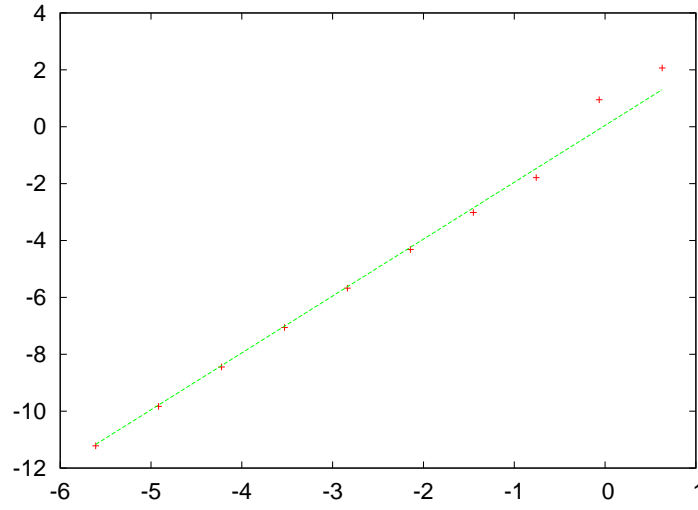


FIG. 80 – Diagramme de phase RDR raide :  $\ln(\|\beta'_{sp} - \beta'_{qe}\|_2)$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 2. En bleu : droite de pente 1,5

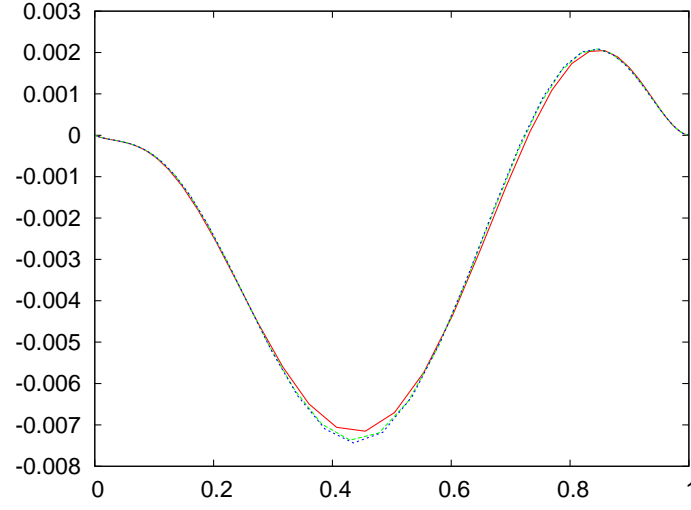


FIG. 81 – KPP raide : tracé de  $\beta'_{sp} - \beta'_{qe}$  pour DRD 512 (rouge),  $4(\beta'_{sp} - \beta'_{qe})$  pour DRD 1024 (vert), et  $16(\beta'_{sp} - \beta'_{qe})$  pour DRD 1024 (bleu) en fonction de  $\beta$ .

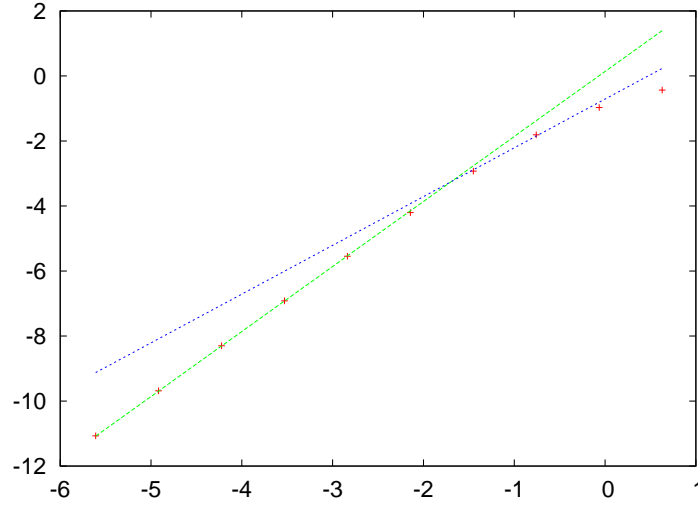


FIG. 82 – Diagramme de phase DRD raide :  $\ln(\|\beta'_{sp} - \beta'_{qe}\|_2)$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 2. En bleu : droite de pente 1,5



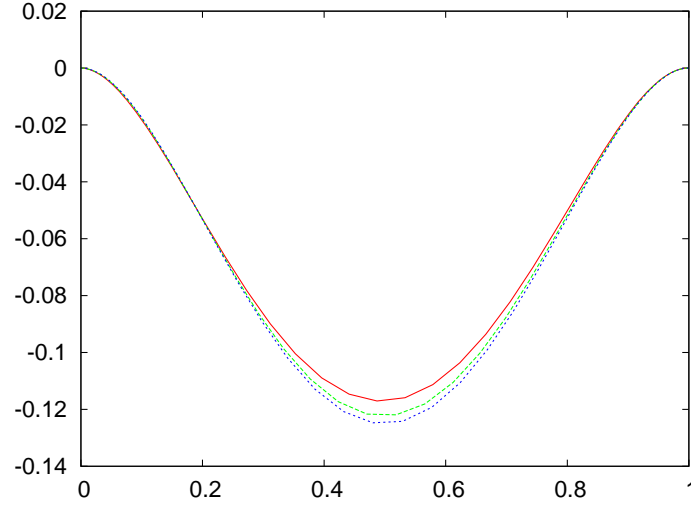


FIG. 83 – KPP raide : tracé de  $\beta'_{sp} - \beta'_{qe}$  pour RD 512 (rouge),  $2(\beta'_{sp} - \beta'_{qe})$  pour RD 1024 (vert), et  $4(\beta'_{sp} - \beta'_{qe})$  pour RD 1024 (bleu) en fonction de  $\beta$ .

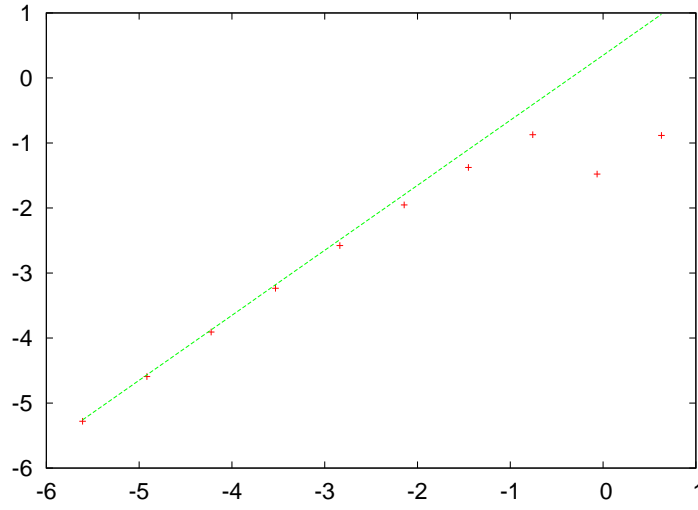


FIG. 84 – Diagramme de phase RD raide :  $\ln(\|\beta'_{sp} - \beta'_{qe}\|_2)$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 1

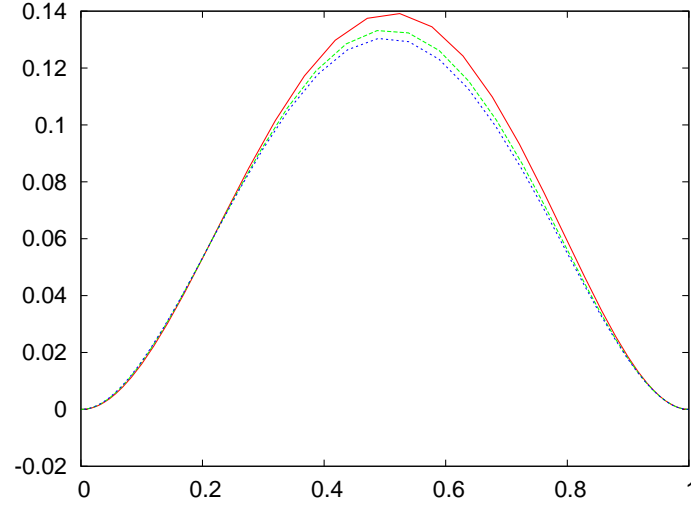


FIG. 85 – KPP raide : tracé de  $\beta'_{sp} - \beta'_{qe}$  pour DR 512 (rouge),  $2(\beta'_{sp} - \beta'_{qe})$  pour DR 1024 (vert), et  $4(\beta'_{sp} - \beta'_{qe})$  pour DR 1024 (bleu) en fonction de  $\beta$ .

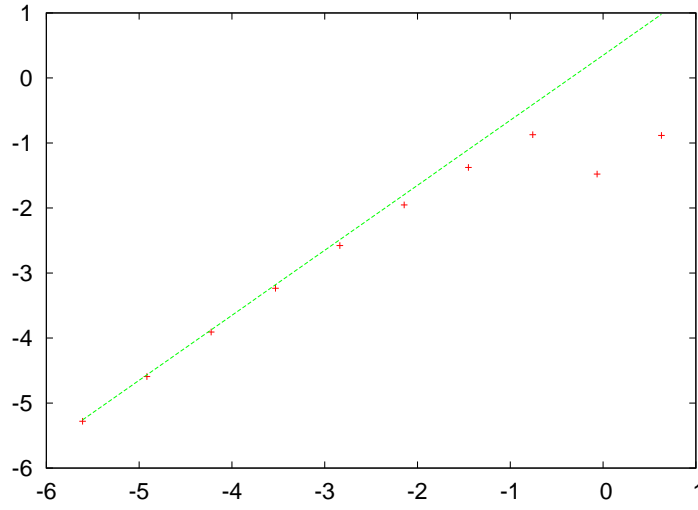


FIG. 86 – Diagramme de phase DR raide :  $\ln(\|\beta'_{sp} - \beta'_{qe}\|_2)$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 1

## 10.4 Etude du plan de phase du modèle de combustion

Regardons d'abord l'allure du diagramme de phase de notre modèle de combustion. On voit qu'il ressemble beaucoup à celui de KPP, mais que les pentes prennent des valeurs beaucoup plus importantes (30 fois plus importantes que KPP raide !)

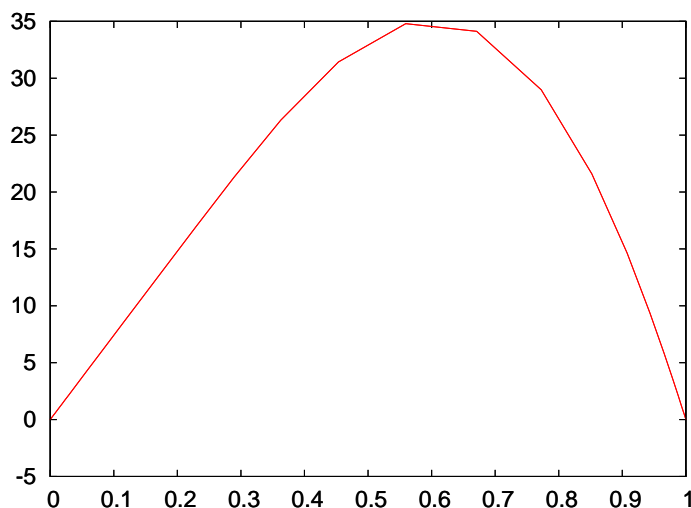


FIG. 87 – Diagramme de phase de l'onde quasi-exacte solution du modèle de combustion

Essayons maintenant de retrouver une fonction de forme, comme dans le cas de KPP. Dans le cas RDR, on arrive presque à faire se superposer les courbes pour les pas de temps  $\frac{0,04}{32768}$  et  $\frac{0,04}{16384}$  (les tracés sont fait pour la température réduite  $\Theta$  mais ils seraient les mêmes pour les deux autres variables). Cela est confirmé par la droite de pente 2 qui passe par les deux points correspondants (les plus à gauche) sur la courbe des logarithmes. Dans le cas RD, on obtient bien une droite de pente proche de 1, ce que confirme l'observation des courbes qui se superposent à peu près pour les pas de temps  $\frac{0,04}{65536}$ ,  $\frac{0,04}{32768}$  et  $\frac{0,04}{16384}$  (trois points de gauche sur le graphe logarithmique).

Dans tous les cas apparaît bien une fonction de forme. Notons qu'ici l'interpolation linéaire n'est que très peu précise : cela est dû au fait que, d'une part, nous n'avons que très peu de points dans la zone du front d'onde, et d'autre part, nous n'avons pas l'expression analytique du diagramme de phase. Nous pourrions aussi améliorer en faisant une interpolation par splines.

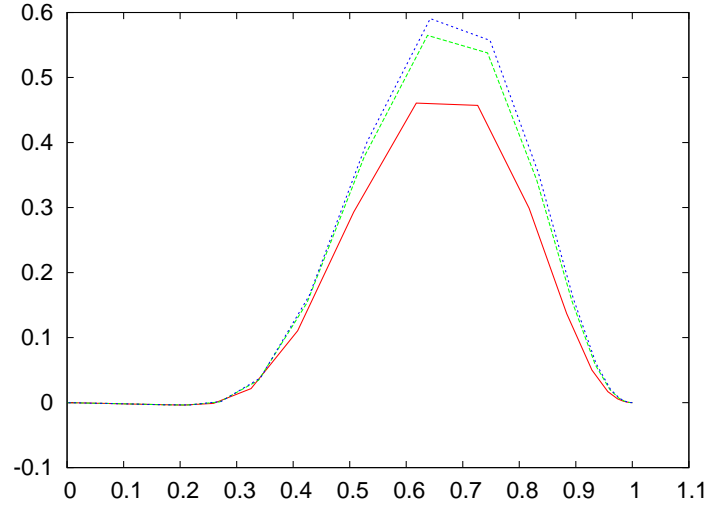


FIG. 88 – Combustion RDR : tracé de  $\Theta'_{sp} - \Theta'_{qe}$  pour  $\Delta t = \frac{0.04}{8192}$  (rouge),  $4(\Theta'_{sp} - \Theta'_{qe})$  pour  $\Delta t = \frac{0.04}{16384}$  (vert), et  $16(\Theta'_{sp} - \Theta'_{qe})$  pour  $\Delta t = \frac{0.04}{32768}$  (bleu) en fonction de  $\beta$ .

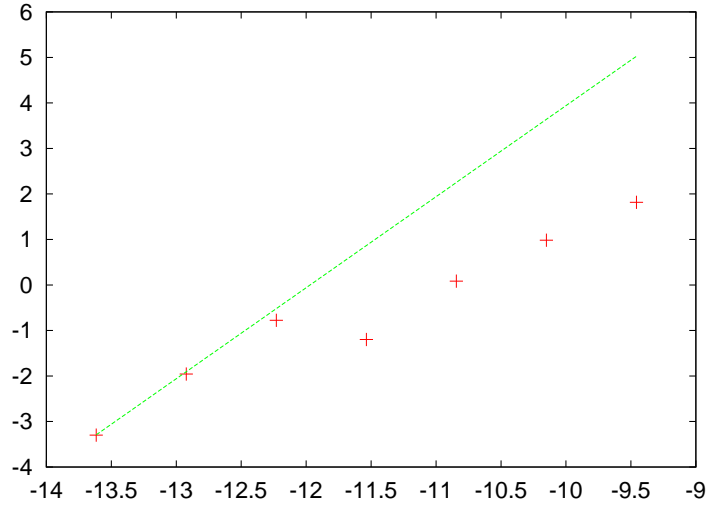


FIG. 89 – Combustion RDR : diagramme de phase :  $\ln(\|\Theta'_{sp} - \Theta'_{qe}\|)$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 2

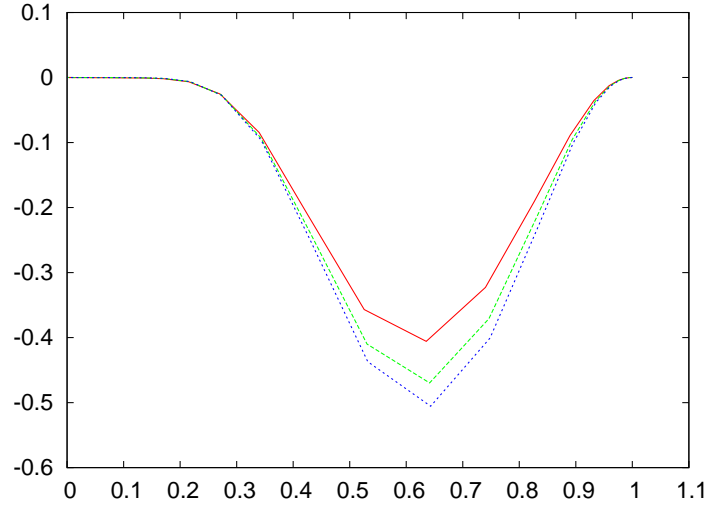


FIG. 90 – Combustion RD : tracé de  $\Theta'_{sp} - \Theta'_{qe}$  pour  $\Delta t = \frac{0.04}{16384}$  (rouge),  $2(\Theta'_{sp} - \Theta'_{qe})$  pour  $\Delta t = \frac{0.04}{32768}$  (vert), et  $4(\Theta'_{sp} - \Theta'_{qe})$  pour  $\Delta t = \frac{0.04}{65536}$  (bleu) en fonction de  $\beta$ .

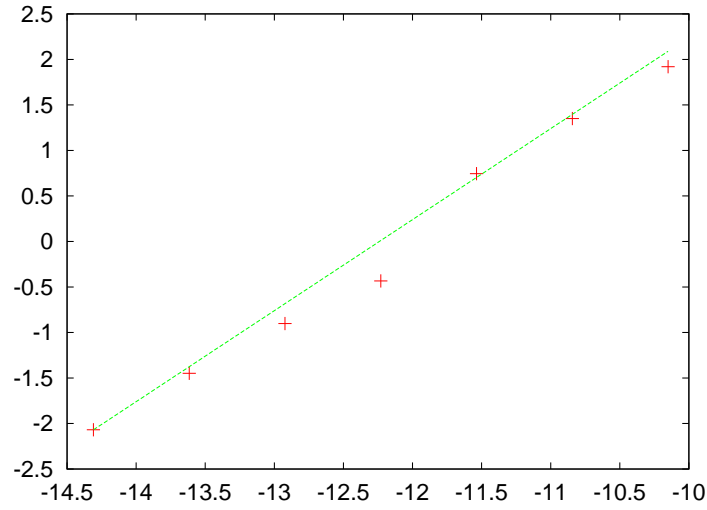


FIG. 91 – Combustion RD : diagramme de phase :  $\ln(\|\Theta'_{sp} - \Theta'_{qe}\|)$  en fonction de  $\ln(\Delta t)$ . En vert : droite de pente 1

## 10.5 Conclusion

Les études dans le plan de phase nous montrent que la dérivée spatiale de l'onde splittée vérifie une équation de type

$$\beta'_{sp} = \beta'_{qe} + \alpha_p \Psi(\beta)(\Delta t)^\delta$$

Où  $\Psi(\beta)$  est une fonction de forme normalisée à 1, spécifique au schéma de splitting, et  $\delta$  est l'ordre du schéma.

Nous avons donc deux constantes intéressantes intervenant dans la structure de l'erreur :  $\alpha_p$  et  $\alpha_v$  dans l'équation (53) :

$$v_{sp} = v_{qe} + \alpha_v (\Delta t)^\delta$$

La comparaison de ces deux constantes nous donnera donc des informations intéressantes sur la manière dont se construit l'erreur. Nous tirons des équations ci-dessus que :

$$\begin{aligned} \ln(\|\beta'_{sp} - \beta'_{qe}\|_2) &= \ln(\alpha_p) + \delta \ln(\Delta t) \quad \text{car} \quad \|\Psi\|_2 = 1 \\ \ln(v_{sp} - v_{qe}) &= \ln(\alpha_v) + \delta \ln(\Delta t) \end{aligned}$$

ce qui nous donne un moyen simple de calculer les coefficients  $\alpha_v$  et  $\alpha_p$  à partir des courbes déjà tracées.

À partir des équations des droites de régression on obtient les résultats résumés dans le tableau 3. Il apparaît que les ordres de grandeur pour  $\alpha_p$  et  $\alpha_v$  sont tout à fait comparables dans le cas de KPP  $k = 1$ ,  $D = 1$ , ce qui est loin du résultat attendu.

À ce stade, nous ne pouvons pas conclure sur la manière dont se structure l'erreur. De nombreuses choses restent encore à être examinées, notamment le problème de la mise à l'échelle lorsqu'on passe de KPP non raide à KPP raide, que nous n'avons pas résolu.

Modèle	Schéma	$\ln(\alpha_v)$	$\ln(\alpha_p)$	$\alpha_v \cdot 10^3$	$\alpha_p \cdot 10^3$
KPP non raide	RD	-5,12	-4,26	5,97	14,12
	DR	-5,15	-4,24	5,80	14,40
	RDR	-6,48	-6,82	1,53	1,09
	DRD	-6,08	-6,82	2,28	1,09
KPP raide	RD	-5,1	0,39	6,10	1476
	DR	-5,2	0,37	5,51	1447
	RDR	-1,9	0,05	150	1051
	DRD	-1,9	0,14	150	1150

TAB. 3 – Estimation des constantes  $\alpha_v$  et  $\alpha_p$

## Conclusion

En conclusion, les nombreux calculs que nous avons menés ont ouvert beaucoup plus de questions qu'ils n'en ont résolues. Dans le cas du système KPP non raide, les résultats théoriques du cas linéaire se retrouvent très bien. L'introduction de raideur (KPP raide et combustion) a pour effet de perturber nettement les résultats du cas linéaire, et, bien que l'on puisse proposer une théorie de la perte d'ordre, les résultats que l'on observe ne lui sont pas très conformes. Ici, de nombreuses questions se posent pour savoir à quoi sont dus ces "pertes" d'ordres (qui, tels qu'on les observe, sont parfois des "gains d'ordre"!). Il est possible qu'interviennent effets non linéaires complexes, ou qu'il y ait un cumul d'erreur numériques non maîtrisé lors du passage du local au global. Des erreurs au niveau du code sont possibles, mais celui donne pourtant des résultats acceptables dans le cas de KPP raide, et de plus, les erreurs de "manipulation" au niveau des fichiers sont peu probables car nous avons vérifié de nombreuses fois les résultats.

En ce qui concerne la question de la structure de l'erreur, nous avons commencé à comprendre et à vérifier certains résultats : rôles spécifiques et structure de la vitesse et du diagramme de phase. Mais la question de la prépondérance de l'un par rapport à l'autre n'est pas résolue.

En conclusion, ce projet peut servir de base à des recherches plus approfondies sur les questions de pertes d'ordre et de structure de l'erreur engendrée par la méthode de séparation d'opérateurs appliquée aux systèmes de réaction-diffusion dont la condition initiale présente de forts gradients spatiaux. En ce qui nous concerne, il nous a donné la chance de pouvoir prendre contact avec le milieu de la recherche en mathématiques appliquées à l'interface, et de travailler sur des sujets actuels, en lien étroit avec notre encadrant et ses collègues, ce qui a été très enrichissant.

## Références

- [1] S. DESCOMBES, M. MASSOT, Operator splitting for nonlinear reaction-diffusion systems with an entropic structure : singular perturbation and order reduction, *Numerische Mathematik*, Vol 97 :4, pp 667-698, 2004
- [2] T. GALLAY, Ondes progressives dans les systèmes de réaction-diffusion, Cours à l'école d'été 2005 de l'Université Joseph Fourier.
- [3] P. GRAY, S.K. SCOTT, Chemical Oscillations and Instabilities : nonlinear chemical kinetics, Oxford University Press, 1990. Chapter 11 : Travelling Waves. pp. 292-312
- [4] E. HAIRER, Cours d'Analyse Numérique à l'Université de Genève, Chapitre III : Equations Différentielles Ordinaires, 2004
- [5] A. N. KOLMOGOROV, I. G. PETROVSKII, N. S. PISKUNOV, A Study of the diffusion equation with increase in the amount of substance, and its application to a biological problem, *Bjul. Moskovskogo Gos. Univ*, Série Internationale, Section A. 1, 1937, pp. 1-26
- [6] T. POINSOT, D. VEYNANTE, Theoretical and Numerical Combustion, Edwards, 2005.
- [7] B. SPORTISSE, An analysis of operator splitting techniques in the stiff case. *J.Comp. Phys.*, 141 :140-168, 2000.

- [8] B. SPORTISSE, V. MALLET, Calcul scientifique pour l'environnement, cours ENSTA, 2005.
- [9] A. VOLPERT, V. VOLPERT, V. VOLPERT, Travelling Wave Solutions of Parabolic Systems, American Mathematical Society, 1994

## **Annexes**

En annexe sont attachés deux rapports que nous avons écrits dans le cadre d'une étude préliminaire au projet. L'un porte sur la théorie des équations différentielles ordinaires, et l'autre les méthodes numériques de résolution de ces équations.



# Théorie des équations différentielles ordinaires

Adrien Auclert

23 décembre 2005

## 1 Définitions et généralités

### 1.1 Equation différentielle résolue

Le premier problème qui se pose lors de l'étude des équations différentielles est leur définition précise. Nous faisons le choix de travailler avec des équations *résolues* et des fonctions qui prennent leurs valeurs dans  $\mathbb{R}^n$  (mais on pourrait prendre plus généralement un espace de Banach).

**Définition 1** Soit  $(n, r) \in \mathbb{N}^{*2}$ ,  $\Omega$  un ouvert de  $\mathbb{R} \times (\mathbb{R}^n)^r$  et une application

$$f : \Omega \subset \mathbb{R} \times (\mathbb{R}^n)^r \rightarrow \mathbb{R}^n$$

On appelle solution de l'équation différentielle (explicite) d'ordre  $r$  :

$$\mathcal{E} : y^{(r)} = f(t, y, y', \dots, y^{(r-1)})$$

toute application  $\varphi : I \rightarrow \mathbb{R}^n$  (où  $I$  est un intervalle de  $\mathbb{R}$ ),  $r$  fois dérivable et vérifiant

$$\forall t \in I, (t, \varphi(t), \varphi'(t), \dots, \varphi^{(r-1)}(t)) \in \Omega \quad (1)$$

$$\forall t \in I, f(t, \varphi(t), \varphi'(t), \dots, \varphi^{(r-1)}(t)) = \varphi^{(r)}(t) \quad (2)$$

### 1.2 Problème de Cauchy

L'introduction d'une condition supplémentaire sur une solution va permettre d'obtenir, dans certains cas, son unicité.

**Définition 2** On dit qu'une solution  $\varphi$  de  $\mathcal{E}$  vérifie la condition initiale (ou de Cauchy)  $(t_0, x_0, x_1, \dots, x_{r-1}) \in \Omega$  si

$$t_0 \in I, \quad \varphi(t_0) = x_0, \quad \dots, \quad \varphi^{(r-1)}(t_0) = x_{r-1} \quad (3)$$

### 1.3 Passage à une équation d'ordre 1

On peut ramener toute équation différentielle à une équation d'ordre 1. Pour cela, on définit

$$\Phi : I \rightarrow (\mathbb{R}^n)^r \quad t \mapsto (\varphi(t), \varphi'(t), \dots, \varphi^{(r-1)}(t))$$

$$F : \Omega \rightarrow (\mathbb{R}^n)^r \quad (t, y) = (t, x_0, x_1, \dots, x_{r-1}) \mapsto (x_1, \dots, x_{r-1}, f(t, y))$$

Il est alors équivalent de dire que  $\varphi$  est solution de  $\mathcal{E}$  ou que  $\Phi$  est solution de l'équation d'ordre 1,  $\mathcal{E}_1 : \Phi' = F(t, \Phi)$

Cette opération est compatible avec les conditions de Cauchy :  $\varphi$  est solution de  $\mathcal{E}$  avec la condition de Cauchy  $(t_0, x_0, x_1, \dots, x_{r-1})$  si, et seulement si la solution correspondante  $\Phi$  est solution de  $\mathcal{E}_1$  avec la condition de Cauchy  $(t_0, y_0)$ , où  $y_0 = (x_0, x_1, \dots, x_{r-1})$

Grâce à ce résultat, on ne s'intéresse plus qu'au cas des équations différentielles de degré 1 ( $r = 1$ )

## 2 Existence et unicité des solutions

### 2.1 Forme intégrale

Une formulation équivalente à un problème de Cauchy va permettre de démontrer plus facilement l'existence et l'unicité des solutions.

Soit  $f : \Omega \rightarrow \mathbb{R}^n$  (avec  $\Omega$  ouvert de  $\mathbb{R} \times \mathbb{R}^n$ ) une fonction *continue*. Les solutions de

$$y' = f(t, y) \tag{4}$$

avec la condition de Cauchy

$$\varphi(t_0) = x_0 \tag{5}$$

sont alors les fonctions  $\varphi \in \mathcal{C}^1$  vérifiant :

$$\forall t \in I, \quad (t, \varphi(t)) \in \Omega \quad \text{et} \quad \varphi(t) = x_0 + \int_{t_0}^t f(s, \varphi(s)) \, ds \tag{6}$$

### 2.2 Théorème de Cauchy-Lipschitz

L'existence et l'unicité des solutions ne s'obtient que pour un certain type de fonction que nous définissons maintenant :

**Définition 3** On dit qu'une application  $f$  d'un ouvert  $\Omega$  de  $\mathbb{R} \times \mathbb{R}^n$  est localement lipschitzienne en  $x$  si pour tout  $(t_0, x_0) \in \Omega$ , il existe un voisinage  $V$  de  $(t_0, x_0)$  et  $k > 0$  tels que

$$\forall (t, x) \in V, \forall x' : (t, x') \in V, \|f(t, x) - f(t, x')\| \leq k \|x - x'\|$$

Introduisons maintenant une notion essentielle pour démontrer le théorème de Cauchy-Lipschitz :

**Définition 4 (Cylindre de sécurité)** Soit  $(t_0, x_0) \in \Omega$ , on pose

$$\mathcal{B}_r = \{x \in \mathbb{R}^n / \|x - x_0\| \leq r\}$$

$$I_\alpha = ]t_0 - \alpha, t_0 + \alpha[$$

Soit  $f : \Omega \rightarrow \mathbb{R}^n$  une application continue, localement lipschitzienne en  $x$ ,  
Soit  $V$  un voisinage compact de  $(t_0, x_0)$  sur lequel  $f$  est  $k$ -lipschitzienne  
 $V$  étant compact, on peut majorer  $\|f(t, x)\|$  par  $M$  sur  $V$   
On peut alors prendre  $r > 0$  et  $\alpha > 0$  tels que

$$I_\alpha \times \mathcal{B}_r \subset V \quad \text{et} \quad \alpha M < r$$

Un tel voisinage  $I_\alpha \times \mathcal{B}_r$  est appelé cylindre de sécurité pour  $f$  en  $(t_0, x_0)$

Un cylindre de sécurité étant fixé, l'opérateur  $\mathcal{T}$  défini sur  $\mathcal{C}^0(I_\alpha, \mathbb{R}^n)$  par

$$\mathcal{T}\Psi(t) = x_0 + \int_{t_0}^t f(s, \Psi(s)) \, ds$$

laisse stable l'ensemble fermé  $\mathcal{X} = \{\Psi \in \mathcal{C}^0(I_\alpha, \mathbb{R}^n) / \Psi(I_\alpha) \subset \mathcal{B}_r\}$  : en effet, si  $\Psi \in \mathcal{X}$ , alors  $\forall t \in I_\alpha$ ,

$$\|\mathcal{T}\Psi(t) - x_0\| \leq \left| \int_{t_0}^t \|f(s, \Psi(s))\| \, ds \right| \leq \alpha M < r$$

et donc  $\mathcal{T}\Psi \in \mathcal{X}$ .

Le fait que  $f$  soit lipschitzienne sur le cylindre de sécurité va alors permettre de montrer que  $\mathcal{T}$  (ou l'un de ses itérés) est une contraction de  $\mathcal{X}$  qui est complet ; son unique point fixe vérifie alors (6), ce qui permet de prouver le

**Théorème 1 (Cauchy-Lipschitz)** Soit  $\Omega$  un ouvert de  $\mathbb{R} \times \mathbb{R}^n$ ,  $f : \Omega \rightarrow \mathbb{R}^n$  une application continue, localement lipschitzienne en  $x$  et soit  $I_\alpha \times \mathcal{B}_r$  un cylindre de sécurité pour  $f$  en  $(t_0, x_0) \in \Omega$   
Alors il existe une unique solution  $\varphi : I_\alpha \rightarrow \mathcal{B}_r$  de  $\mathcal{E} : y' = f(t, y)$  vérifiant la condition de Cauchy  $\varphi(t_0) = x_0$

## 2.3 Solution maximale

A partir du théorème de Cauchy-Lipschitz, qui est un théorème local, nous cherchons maintenant à définir des solutions globales. Tout repose sur le lemme suivant, qui est une de ses conséquences :

**Lemme 2 (de recollement)** Soient  $\varphi_1 : I_1 \rightarrow \mathbb{R}^n$  et  $\varphi_2 : I_2 \rightarrow \mathbb{R}^n$  deux solutions de  $\mathcal{E}$ , soit  $t_0 \in I_1 \cap I_2$  tel que  $\varphi_1(t_0) = \varphi_2(t_0)$ .  
Alors  $\forall t \in I_1 \cap I_2$ ,  $\varphi_1(t) = \varphi_2(t)$

L'idée de solution maximale, ainsi que le théorème qui suit et qui prouve leur existence, découlent naturellement du lemme de recollement :

**Définitions 5** Une solution  $\varphi_1 : I_1 \rightarrow \mathbb{R}^n$  prolonge une solution  $\varphi_2 : I_2 \rightarrow \mathbb{R}^n$  si  $I_2 \subset I_1$  et  $\forall t \in I_1, \varphi_1(t) = \varphi_2(t)$

$\varphi_1$  est dite solution maximale de  $\mathcal{E}$  s'il n'existe pas d'autre solution de  $\mathcal{E}$  qui la prolonge.

**Théorème 3 (Existence et unicité globale)** Pour tout  $(t_0, x_0) \in \Omega$ , il existe une unique solution maximale  $\varphi$  de  $\mathcal{E}$  prenant la valeur  $x_0$  en  $t_0$ . Cette solution maximale est définie sur un intervalle ouvert de  $\mathbb{R}$  noté  $J = ]\omega_-, \omega_+[$

Enfin le théorème qui suit explique dans quels cas une solution ne peut pas être prolongeable :

**Théorème 4** Soit  $\varphi : J = ]\omega_-, \omega_+[ \rightarrow \mathbb{R}^n$  une solution maximale de  $\mathcal{E}$ . Lorsque  $t \rightarrow \omega_\pm, (\varphi(t), t)$  sort de tout compact de  $\Omega$  (tend vers sa frontière  $\partial\Omega$ ).

Plus précisément, pour tout compact  $K$  de  $\Omega$  il existe un voisinage  $v$  de  $\omega_\pm$  dans  $\mathbb{R}$  tel que  $(t, \varphi(t)) \notin K$  si  $t \in v \cap J$

Ainsi, soit la solution est définie à tout instant ( $\omega_\pm = \pm\infty$ ), soit elle tend vers l'infini en temps fini.

## 2.4 Théorème de Péano

L'existence au problème de Cauchy reste encore vraie si  $f$  est seulement supposée continue (théorème de Péano). Ce théorème se prouve en utilisant la suite des polygones d'Euler, qui approximent localement la solution par sa tangente ; cette suite est ainsi définie : on se donne  $h > 0$  et la suite  $(t_n, x_n)$  :

$$t_{n+1} = t_n + h \quad \text{et} \quad x_{n+1} = x_n + hf(t_n, x_n)$$

On note  $x_h(t)$  la fonction linéaire par morceaux qui passe par les points  $(t_n, x_n)$ .

La suite des polygones d'Euler est alors la suite des  $x_h(t)$  pour  $h = \frac{\alpha}{n}$  où  $n \in \mathbb{N}^*$  et  $\alpha$  est celui d'un cylindre de sécurité. Cette suite possède une sous-suite qui converge uniformément vers une solution du problème de Cauchy. Cependant il n'y a plus unicité.

## 3 Equations différentielles linéaires et affines

On considère l'équation affine

$$\mathcal{E}_a : \quad y' = \mathcal{A}(t) \cdot y + b(t) \quad \text{où} \quad \mathcal{A} : I \rightarrow \mathcal{L}(\mathbb{R}^n) \quad \text{et} \quad b : I \rightarrow \mathbb{R}^n$$

On est dans le cas d'application du théorème de Cauchy-Lipschitz, il y a donc existence et unicité pour tout problème de Cauchy. De plus, on a le résultat suivant :

**Théorème 5** *Les solutions maximales de  $\mathcal{E}_a$  sont définies sur tout  $I$*

### 3.1 Equations différentielles linéaires dépendant du temps

Soit l'équation différentielle linéaire homogène associée à  $\mathcal{E}_a$

$$\mathcal{E}_l : y' = \mathcal{A}(t) \cdot y \quad \text{avec} \quad \mathcal{A} : I \rightarrow \mathcal{L}(\mathbb{R}^n)$$

Soit  $(t_0, x_0) \in I \times \mathbb{R}^n$ , on note  $\varphi_{t_0, x_0}$  la solution maximale de  $\mathcal{E}_l$  qui vérifie la condition de Cauchy  $(t_0, x_0)$  :

$$\varphi_{t_0, x_0} : I \rightarrow \mathbb{R}^n \quad \text{et} \quad \varphi_{t_0, x_0}(t_0) = x_0$$

La proposition suivante découle de la linéarité de l'équation et du théorème de Cauchy-Lipschitz. Elle est essentielle, car elle décrit de manière très précise l'ensemble des solutions dans le cas linéaire.

**Proposition 6** *L'ensemble  $\mathcal{S}$  des solutions maximales de  $\mathcal{E}_l$  est un  $\mathbb{R}$ -espace vectoriel et pour tout  $t \in I$ , l'application*

$$\theta_t : \mathcal{S} \rightarrow \mathbb{R}^n \quad \varphi \mapsto \varphi(t)$$

*est un isomorphisme de  $\mathcal{S}$  sur  $\mathbb{R}^n$*

Nous introduisons maintenant un outil qui va permettre de calculer facilement toutes les solutions de  $\mathcal{E}_l$ , quelle que soit la condition de Cauchy :

**Résolvante.** Soit  $(t_0, t) \in I^2$ , l'application  $\theta_t \circ \theta_{t_0}^{-1} = R(t, t_0) \in \mathcal{GL}(\mathbb{R}^n)$  d'après ce qui précède. De plus :

$$\forall \varphi \in \mathcal{S}, \quad \varphi(t) = \theta_t(\varphi) = R(t, t_0) \circ \theta_{t_0}(\varphi) = R(t, t_0) \cdot \varphi(t_0)$$

Donc on obtient la forme très intéressante :

$$\varphi_{t_0, x_0}(t) = R(t, t_0) \cdot x_0 \tag{7}$$

**Définition 6** *L'application  $R : I^2 \rightarrow \mathcal{GL}(\mathbb{R}^n)$  est appelée résolvante de  $\mathcal{E}_l$*

**Proposition 7** *La résolvante vérifie les propriétés suivantes :*

$$R'(t, t_0) = \mathcal{A}(t) \circ R(t, t_0) \tag{8}$$

$$R(t, t) = id_{\mathbb{R}^n} \tag{9}$$

$$R(t, t') \circ R(t', t'') = R(t, t'') \tag{10}$$

$$R(t, t_0)^{-1} = R(t_0, t) \tag{11}$$

(8) se déduit de la dérivation de (7), les autres propriétés sont immédiates. Ainsi, on obtient un moyen de calculer la résolvante :

**Corollaire 8** *L'application  $t \rightarrow R(t, t_0)$  est l'unique solution de l'équation différentielle dans  $\mathcal{L}(\mathbb{R}^n)$  :  $U' = \mathcal{A}(t) \circ U$  satisfaisant la condition  $U(t_0) = id_{\mathbb{R}^n}$*

**Variation des constantes** La question qui se pose maintenant est celle de la résolution de l'équation avec second membre  $\mathcal{E}_a$ . Or, cela peut se faire dès que l'on connaît la résolvante de  $\mathcal{E}_l$ . L'idée de la *méthode de la variation de la constante* est de chercher la solution de  $\mathcal{E}_a$  sous la forme :

$$\varphi(t) = R(t, t_0) \cdot \psi(t)$$

En dérivant, on obtient une équation nécessairement vérifiée par  $\psi'$ , ce qui permet d'obtenir l'expression de la solution  $\varphi$  de  $\mathcal{E}_a$  qui vérifie  $\varphi(t_0) = x_0$  :

$$\varphi(t) = R(t, t_0) \cdot [x_0 + (\int_{t_0}^t R(t_0, s) \cdot b(s) ds)] \quad (12)$$

### 3.2 Equations différentielles à coefficients constants

Dans le cas où  $\mathcal{A}$  ne dépend pas de  $t$ , on peut résoudre  $\mathcal{E}_l$  à l'aide d'une exponentielle d'endomorphisme. En effet, notons

$$\forall \mathcal{A} \in \mathcal{L}(\mathbb{R}^n), \quad \exp(\mathcal{A}) = \sum_{k=0}^{+\infty} \frac{\mathcal{A}^k}{k!}$$

Alors on peut montrer à l'aide du corollaire (8) que :

$$R(t, t_0) = \exp((t - t_0)\mathcal{A})$$

donc que la solution  $\varphi_{t_0, x_0}$  qui prend la valeur  $x_0$  en  $t_0$  a pour expression :

$$\varphi_{t_0, x_0}(t) = e^{(t-t_0)\mathcal{A}} \cdot x_0 \quad (13)$$

La réduction de  $\mathcal{A}$  permet de calculer facilement l'endomorphisme  $\exp(\mathcal{A})$  et donc d'obtenir les solutions de  $\mathcal{E}_l$  et de  $\mathcal{E}_a$  par (12).

**Forme générale des solutions** . Il est possible de préciser la forme générale des solutions. Factorisons le polynôme caractéristique de  $\mathcal{A}$  dans  $\mathbb{C}[X]$  sous la forme :  $\prod_{i=1}^r (X - \lambda_i)^{m_i}$ . Alors on peut montrer à l'aide de résultats d'algèbre linéaire que les solutions de  $\mathcal{E}_l$  sont les applications :

$$t \mapsto \sum_{i=1}^r e^{\lambda_i t} P_i(t) \quad \text{où} \quad P_i \in \mathbb{C}[X], \deg(P_i) < m_i$$

ce qui permet de trouver la forme générale des solutions sur  $\mathbb{R}$  comme somme d'expressions du type :

$$t \mapsto t^p e^{\alpha t} (A \cos \beta t + B \sin \beta t)$$

avec  $A, B \in \mathbb{R}^n$ ,  $\lambda = \alpha + i\beta$  valeur propre de  $\mathcal{A}$  d'ordre  $m_\lambda$  et  $p < m_\lambda$ .

**Comportement asymptotique des solutions** Le résultat suivant permet de décrire le comportement des solutions de  $\mathcal{E}_l$  lorsque  $t \rightarrow \infty$  :

**Théorème 9**  $\lim_{t \rightarrow \infty} e^{t\mathcal{A}} = 0$  si, et seulement si toutes les valeurs propres de  $\mathcal{A}$  ont une partie réelle négative.

## 4 Equations différentielles autonomes

### 4.1 Définition et propriété fondamentale

**Définition 7** L'équation différentielle  $y' = f(t, y)$  est dite autonome si  $f$  est indépendante du temps :  $f(t, y) = f(y)$

Les équations différentielles autonomes présentent la propriété très intéressante d'*invariance par translation* : si  $\varphi$  est solution de  $\mathcal{E} : y' = f(y)$  sur l'intervalle  $I$  alors  $\tau_\alpha \varphi : t \mapsto \varphi(t + \alpha)$  est aussi solution de  $\mathcal{E}$  sur l'intervalle  $I - \alpha$ .

Ce n'est donc pas une restriction de prendre les problèmes de Cauchy en  $t = 0$  : nous notons

$$\varphi_x : J_x = ]\alpha_x, \omega_x[ \rightarrow \mathbb{R}^n$$

la solution maximale de  $y' = f(y)$  vérifiant  $\varphi_x(0) = x$ .

### 4.2 Portrait de phase

**Définition 8** L'image  $\gamma(x)$  de  $J_x$  par  $\varphi_x$  est appelée orbite du point  $x$  pour  $f$

Les orbites de  $f$  définissent une partition de l'ouvert  $U \subset \mathbb{R}^n$  sur lequel  $f$  est définie. En effet, soit  $x_1 = \varphi_x(t_1)$  un point de  $\gamma(x)$  : alors  $\tau_{t_1} \varphi_x$  vérifie  $\tau_{t_1} \varphi_x(0) = \varphi_x(t_1) = x_1$  : c'est la solution maximale  $\varphi_{x_1}$  de  $\mathcal{E}$ . Donc  $\gamma(x) = \gamma(x_1)$  dès que  $x_1 \in \gamma(x)$ .

Le *portrait de phase* de  $f$  est la partition de  $U$  en orbites  $\gamma(x)$ .

### Typologie des orbites

**Définition 9** On dit que  $x \in U$  est un point singulier de  $f$  si  $f(x) = 0$

1. Si  $x$  est un point singulier, son orbite  $\gamma(x)$  vaut  $\{x\}$ .

2. Sinon,  $\varphi_x$  a une dérivée qui ne s'annule jamais. On a alors deux cas :
- Si  $\varphi_x$  est non injective, on montre qu'elle est périodique et que  $\gamma(x)$  est homéomorphe au cercle  $S^1$ . On dit que  $\gamma(x)$  est un *cycle* de  $f$ .
  - Si  $\varphi_x$  est injective,  $\gamma(x)$  est une *orbite non compacte*. On fait alors séparément l'étude des *demi-orbites* : notant  $J_x = ]\alpha_x, \omega_x[$  on pose

$$\gamma^+(x) = \varphi_x([0, \omega_x[) \quad \text{et} \quad \gamma^-(x) = \varphi_x(]\alpha_x, 0])$$

Quitte à changer  $f$  en  $-f$ , limitons-nous à l'étude de  $\gamma^+(x)$

Si  $\omega_x < +\infty$ ,  $\varphi_x$  tend vers la frontière de  $U$  quand  $t \rightarrow \omega_x$  d'après le th. 4

Sinon, de nombreux cas peuvent se présenter. On appelle *ensemble  $\omega$ -limite* de  $\gamma(x)$  l'ensemble des points adhérents à  $\gamma(x)$  au voisinage des temps infinis :

$$\omega(x) = \{a \in U / \exists (t_n)_{n \in \mathbb{N}} : \lim t_n = +\infty \text{ et } \lim \varphi_x(t_n) = a\}$$

L'ensemble  $\alpha$ -limite de  $x$ , noté  $\alpha(x)$  est l'ensemble  $\omega$ -limite de  $x$  pour le champ de vecteurs  $-f$ .

### 4.3 Flot

Soit

$$V = \bigcup_{x \in U} J_x \times \{x\}$$

$V$  est un ouvert de  $\mathbb{R} \times U$ . Le *flot* de  $f$  est l'application :

$$\phi : V \rightarrow U \quad \phi(t, x) = \varphi_x(t)$$

Le flot a les propriétés suivantes (la deuxième est due à l'invariance par translation), qui sont celles d'un *groupe local de transformations agissant sur  $U$*  (voir [4] p 21) :

$$\forall x \in U \quad \phi(0, x) = x \tag{14}$$

$$\forall (t, x), (t + t', x) \in V \quad \phi(t, \phi(t', x)) = \phi(t + t', x) \tag{15}$$

En notant  $\phi(t, x) = t \cdot x$  on voit que l'on a presque défini une action de groupe, l'orbite de  $x$  sous cette action coïncide alors avec la définition 8.

Si  $V = \mathbb{R} \times U$ , on dit que  $f$  est *complet*. On peut montrer, à l'aide du théorème 4, que si  $U$  est borné, alors  $f$  est complet. Dans ce cas, on définit

$$\forall t \in \mathbb{R}, \quad \phi_t : U \rightarrow U \quad \phi_t(x) = \phi(t, x) = \varphi_x(t)$$

(C'est-à-dire que  $\phi_t$  associe à tout point  $x \in U$ , la valeur à  $t$  de la solution qui valait  $x$  à  $t = 0$ ) Les propriétés 14 et 15 se réécrivent :

$$\phi_0 = id_U \quad \text{et} \quad \forall (t, t') \in \mathbb{R}^2, \quad \phi_{t+t'} = \phi_t \circ \phi_{t'} \quad \text{donc} \quad \phi_t \circ \phi_{-t} = id_U$$

$t \rightarrow \phi_t$  est donc un morphisme de  $(\mathbb{R}, +)$  sur le groupe des difféomorphismes de  $U$  : on dit que  $(\phi_t)_{t \in \mathbb{R}}$  est un *groupe à un paramètre*.



## 5 Etude locale du flot

### 5.1 Au voisinage d'un point régulier

Au voisinage d'un point  $a$  régulier, le théorème de redressement décrit bien le comportement du flot.

**Définition 10** Soit  $f$  un champ de vecteurs défini sur un ouvert  $U$  de  $\mathbb{R}^n$  et  $g$  un champ de vecteurs défini sur  $V$ . Soit  $\psi$  une application différentiable de  $U$  sur  $V$ . On dit que  $f$  et  $g$  sont reliés par  $\psi$  si :  
 $\varphi$  solution de  $y' = f(y)$  sur  $U \iff \psi \circ \varphi$  solution de  $y' = g(y)$  sur  $V$

**Théorème 10 (de redressement du flot)** Soit  $f$  un champ de vecteurs de classe  $C^r$  sur  $U$  et  $a \in U$  tel que  $f(a) \neq 0$ . Alors il existe un  $C^r$ -difféomorphisme  $\psi$  d'un voisinage de  $a$  sur un voisinage de  $0$  qui relie  $f$  au champ de vecteurs  $(1, 0, \dots, 0)$

Au voisinage de  $a$ , la structure du flot est donc très simple car elle est équivalente à celle du flot  $\phi(t, x) = x + (t, 0, \dots, 0)$ .

### 5.2 Au voisinage d'un point singulier hyperbolique

**Définition 11** Soit  $a \in U$  un point singulier :  $f(a) = 0$ . On dit que  $a$  est une singularité hyperbolique de  $f$  si les valeurs propres de  $df_a$  ont toutes une partie réelle non nulle.

Au voisinage d'un point singulier hyperbolique, nous allons voir que l'on peut aussi donner un modèle au flot de  $f$ .

**Définition 12** On dit que deux champs de vecteurs  $f$  et  $g$  sont  $C^r$ -conjugués ( $r \geq 1$ ) s'ils sont reliés (au sens de la def 10) par un  $C^r$ -difféomorphisme  $h$ .

Soient  $\phi$  et  $\psi$  les flots respectifs de  $f$  et  $g$  : cette propriété s'écrit  $\psi(t, h(x)) = h(t, \phi(x))$ <sup>1</sup>. Cette égalité garde un sens si  $h$  est un homéomorphisme et on dit alors que  $f$  et  $g$  sont  $C^0$ -conjugués par  $h$ .

**Théorème 11 (Hartman - Grobman)** Soit  $a \in U$  une singularité hyperbolique de  $f$ , alors au voisinage de  $a$ ,  $f$  est  $C^0$ -conjugué à son linéarisé  $x \mapsto df_a(x - a)$

Il est donc très important de classifier les champs de vecteurs *linéaires* hyperboliques. Ceci s'effectue à l'aide du théorème suivant :

---

<sup>1</sup>soit encore  $\psi \cdot h = h \cdot \phi$  ce qui justifie le terme *conjugués*

**Théorème 12 (Classification des champs linéaires hyperboliques)** *Soient  $\mathcal{A}$  et  $\mathcal{B}$  deux endomorphismes de  $\mathbb{R}^n$  dont toutes les valeurs propres ont des parties réelles non nulles. Alors les champs de vecteurs  $x \mapsto \mathcal{A}(x)$  et  $x \mapsto \mathcal{B}(x)$  sont  $\mathcal{C}^0$ -conjugués si, et seulement si  $\mathcal{A}$  et  $\mathcal{B}$  ont le même nombre de valeurs propres à partie réelle négative.*

Il y a donc  $(n+1)$  classes d'équivalence pour la relation de  $\mathcal{C}^0$ -conjugaison, chacune ayant pour représentant le champ de vecteurs associé à l'application linéaire définie, en notant  $(e_1, \dots, e_n)$  la base canonique de  $\mathbb{R}^n$ , par

$$\mathcal{R}_p(e_i) = \begin{cases} -1 & \text{si } i \leq p \\ 1 & \text{si } i > p \end{cases} \quad p = 0, \dots, n$$

Les résultats qui précèdent sont résumés par le

**Théorème 13 (Classification topologique des singularités hyperboliques)**

*Soit  $f$  un champ de vecteurs  $\mathcal{C}^r, r \geq 1$  sur un ouvert  $U$  de  $\mathbb{R}^n$ , soit  $a$  un point singulier hyperbolique de  $f$ , soit  $p$  le nombre de valeurs propres de  $df_a$  à partie réelle strictement négative. Alors il existe un homéomorphisme  $h$  d'un voisinage de 0 sur un voisinage  $U'$  de  $a$  tel que le flot  $\phi$  de  $f$  sur  $U$  vérifie*

$$\forall x \in U', \quad \phi(t, h(x)) = h(e^{t\mathcal{R}_p} \cdot x)$$

En particulier, en dimension 2 :

- Si les parties réelles des valeurs propres de  $df_a$  sont toutes deux négatives,  $a$  est un attracteur.
- Si les parties réelles des valeurs propres de  $df_a$  sont toutes deux positives,  $a$  est un répulseur.
- Sinon l'une est positive et l'autre négative,  $a$  est un col.

## 6 Etude globale du flot en dimension 2

Après avoir fait l'étude du flot au voisinage de points réguliers et de points singuliers hyperboliques, nous cherchons à faire une étude des propriétés globales du champ de vecteurs en essayant de déterminer les ensemble  $\alpha$ - et  $\omega$ -limites. Le théorème de Poincaré-Bendixson montre que ceux-ci ont une forme simple en dimension 2, ce qui n'est pas le cas en dimension  $> 2$ .

On suppose que  $f$  est un champ de vecteurs complet sur  $U \subset \mathbb{R}^2$  et on considère le groupe à un paramètre  $(\phi_t)_{t \in \mathbb{R}}$ .

### 6.1 Ensembles invariants, ensembles limites

**Définition 13** *Une partie  $A$  de  $U$  est dite invariante si  $\forall t \in \mathbb{R}, \phi_t(A) \subset A$ .  $A$  est positivement-invariante si  $\forall t > 0, \phi_t(A) \subset A$ .*

Une partie invariante piège les trajectoires ; une trajectoire est l'exemple le plus simple d'un ensemble invariant. Le théorème du point fixe permet de montrer le

**Théorème 14** *Soit  $A$  un ensemble positivement invariant homéomorphe à la boule unité compacte, alors  $f$  possède un point singulier  $a \in A$*

Intéressons-nous maintenant à l'action du flot sur l'ensemble  $\omega$ -limite :

**Théorème 15** *Soit  $x \in U$ , alors*

$$\forall t \in \mathbb{R}, \quad \omega(x) = \omega(\phi_t(x))$$

$$\forall y \in \omega(x), \forall t \in \mathbb{R}, \quad \phi_t(y) \in \omega(x) \quad \text{donc} \quad \gamma(y) \subset \omega(x)$$

*On dit que  $\omega(x)$  est invariant sous le flot.*

On peut aussi obtenir des propriétés topologiques sur  $\omega(x)$  avec une hypothèse supplémentaire :

**Théorème 16** *S'il existe un compact  $K \subset U$  tel que la demi-orbite positive  $\gamma^+(x) \subset K$ , alors  $\omega(x)$  est non vide, fermé (donc compact), et connexe.*

## 6.2 Théorème de Poincaré-Bendixson

Le théorème qui suit repose sur le lemme de Jordan (qui dit qu'un arc fermé simple sépare le plan en deux composantes connexes), ce qui explique qu'il n'est valable qu'en dimension 2.

**Théorème 17 (Poincaré-Bendixson)** *Soit  $x \in U$  et un compact  $K$  de  $U$  tel que  $\gamma^+(x) \subset K$ . Alors l'ensemble  $\omega$ -limite est l'un des trois types suivants :*

1.  $\omega(x)$  est un point critique de  $f$
2.  $\omega(x)$  est un cycle de  $f$
3.  $\omega(x)$  est la réunion d'un nombre fini de points singuliers et  $f$  et d'orbites de  $f$  qui joignent ces points.

## Références

- [1] R. MOUSSU, Systèmes dynamiques : aspects théoriques, Cours de l'Ecole Polytechnique, 1991
- [2] M. SCHATZMAN, Analyse Numérique. Approche Mathématique, Dunod, 2001
- [3] X. GOURDON, Les maths en tête, Analyse, Ellipses, 1994
- [4] P. J. OLVER, Applications of Lie Groups to Differential Equations, Springer-Verlag, 2000
- [5] C DOSS-BACHELET, J.P. FRANCOISE, C. PIQUET, Géométrie différentielle avec 80 figures, Ellipses, 2000
- [6] S CANTAT, Théorème de Poincaré-Bendixson, Le journal de maths des élèves - ENS Lyon, Volume 1, No. 3, 1995

# Equations différentielles ordinaires

## Méthodes de résolution numérique

Miguel GILLOT

11 juin 2006

## 1 Introduction

On s'intéresse à la résolution numérique d'un système différentiel

$$u(t_0) = u_0 \in \mathbb{R}^d \quad (1)$$

$$\forall t \in [t_0, T], u'(t) = f(t, u(t)) \quad (2)$$

On suppose que ce système vérifie les conditions de Cauchy. Les méthodes utilisées pour calculer des fonctions approchant la solution ( que l'on notera  $u$  ) de cette équation consistent à prendre une subdivision  $(t_n)_{n \in [0, N]}$  de l'intervalle  $[t_0, T]$  et à calculer des valeurs approchées de  $u(t_n)$ , que l'on notera  $U_n$ . Elles se fondent sur des méthodes d'approximation de l'intégrale :

$$u_{n+1} = u_n + \int_{t_n}^{t_{n+1}} f(t, u(t)) \quad (3)$$

On distingue les méthodes à un pas qui déterminent  $U_{n+1}$  uniquement à partir de la valeur de  $U_n$  et les méthodes multipas qui déterminent  $U_{n+1}$  à partir de  $q$  valeurs  $U_n, U_{n-1}, \dots, U_{n-q+1}$  avec  $q \geq 2$ . L'étude de l'erreur tient une place prépondérante dans ces schémas, comme nous le verrons dans les définitions de la consistance et de la stabilité d'un schéma.

## 2 Schéma à un pas

### 2.1 Quelques définitions

**Définition 1.** Soit  $(t_n)_{n \in [0, N]}$  une subdivision de l'intervalle  $[t_0, T]$

On appelle schéma à un pas la donnée d'une fonction

$$F : [t_0, T] \times \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}^d$$

et d'une suite récurrente de  $N + 1$  éléments de  $\mathbb{R}^d$  vérifiant

$$U_0 \in \mathbb{R}^d$$

$$\forall k \in [0, N-1], U_{k+1} = U_k + h_k F(t_k, U_k, h) \quad (4)$$

Avec :

$$t_{k+1} = t_k + h_k$$

$U_k$  constitue une approximation de  $u(t_k)$ , solution exacte estimée au point  $t_k$ . Il est important de noter que dans ces schémas numériques, on distingue trois types d'erreurs :

- L'erreur systématique correspondant à l'approximation de  $u(t_{k+1}) - u(t_k)$  par  $h_k F(t_k, u_k, h_k)$
- L'erreur d'arrondi due aux calculs approchés effectués par ordinateur.
- L'erreur héritée due à la propagation à  $U_{k+1}$  des erreurs commises sur  $U_k$ .

Dans toute la suite, on notera  $h = \sup_{k \in [0, N]} h_k$  et on appellera cette quantité le pas de la méthode.

On définit la convergence d'un schéma comme étant la convergence de chaque valeur du schéma vers la valeur de la solution du système lorsque le pas tend vers 0 et la condition initiale du schéma tend vers la condition initiale du système.

**Définition 2.** *L'approximation de (1) – (2) définie par le schéma à un pas (3) est dite convergente si*

$$\forall u_0 \in \mathbb{R}^d, \lim_{U_0 \rightarrow u_0; h \rightarrow 0} \sup_{0 \leq k \leq N} |u(t_k) - U_k| = 0$$

Lorsqu'on calcule une approximation de la fonction en un point, on commet une erreur qui est susceptible d'augmenter au fur et à mesure que l'on calcule les approximations successives. La condition de stabilité nous dit que l'erreur cumulée qu'apporte le schéma ne devient "pas trop grande".

**Définition 3.** *Un schéma à un pas est dit stable s'il existe une constante  $M$  telle que pour tout  $U_0 \in \mathbb{R}^d$ , pour tout  $V_0 \in \mathbb{R}^d$ , pour tout  $h$  et pour toute suite de vecteur  $\varepsilon_j$ , les suites  $U_j$  et  $V_j$  définies par*

$$U_{j+1} = U_j + h_j F(t_j, U_j, h)$$

$$V_{j+1} = V_j + h_j F(t_j, V_j, h) + \varepsilon_j$$

*vérifient :*

$$\forall j \leq N, |U_j - V_j| \leq M(|U_0 - V_0| + \sum_{0 \leq j \leq j-1} |\varepsilon_j|)$$

Par ailleurs, il est important que le schéma calcule quelque chose qui soit proche de la solution de l'équation. La condition de consistance que nous allons voir implique que le schéma s'écarte peu localement de la solution.

**Définition 4.** Un schéma à un pas est dit consistant avec (2) si pour toute solution de (2), on a

$$\lim_{\sup h_k \rightarrow 0} \sum_{0 \leq j \leq N-1} |u(t_{j+1}) - u(t_j) - h_j F(t_j, u(t_j), h)| = 0$$

On peut interpréter graphiquement le vecteur

$$\varepsilon_j = u(t_{j+1}) - u(t_j) - h_j F(t_j, u(t_j), h)$$

comme étant l'erreur qu'on commet en remplaçant  $u_{j+1}$  par la quantité calculée à l'aide du schéma  $(u(t_j) + h_k F(t_j, u(t_j), h))$ . Ce vecteur est appelé l'erreur locale. Une autre formulation de la consistance est de dire que les erreurs locale cumulées tendent vers 0 quand le pas du schéma tend vers 0

L'intérêt des conditions de stabilité et de consistance est qu'elles sont suffisantes pour garantir la convergence du schéma.

## 2.2 Conditions de convergence

**Théorème 2.1** (Consistance plus stabilité impliquent convergence). *Soit  $f$  une fonction satisfaisant les conditions de Cauchy-Lipschitz et soit  $F : [t_0, T] \times \mathbb{R}^d \times [0, h^*] \rightarrow \mathbb{R}^d$  une fonction **continue** définissant un schéma à un pas. Si ce schéma est consistant avec et s'il est stable, il est convergent.*

Nous allons donner des conditions qui assurent la stabilité et la convergence.

**Théorème 2.2** (consistance). *Soit  $F : [t_0, T] \times \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}^d$  une fonction **continue** définissant un schéma à un pas. Ce schéma est consistant si et seulement si :*

$$\forall t \in [t_0, T], \forall u \in \mathbb{R}^d, \lim_{h \rightarrow 0} F(t, u, h) = f(t, u)$$

**Théorème 2.3** (stabilité). *Pour qu'un schéma à un pas soit stable, il suffit qu'il existe une constante  $\Lambda$  telle que*

$$\forall t \in [t_0, T], \forall u, v \in \mathbb{R}^d, \forall h \in [0, h^*],$$

$$|F(t, u, h) - F(t, v, h)| \leq \Lambda |u - v|$$

## 2.3 Ordre d'un schéma à un pas

L'ordre d'un schéma à un pas majore l'erreur de consistance et permet de caractériser sa vitesse de convergence.

**Définition 5** (Ordre d'un schéma). *Soit  $p$  un entier  $\geq 1$ . Un schéma à un pas est dit d'ordre  $p$  si pour toute solution  $u$  de l'équation, il existe une constante  $C$  positive tel que*

$$\sum_{0 \leq j \leq N-1} |u(t_{j+1}) - u(t_j) - h_j F(t_j, u(t_j), h)| \leq Ch^p$$

Notons qu'un schéma est consistant si et seulement si il est d'ordre 1.

**Théorème 2.4** (vitesse de convergence). *Soit  $f$  une fonction satisfaisant les conditions de Cauchy-Lipschitz et soit  $F : [t_0, T] \times \mathbb{R}^d \times [0, h^*] \rightarrow \mathbb{R}^d$  une fonction **continue** définissant un schéma à un pas. Si ce schéma est d'ordre  $p$ , s'il est stable et si  $|u_0 - U_0| \leq C'h^p$ , alors*

$$\sup_{j \leq N} |u(t_j) - U(t_j)| \leq M(C + C')h^p$$

Enfin, avec des fonctions suffisamment régulières, on peut donner une condition nécessaire et suffisante pour que le schéma soit d'ordre  $p$

**Théorème 2.5.** *Soit  $f$  une fonction de classe  $C^p$ . On définit la suite de fonctions  $(f_k)_{0 \leq k \leq p}$  par*

$$\begin{aligned} f_0(t, u) &= f(t, u) \\ f_{k+1}(t, u) &= \frac{\partial f_k}{\partial t}(t, u) + Df_k(t, u) \cdot f(t, u) \end{aligned}$$

*Si  $F$  est  $p$  fois dérivable par rapport à  $h$  et que les dérivées  $p$ -ième par rapport à  $h$  sont des fonctions continues de toutes les variables, alors le schéma est d'ordre  $p$  si et seulement si*

$$\forall k \in [0, p-1], \forall t \in [t_0, T], \forall u \in \mathbb{R}^d, \frac{\partial F_k}{\partial h^k}(t, u, 0) = \frac{f_k}{k+1}$$

## 2.4 les méthodes de Runge-Kutta

Les méthodes de Runge-Kutta sont fondées sur le calcul de l'intégrale (3) par des méthodes de quadrature qui introduisent des points intermédiaires sur le segment d'intégration et qui détermine la valeur de l'intégrale à partir des valeurs de la fonction en ces points intermédiaires.

**Définition 6** (Méthode de Runge-Kutta). *Une méthode de Runge-Kutta à  $q$  étages est donnée par :*

$$\begin{aligned} t_{k,i} &= t_k + c_i h_k \\ U_{k,i} &= U_k + h_k \sum_{j=1}^q a_{ij} f(t_{k,j}, U_{k,j}) \\ U_{k+1} &= U_k + h_k \sum_{j=1}^q b_j f(t_{k,j}, U_{k,j}) \end{aligned}$$



La méthode est dite explicite si  $\forall(i, j), j \geq i \implies a_{ij} = 0$  (les termes  $U_{k+1}$  se calculent itérativement à partir des  $U_j, j \leq k$ ).

Elle est dite implicite sinon (pour calculer  $U_{k+1}$ , il faut résoudre un système d'équations non linéaires couplées sur les  $U_{k,i}$ ).

On résume les coefficients dans un tableau de la forme :

$c_1$	$a_{11}$	$a_{12}$	$\dots$	$a_{1q}$
$c_2$	$a_{21}$	$a_{22}$	$\dots$	$a_{2q}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$c_q$	$a_{q1}$	$a_{q2}$	$\dots$	$a_{qq}$
	$b_1$	$b_2$	$\dots$	$b_q$

Le nombre d'étage représente le nombre de points intermédiaires (compris dans l'intervalle  $[t_0, T]$ , ) nécessaires au calcul de  $U_{k+1}$ .

On montre très simplement le résultat suivant :

**Théorème 2.6** (Convergence des méthodes de Runge-Kutta). *Si la fonction  $f$  est continuellement dérivable, les méthodes de Runge-Kutta sont consistantes si et seulement si :*

$$\sum_{j=1}^q a_{ij} = 1$$

Présentons la plus simple des méthodes de Runge-Kutta : la méthode d'Euler. Elle revient à approcher l'intégral ( 3 ) par une méthode des rectangles.

**Définition 7** (Méthode d'Euler explicite). *La méthode d'Euler explicite est définie par :*

$$U_{k+1} = U_k + h_k F(t_k, U_k, h)$$

Avec

$$F(t_k, U_k, h) = f(t_k, U_k)$$

La méthode d'Euler implicite est définie par :

$$U_{k+1} = U_k + h_k f(t_{k+1}, U_{k+1})$$

On montre que les méthodes d'Euler explicite et implicite sont stables et d'ordre 1.

## 2.5 Calcul du pas dans les méthodes à pas variable

Pour résoudre un problème réaliste, un calcul à pas constant est en général inefficace. On utilise donc des méthodes à pas variable. Mais comment choisir la subdivision ? La méthode que nous allons voir consiste à choisir l'erreur locale de façon à ce qu'elle soit partout égale à une valeur donnée, que l'on nommera  $tol$ . La difficulté consiste à trouver une bonne estimation

de l'erreur locale. La méthode consiste à introduire deux schémas  $(U_n)$  et  $(\hat{U}_n)$  et on utilise alors la différence  $(U_n) - (\hat{U}_n)$  comme estimation de l'erreur locale du moins bon résultat (le résultat d'ordre le plus faible). En effet si on note  $p$  (resp.  $\hat{p}$ ) l'ordre du schéma  $U$  (resp.  $\hat{U}$ ), et si on suppose que  $\hat{p} \leq p$ , on a l'égalité suivante :

$$\|U_n - \hat{U}_n\| \leq \|U_n - u(t_n)\| + \|u(t_n) - \hat{U}_n\| \leq Ch^{p+1} + \hat{C}h^{\hat{p}+1}$$

Donc :

$$\|U_n - \hat{U}_n\| = O(h^{p+1}) + O(h^{\hat{p}+1}) = O(h^{\hat{p}}) \approx Ch^{\hat{p}}$$

D'après la définition de la tolérance, on a alors, en notant  $h_{opt}$  le pas optimal :

$$tol = Ch_{opt}^{\hat{p}}$$

On en déduit une expression de  $h_{opt}$ .

$$h_{opt} = h \left( \frac{tol}{\|U_n - \hat{U}_n\|} \right)^{\frac{1}{\hat{p}+1}}$$

On peut de cette manière définir un algorithme de calcul du schéma, qui prend initialement un  $h$  donné par l'utilisateur puis qui calcule pour chaque itération du schéma  $h_{opt}$ .

### 3 Schéma multipas

Les schémas à un pas calculent la valeur de l'estimation au temps  $t_{n+1}$  uniquement à partir de la valeur au temps  $t_n$ . Les méthodes multipas au contraire se servent systématiquement des résultats obtenus à des temps antérieurs. Nous étudierons dans ce chapitre la théorie à pas constant pour les méthodes multipas. L'idée est de partir de l'égalité

$$u_{n+1} = u_n + \int_{t_n}^{t_{n+1}} f(t, u(t))$$

et d'approximer le terme sous forme intégrale par des interpolations polynomiales. On distingue deux méthodes : la méthode d'Adams et la méthode des différenciations rétrogrades (BCF).

Dans tout ce paragraphe, on se donne une subdivision  $(t_i)_{i \in [1, n]}$  de l'intervalle  $[t_0, T]$  et  $q$  un entier non nul qui comptera le nombre d'approximations dont on va se servir.

### 3.1 Définitions

**Définition 8** (Méthodes multipas). *Une méthode multipas est de la forme*

$$\sum_{j=0}^q \alpha_j U_{n+j} = h \sum_{j=0}^q \beta_j f(t_{n+j}, U_{n+j})$$

avec  $\alpha_q \neq 0$ ,  $|\alpha_0| + |\beta_0| \neq 0$ .

Lorsque  $\beta_q = 0$  on dit que la méthode est explicite. Dans le cas contraire elle est dite implicite (car pour calculer la valeur  $U_{n+q}$ , il faut résoudre un système linéaire de la forme  $U_{n+q} = v_n + \frac{\beta_q}{\alpha_q} f(t_{n+q}, U_{n+q})$ ).

La notion de consistance pour les méthodes multipas que nous abordons est bien plus simple que pour les schémas à un pas. Cela est dû au caractère polynomial de l'approximation choisie.

**Théorème 3.1** (Consistance des méthodes multipas). *Le schéma multipas est d'ordre  $p$  si l'une des conditions équivalentes suivantes est vérifiée :*

1) *Le schéma d'intégration est exact pour tout polynôme de degré inférieur ou égal à  $p$ , soit :*

$$\sum_{j=1}^q \alpha_j = 0$$

$$\forall l \in [1, q], \sum_{j=0}^q j^l \alpha_j - l \sum_{j=0}^q j^{l-1} \beta_j = 0$$

Avec la convention  $0^0 = 1$

2) *Les polynômes  $\rho$  et  $\sigma$  que l'on définit de la manière suivante*

$$\rho(x) = \sum_{j=0}^q \alpha_j x^j \quad \sigma(x) = \sum_{j=0}^q \beta_j x^j$$

*vérifient au voisinage de 0*

$$\rho(e^x) - x\sigma(e^x) = O(x^{p+1})$$

**Théorème 3.2** (Stabilité des méthodes multipas). *Les méthodes multipas sont stables si et seulement si le polynôme  $\rho$  défini dans le théorème ci-dessus a toutes ses racines dans le disque unité, que que ses racines de module 1 sont simples.*

Examinons les exemples de méthodes.

### 3.2 Méthodes D'Adams

**Définition 9** (Méthodes D'Adams). *On définit le polynôme d'interpolation  $P$  de degré au plus  $p-1$  qui vérifie*

$$\forall j \in [n-q+1, n+l], P(t_j) = F_j$$

*Si  $l = 0$  la méthode est explicite, si  $l = 1$  est est implicite*

### 3.3 Méthodes des différenciations rétrogrades ou BCF

**Définition 10** (Méthodes des différenciations rétrogrades ou BCF). *On définit le polynôme d'interpolation  $Q$  de degré inférieur ou égal à  $q$  qui interpole les valeurs de  $U_i$  aux points  $t_i$ . Par conséquent il vérifie*

$$\forall j \in [n - q + 1, n + 1], \quad Q(t_j) = U_j$$

*Pour déterminer la valeur de  $U_{n+1}$ , on impose la relation*

$$Q'(t_{n+1}) = f(t_{n+1}, U_{n+1})$$

*On montre alors que  $U_{n+1}$  vérifie l'équation suivante :*

$$\sum_{j=1}^k \frac{1}{j} \nabla^j U_{n+1} = hf(t_{n+1}, U_{n+1})$$

### Références

- [1] E. HAIRER, S. P. NORSETT, G. WANNER, Solving ordinary differential equations. I. Nonstiff problems., 1993
- [2] M. SCHATZMAN, Analyse numérique, Une approche mathématique, 2001

# Rapport intermédiaire de projet

## 14/12/2005

### Rappel des objectifs

L'objectif de ce projet est de faire découvrir aux élèves le monde de la recherche en mathématiques appliquées (plus précisément, à l'interface avec d'autres sciences), à travers l'étude théorique et la résolution par la méthode du splitting d'équation aux dérivées partielles, qui jouent un rôle important à cette interface, et d'acquérir la méthodologie propre à la démarche scientifique.

En vue d'accomplir l'objectif principal, les élèves devront :

- Assimiler un certain nombre de notions, à savoir, la théorie générale des équations différentielles ordinaires, les méthodes courantes de résolution numérique et l'analyse de ces dernières.
- Se familiariser avec les outils : le langage FORTRAN 77, le solveur d'équations aux dérivées partielles LSODE, etc...
- Savoir modéliser des problèmes, programmer leur résolution numérique, vérifier la cohérence des résultats

### Travail effectué à ce jour

**A ce jour, nous avons terminé la partie préliminaire du projet consistant à se familiariser avec les outils et les notions, et nous abordons le cœur du projet.** Plus précisément, voici ce qui a été effectué :

- Nous avons réalisé une étude bibliographique en nous répartissant les tâches : l'un d'entre nous s'est concentré sur la partie théorique des équations différentielles ordinaires, l'autre sur la partie numérique. Suite à cela, nous avons chacun réalisé un rapport de synthèse (les deux rapports sont fournis en annexe).
- Nous nous sommes familiarisés avec des logiciels comme LaTeX pour l'écriture de rapports, l'environnement et les outils Unix, comme Gnuplot qui sert à visualiser les résultats du code Fortran, etc...
- Nous avons pris en main le code Fortran, fourni par notre encadrant, qui modélise la propagation d'une onde qui se forme en chimie par la compétition d'une réaction chimique avec le phénomène de diffusion, et nous avons adapté un modèle à deux variables (dit BZ) à un modèle à une variable (dit KPP), en vérifiant la cohérence des résultats. Dans ce code, la résolution se fait de manière quasi-exacte, en utilisant le solveur d'équations aux dérivées partielles LSODE.
- Stéphane Descombes nous a fait l'exposé théorique de la méthode du splitting et nous avons pu discuter de son intérêt, et voir sa nécessité car le temps de calcul pour une résolution quasi-exacte est relativement long même sur une machine et avec un compilateur performants.
- Parallèlement, dans le cadre de notre étude sur la recherche à l'interface, nous avons eu amplement le temps de discuter avec notre encadrant, avec Stéphane Descombes et avec Violaine Louvet sur ce qu'est le travail de chercheur à l'interface et les différents problèmes auquel il est confronté.

## **Travail à venir**

- Dans les semaines qui vont suivre, nous allons nous concentrer sur le code, en faisant de nombreux tests sur les modèles BZ et KPP résolus, d'une part dans le cas quasi-exact et d'autre part dans le cas splitté, en essayant de voir où peuvent apparaître des pertes d'ordre.
- Nous étudierons ensuite le passage du code en chimie complexe et verrons en quoi nous pouvons contribuer à cette nouvelle étude.
- En ce qui concerne l'étude théorique, il reste à voir exactement la manière dont la géométrie différentielle intervient, notamment le crochet de Lie, dans l'étude théorique du splitting dans le cas non-linéaire. Ceci nécessite de se familiariser avec des notions relativement abstraites de géométrie différentielle et fera certainement l'objet d'un nouveau rapport.
- Enfin, nous allons chercher à rencontrer d'autres chercheurs ou thésards travaillant à l'interface pour enrichir notre compréhension de ce domaine

## **Conclusion**

En résumé, le projet a beaucoup progressé depuis le début de l'année. Nous entrons dans la partie la plus intéressante qui touche aux travaux actuels de recherche de nos encadrants. Les perspectives sont bonnes puisqu'il est possible, si les travaux en chimie complexe avancent bien et aboutissent à des résultats intéressants, que nous contribuions à l'élaboration d'une présentation de qui s'effectuera à Grenade en avril 2006.

## **Annexes**

Rapport sur la théorie des équations différentielles ordinaires (EDO) (Adrien Auclert)  
Rapport sur les méthodes numériques de résolution des EDO (Miguel Gillot)