UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Mathématiques et Interactions*

Par

Léopold TRÉMANT

Méthodes d'analyse asymptotique et d'approximation numérique

Problèmes d'évolution multi-échelles de type oscillatoire ou dissipatif

Thèse présentée et soutenue à « Lieu », le « date »
Unité de recherche : « voir liste sur le site de votre école doctorale »

Rapporteurs avant soutenance :

Pauline LAFITTE Professeur des universités – CentraleSupélec
Katharina SCHRATZ Professeur des universités – Sorbonne Université

Composition du Jury :

Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition du jury doit être revue pour s'assurer qu'elle est conforme et devra être répercutée sur la couverture de thèse

Président :	Prénom NOM	Fonction et établissement d'exercice *(à préciser après la soutenance)*
Examinateurs :	Prénom NOM	Fonction et établissement d'exercice
	Prénom NOM	Fonction et établissement d'exercice
	Prénom NOM	Fonction et établissement d'exercice
	Prénom NOM	Fonction et établissement d'exercice
Dir. de thèse :	Philippe CHARTIER	Fonction et établissement d'exercice
Co-dir. de thèse :	Mohammed LEMOU	Fonction et établissement d'exercice *(si pertinent)*

Invité(s) :

Prénom NOM Fonction et établissement d'exercice

Acknowledgement

Je tiens à remercier
I would like to thank. my parents..
J'adresse également toute ma reconnaissance à ....
....

Table of Contents

INTRODUCTION

Le développement de modèles mathématiques en sciences naturelles bénéficie d'avancées mathématiques qui permettent de vérifier le caractère bien posé des équations, ou le bon comportement des solutions. Une classe de modèles très prisés depuis quelques dizaines d'années sont les modèles multi-échelles, dont l'étude concerne principalement les modèles double-échelle. Dans ces modèles, on distingue deux dynamiques : une d'échelle caractéristique « rapide » ε et l'autre d'échelle 1. Dans ce manuscrit, on s'intéresse essentiellement à une sous-classe de ces modèles : ceux dont la dynamique rapide est une relaxation. Pour nos résultats, on s'inspire de méthodes développées pour les problèmes dont la dynamique rapide est oscillatoire.

Ces systèmes à relaxation rapide apparaissent en physique dans un cadre fonctionnel de modèles cinétiques [BGK54 ; LM08] ou dans certains systèmes dérivés de problèmes non-linéaires [JX95]. On les observe également en dynamique des populations, e.g. dans [GHM94 ; AP96 ; SAAP00 ; CCS18]. Les systèmes hautement oscillants sont également fréquents en physique. Certains exemples sont présentés dans l'ouvrage de référence [HLW06, Chap. I], comme le modèle de Hénon-Heiles [HH64] dans un contexte de mouvements célestes, ou le problème de Fermi-Pasta-Ulam-Tsingou [For92] en théorie du chaos. On peut aussi étudier certains phénomènes de dynamique quantique non-linéaires, tels que le modèle de Klein-Gordon [BD12], l'équation de Schrödinger [GV11] ou le modèle de Wigner en milieu périodique [CJL17 ; MS11].

La simulation de tels systèmes présente des défis particuliers, qui peuvent se ramener aux concepts de base des méthodes numériques de *stabilité* et de *convergence*. La stabilité consiste à déterminer une condition pour que la solution numérique soit bien définie, qu'elle ne diverge pas. La convergence trace un lien direct entre le coût de calcul et la précision de l'approximation numérique. Dans ce chapitre d'introduction, on introduit mathématiquement les modèles qui nous concernent, et on fait une brève description de leur comportement. À cet égard, on introduit deux exemples de systèmes « jouet » que nous suivrons tout au long du chapitre. Ensuite, on illustre les limitations des méthodes numériques de l'état de l'art, en introduisant certains concepts de convergence liés à la présence du paramètre ε. Enfin, on présente brièvement la contribution de ce travail de

thèse, et on annonce le plan pour la suite du manuscrit.

Introduction mathématique

Ce manuscrit se concentre sur des problèmes de Cauchy de la forme

$$\partial_t u^\varepsilon = -\frac{1}{\varepsilon} A u^\varepsilon + f(u^\varepsilon), \qquad u^\varepsilon(0) = u_0 \tag{1}$$

où t évolue dans l'intervalle $[0, 1]$. On considère ce problème dans un Banach $(E, |\cdot|)$ de dimension finie $d > 0$, avec $A : E \to E$ un opérateur linéaire et $f : E \to E$ un champ de vecteurs régulier. On se concentre sur le cas où A est diagonal avec des valeurs propres positives entières. Souvent, on séparera le problème entre le noyau et l'image de A, pour obtenir un problème de la forme

$$\begin{cases} \partial_t x^\varepsilon = a(x^\varepsilon, z^\varepsilon), & x^\varepsilon(0) = x_0, & \text{(2a)} \\ \partial_t z^\varepsilon = -\frac{1}{\varepsilon}\Lambda z^\varepsilon + b(x^\varepsilon, z^\varepsilon), & z^\varepsilon(0) = z_0 & \text{(2b)} \end{cases}$$

avec $u = \begin{pmatrix} x \\ z \end{pmatrix}$, $A = \begin{pmatrix} 0 & 0 \\ 0 & \Lambda \end{pmatrix}$ et $f = \begin{pmatrix} a \\ b \end{pmatrix}$. En général, on omettra l'exposant ε. Le lecteur peut supposer que, sauf mention contraire, toutes les variables dépendent de ε. D'ailleurs, on peut supposer que le champ de vecteurs f évolue de manière régulière en fonction de ε sans impacter les résultats.

Dans cette section on décrit et illustre le comportement de la solution $(x^\varepsilon, z^\varepsilon)$ à travers deux exemples. En particulier, on énonce le théorème de variété centrale, qui décrit le comportement de la solution en temps long, et on présente rapidement une méthode pour calculer cette variété centrale. On introduit ensuite quelques hypothèses qui permettront de citer des résultats d'estimation numérique rigoureux dans la prochaine section.

Quelques observations

Pour démarrer, considérons le problème jouet suivant

$$\partial_t z(t) = -\frac{1}{\varepsilon} z(t) + \sin(t), \qquad z(0) = 1, \tag{3}$$

qui peut être transformé en un problème de la forme (2) en posant $x(t) = t$, soit $\partial_t x = 1,\ x(0) = 0$. Ce problème peut être obtenu à partir de

$$\partial_t y(t) = -\frac{1}{\varepsilon}\Big(y(t) - \cos(t)\Big), \qquad y(0) = 0,$$

en posant $z(t) = y(t) - \cos(t)$. C'est un problème de référence pour l'introduction aux systèmes raides : c'est le premier exemple présenté dans [HW96]. La solution exacte se calcule sans difficulté en intégrant $\partial_t\Big[e^{t/\varepsilon} z(t)\Big]$, ce qui donne

$$z(t) = e^{-t/\varepsilon}\left(1 + \frac{\varepsilon^2}{1+\varepsilon^2}\right) + \frac{\varepsilon}{1+\varepsilon^2}\Big(\sin(t) - \varepsilon\cos(t)\Big). \tag{4}$$

On observe que la solution comporte deux parties de natures différentes, la phase transitoire (en $e^{-t/\varepsilon}$) et la variété centrale (en t) de taille ε. Ces deux phases apparaissent clairement sur la figure ci-dessous où on a tracé la solution et la variété centrale associée pour trois valeurs de ε. En effet, le temps d'atteinte de la variété semble proportionnel à ε pour des petites valeurs.

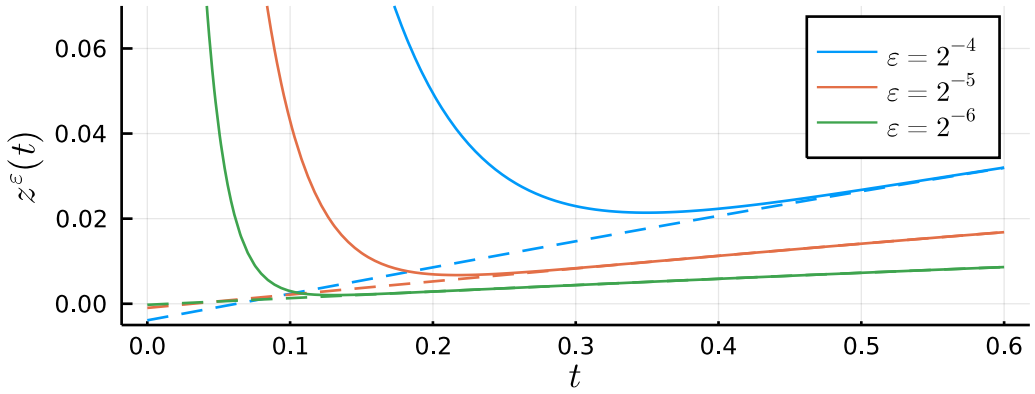


FIGURE 1 – Comportement de la solution (4) pour différentes valeurs de ε, avec chaque variété centrale associée en pointillés.

En fait, dans le cas d'un système de la forme (2), la dynamique en temps long est déterminé entièrement par la variable x. Il y a une réduction de dimension, qui est traduite dans le théorème de variété centrale.

Théorème (Variété centrale, [Car82])**.** *En supposant que le champ f est dans $C^\infty(E)$ et que la donnée initiale $u_0 \in E$ du problème* (1) *est bornée, il existe un temps final $T > 0$ et*

un paramètre limite ε_0 tels que le problème (1) est bien posé sur $[0,T]$. En outre, il existe une application régulière $x \mapsto z = \varepsilon h^\varepsilon(x)$, un taux $\mu > 0$ et une constante $c > 0$ tels que pour tout $t \in [0,T]$,

$$\Big|z(t) - \varepsilon h^\varepsilon\Big(x(t)\Big)\Big| \leq ce^{-\mu t/\varepsilon}.$$

En outre, quitte à modifier c, il existe une donnée initiale « de l'ombre » x_0^ε telle que

$$\Big|x(t) - \varphi_t^\varepsilon\Big(x_0^\varepsilon\Big)\Big| \leq c\varepsilon e^{-\mu t/\varepsilon},$$

où φ_t^ε est le t-flot associé au champ de vecteurs $x \mapsto a\Big(x, \varepsilon h^\varepsilon(x)\Big)$. On appelle l'ensemble des $\Big(x, \varepsilon h^\varepsilon(x)\Big)$ la variété centrale, qui est stable et « attire » la solution.

Par exemple dans le cas (3), le morphisme de variété centrale $\varepsilon h^\varepsilon$ est donné par

$$h^\varepsilon(x) = \frac{1}{1+\varepsilon^2}\Big(\sin(x) - \varepsilon\cos(x)\Big). \tag{5}$$

Il est impossible d'espérer calculer le morphisme de variété centrale explicitement en général. On peut néanmoins appliquer une méthode de point fixe en remarquant qu'en temps long, $z \approx \varepsilon h^\varepsilon(x)$, d'où

$$\partial_t z \approx \varepsilon \partial_x h^\varepsilon(x) \cdot \partial_t x \approx \varepsilon \partial_x h^\varepsilon(x) \cdot a(x, \varepsilon h^\varepsilon(x)).$$

Ainsi on peut poser $h^{[0]} = b(x,0)$ et itérer de manière explicite

$$\varepsilon \partial_x h^{[n]}(x) \cdot a(x, \varepsilon h^{[n]}(x)) = -h^{[n+1]}(x) + b(x, \varepsilon h^{[n]}(x)). \tag{6}$$

Le calcul de la donnée initiale de l'ombre x_0^ε n'est en revanche pas si simple. La calculer demande de faire des développements sur l'intégralité du système et demande un certain investissement. Dans [CCS16], les auteurs font appel à des B-séries [1] pour obtenir un modèle asymptotique sur l'intégralité du problème (1), à partir duquel ils trouvent en particulier cette donnée initiale modifiée, mais les calculs sont bien plus compliqués que (6).

Remarque. *Le phénomène de donnée initiale de l'ombre n'est cependant pas visible sur*

1. Les B-séries sont des séries formelles qui décrivent les solutions d'EDO. Cet outil, souvent utilisé pour développer des schémas numériques, est présenté en détails dans [HLW06, Chap. III] ou de manière plus concise dans [CHV10]. Malgré une apparence encombrante, les B-séries possèdent une structure algébrique élégante basée sur des opérations sur les arbres.

cet exemple, puisque la variable x ne dépend pas de la dynamique sur z (pour rappel, on a $x(t) = t$). L'interdépendance entre x et z apparaîtra dans le cas test de la section suivante.

On remarque deux caractéristiques importantes de la méthode d'approximation de h^ε : elle nécessite de calculer une dérivée, ce qui demande du calcul symbolique potentiellement coûteux, et la convergence de la méthode n'est pas assurée. Même dans l'exemple jouet (3), cette méthode génère

$$h^{[n]}(x) = R_{n+1}(\varepsilon)\sin(x) + \varepsilon R_n(\varepsilon)\cos(x)$$

avec $R_n(\varepsilon) = \sum_{k=0}^{\lfloor n/2 \rfloor} \varepsilon^{2k}$ et par convention $R_{-1} = 0$. En d'autres termes, on construit le développement en série entière en ε de la partie lente dans (4), i.e. $h^{[0]}(x) = \sin(x)$, $h^{[1]}(x) = \sin(x) + \varepsilon\cos(x)$ etc. Ainsi, le développement n'est convergent que pour $\varepsilon < 1$, alors que la solution $z(t)$ converge vers la variété $\varepsilon h^\varepsilon(t)$ définie par (5) pour tout $\varepsilon > 0$.

Cette limitation apparaîtra couramment au cours de ce manuscrit sous le format plus contraignant

$$\varepsilon \leq \frac{\varepsilon_0}{n+1}. \tag{7}$$

Justifions cette borne à l'aide d'un exemple. On applique les itérations (6) avec $a = 1$ et $b(x, z) = (1 + x)^{-1}$, ce qui engendre

$$h^{[n]}(x) = (1+x)^{-1} \sum_{k=0}^{n} k! \left(\frac{\varepsilon}{(1+x)}\right)^k. \tag{8}$$

On cherche maintenant une condition sur ε telle que $h^{[n]}(1) \leq 1$. Pour $n \geq 1$, en partant de (6) on obtient

$$|h^{[n]}(1)| \leq \frac{1}{2} + \varepsilon \left|\partial_x h^{[n-1]}(1)\right|.$$

Pour borner la dérivée, on remarque que $x \mapsto h^{[n-1]}(x)$ est analytique autour de $x = 1$, et la série entière converge avec un rayon 2. Ainsi pour $0 < \delta < 2$, on a par formule de Cauchy

$$\partial_x h^{[n-1]}(1) = \frac{1}{2\pi} \int_{|\xi|=\delta} \frac{h^{[n-1]}(1+\xi)}{\xi^2} \mathrm{d}\xi,$$

d'où par majoration et grâce à l'expression de $h^{[n-1]}(x)$,

$$|\partial_x h^{[n-1]}(1)| \leq \frac{1}{\delta} \sup_{|\xi|\leq\delta} |h^{[n-1]}(1+\xi)| \leq \frac{h^{[n-1]}(1-\delta)}{\delta}.$$

On peut alors injecter cette inégalité dans (8) pour obtenir de proche en proche

$$|h^{[n]}(1)| \leq \sum_{k=0}^{n} \frac{1}{2-k\delta} \left(\frac{\varepsilon}{\delta}\right)^k.$$

Pour que tous les termes $(2-k\delta)^{-1}$ soient bien définis, on choisit alors $\delta = 1/n$ et on applique une inégalité de Hölder dans la somme pour majorer $|h^{[n]}(1)| \leq (1-n\varepsilon)^{-1}$. Enfin, on trouve que cette grandeur est positive et inférieure à 1 si et seulement si

$$\varepsilon \leq \frac{1}{n}.$$

Ce même raisonnement peut être tenu dans le cas général si a et b sont analytiques de même rayon de convergence, d'où la borne (7), valide aussi pour $n = 0$.

Définitions et hypothèses

On considère le problème (1) en dimension finie d, et on suppose que la donnée initiale u_0 appartient à un ensemble $\mathcal{U}_0$ bien choisi de sorte qu'en tout temps, la solution appartient à un ensemble borné $\mathcal{K}$ de la forme

$$\mathcal{K} := \left\{ \begin{pmatrix} x \\ z \end{pmatrix} \in E, \text{ s.t. } x \in \mathcal{K}_x, |z| \leq \rho \right\}$$

avec $\mathcal{K}_x$ borné et $\rho > 0$. Grâce au théorème de variété centrale, ces hypothèses peuvent être considérées valides pour tout $\varepsilon \in (0, \varepsilon_0]$ pour un certain $\varepsilon_0 > 0$ fixé.

En outre, on suppose que le champ de vecteurs $u \mapsto f(u)$ est de classe C^∞ sur $\mathcal{K}_R$ pour un certain rayon $R > 0$, où

$$\mathcal{K}_R = \left\{ u \in \mathbb{R}^d, \ d(u, \mathcal{K}) \leq R \right\}$$

avec $d(u, \mathcal{K}) = \inf_{v \in \mathcal{K}} |u - v|$ la distance de u à $\mathcal{K}$. Cette hypothèse permet d'assurer que les méthodes numériques se comportent bien au voisinage de la solution, sans nécessiter d'être exact pour autant.

Paradigme numérique

En général, on suppose $\varepsilon \ll 1$, et donc le système (1) comporte une dynamique *rapide* par rapport au temps d'étude. À cet égard, des méthodes d'*analyse asymptotique* ont été développées, c'est-à-dire des méthodes qui permettent de caractériser le système dans cette limite ε « petit », en général en découplant ces deux dynamiques. Pour les problèmes hautement oscillants, trois exemples particulièrement célèbres sont les méthodes d'homogénéisation [GM03], de moyennisation [Per69 ; SVM07 ; LM88] et de formes normales [Mur06 ; Bam03]. Pour les problèmes à relaxation rapide, la littérature est moins fournie. Qu'il s'agisse de calculer la variété centrale comme précédemment ou d'un développement de Chapman-Enskog [SBD86 ; Deg04 ; CCLM15], la phase transitoire n'est pas calculée.

Plus récemment dans [CCS16], les auteurs capturent aussi la phase transitoire, mais la méthode est formelle. Dans cette section, on étudie l'application de méthodes numériques « standards » de l'état de l'art, et on observe le comportement de l'erreur numérique non seulement en fonction du pas de temps Δt, mais aussi en fonction du paramètre ε.

On commence par décrire ce qu'on entend par « méthode numérique » et le contexte dans lequel on va les étudier. Ensuite, on présente trois méthodes d'ordre 2 de l'état de l'art : le splitting de Strang, un schéma IMEX-BDF et une méthode de Runge-Kutta exponentielle.

Mise en place de la résolution

Pour étudier le comportement des schémas numériques sur les problèmes de la forme (1), on considère l'exemple jouet suivant

$$\begin{cases} \partial_t v_1 = v_2, & v_1(0) = 1, \quad (9a) \\ \partial_t v_2 = -\dfrac{1}{\varepsilon}\big(v_1 + v_2\big), & v_2(0) = 0. \quad (9b) \end{cases}$$

Cet exemple ressemble à certains problèmes hyperboliques avec relaxation, et sa linéarité le rend simple à étudier. Il prend facilement la forme (2) en posant par exemple $x = v_1$

et $z = v_1 + v_2$, ce qui donne

$$\begin{cases} \partial_t x = -x + z, & x(0) = 1, \quad (10a) \\ \partial_t z = -\dfrac{1}{\varepsilon} z - x + z, & z(0) = 1. \quad (10b) \end{cases}$$

Ce problème est linéaire et se diagonalise sans difficulté pour $\varepsilon < 1/4$, ce qui génère

$$\tilde{u} = \underbrace{\begin{pmatrix} -1 & 1 - r_\varepsilon \\ \varepsilon & 1 - \varepsilon - \varepsilon r_\varepsilon \end{pmatrix}}_{P} \begin{pmatrix} x \\ z \end{pmatrix}, \qquad \text{tel que} \qquad \partial_t \tilde{u} = \begin{pmatrix} -r_\varepsilon & 0 \\ 0 & -\frac{1}{\varepsilon} + r_\varepsilon \end{pmatrix} \tilde{u}$$

avec $r_\varepsilon = \frac{1}{2\varepsilon}\left(1 - \sqrt{1 - 4\varepsilon}\right)$. On obtient ainsi une expression explicite pour $u = \begin{pmatrix} x \\ z \end{pmatrix}$ de la forme

$$u(t) = P^{-1} \begin{pmatrix} e^{-tr_\varepsilon} & 0 \\ 0 & e^{-t/\varepsilon} e^{tr_\varepsilon} \end{pmatrix} P u(0)$$

où $P^{-1} = \frac{1}{\sqrt{1-4\varepsilon}} \begin{pmatrix} -1 + \varepsilon + \varepsilon r_\varepsilon & 1 - r_\varepsilon \\ \varepsilon & 1 \end{pmatrix}$.

Remarque. *On voit bien dans la définition de r_ε que le problème change de nature entre $\varepsilon \leq 1/4$ et $\varepsilon > 1/4$. En effet, dans le premier cas le système est purement dissipatif, alors que dans le second, des oscillations apparaissent. Cette singularité apparaît également dans la matrice de changement de variable, dont le déterminant vaut $-\sqrt{1-4\varepsilon}$.*

En pratique, on ne sait pas résoudre tous les systèmes de la forme (1) de manière exacte. On va donc appliquer des méthodes d'*approximation numérique* pour calculer une solution approchée. Plus spécifiquement, on va implémenter certains *schémas numériques* et observer la qualité d'approximation sur l'exemple (10). Définissons ce qu'on entend par « schéma numérique ».

On démarre par considérer une discrétisation de l'intervalle temporel $[0, T]$, c'est-à-dire qu'au lieu de considérer cet objet comme continu, on le considère comme une suite de $N + 1$ points $(t_n)_{0 \leq n \leq N}$ avec $N \geq 1$. On choisit de se restreindre à une discrétisation *uniforme*, c'est-à-dire que l'intervalle de temps $[0, T]$ est divisé en N intervalles de taille égale notée Δt.

De manière équivalente, les points de séparation $(t_n)_{0\leq n\leq N}$ sont donnés par

$$t_{n+1} = t_n + \Delta t \qquad \text{avec} \qquad t_0 = 0 \quad \text{et} \quad \Delta t = \frac{T}{N},$$

ou encore $t_n = \frac{n}{N}T$. À cette discrétisation, on peut associer une *approximation* $(u_n)_{0\leq n\leq N}$ telle que $u_n \approx u(t_n)$. On peut voir un exemple d'une telle approximation en Figure 2, où les points carrés sont une approximation [2] des points ronds.

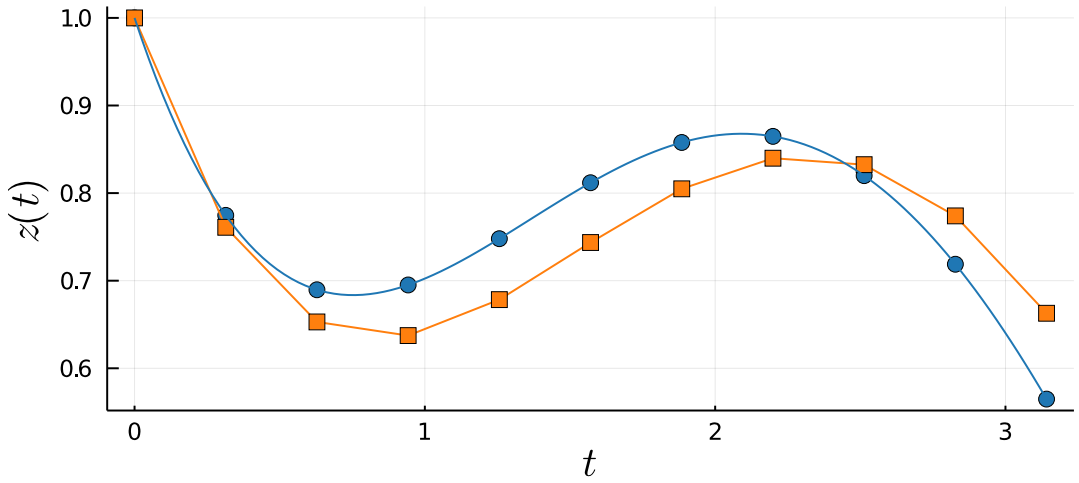


FIGURE 2 – Tracé de la solution exacte (en bleu, marqueurs ronds) au problème (3) avec $\varepsilon = 1$ et d'une approximation (orange, marqueurs carrés) sur une discrétisation uniforme de $[0, \pi]$ à 11 points.

Définition. *On définit l'erreur d'une approximation comme l'erreur maximale sur les points d'interpolation, i.e. étant donnée une discrétisation à $N+1$ points $(t_n)_{0\leq n\leq N}$ et une approximation (u_n) d'une fonction $t \mapsto u(t)$, l'erreur est définie par la formule*

$$\mathrm{err} = \max_{0\leq n\leq N} \left|u_n - u(t_n)\right|. \tag{11}$$

Si en outre la solution et l'approximation dépendent du paramètre $\varepsilon \in (0, \varepsilon_0]$, on écrit $\mathrm{err}(\varepsilon)$ *l'erreur d'approximation, et on définit l'erreur uniforme* $\overline{\mathrm{err}}$ *par*

$$\overline{\mathrm{err}} = \sup_{\varepsilon\in(0,\varepsilon_0]} \mathrm{err}(\varepsilon). \tag{12}$$

2. Cette approximation est obtenue par itération en posant $z_{n+1} = z_n - \dfrac{\Delta t}{\varepsilon} z_{n+1} + \Delta t \sin(t_n)$ avec $z_0 = z(0) = 1$, ce qui correspond à une méthode appelée IMEX-BDF1.

Ces deux erreurs peuvent avoir des comportements différents.

Parfois, l'erreur est définie comme l'erreur au temps final, ce qui peut paraître moins contraignant mais a peu d'influence en pratique car les estimations d'erreur sont croissantes avec l'indice n. Ainsi ces deux définitions coïncident pour les résultats théoriques.

Remarque. *À partir d'une approximation (u_n), on peut obtenir une approximation sur une discrétisation plus fine par interpolation (typiquement avec des splines cubiques ou de manière linéaire comme en Figure 2). Il semble alors naturel de se demander s'il est possible de réduire le coût de calcul en calculant une approximation sur une discrétisation grossière pour l'interpoler ensuite vers une discrétisation plus fine. Néanmoins, si l'erreur telle que définie en* (11) *est mauvaise, on ne peut pas espérer l'améliorer par interpolation. C'est pourquoi ce manuscrit se concentre sur cette approche « simple » de l'erreur.*

Lorsqu'il s'agit de trouver une approximation à une solution d'équation différentielle, il semble naturel de procéder de manière itérative : les seules données accessibles sont la condition initiale $u(0)$ et le champ de vecteurs suivi par la solution, donc il faut trouver un moyen de combiner ces deux informations pour obtenir une approximation $u_1 \approx u(\Delta t)$. Une fois cette information obtenue, on peut l'utiliser avec les autres pour calculer $u_2 \approx u(2\Delta t)$, etc. Dans les faits, on se limite à un nombre fixe $s \geq 1$ de points pour extrapoler le suivant. On parle alors de méthode multipoints. Les méthodes à un point sont appelées méthodes de Runge-Kutta.

La méthode utilisée pour calculer un terme à partir des précédents est appelé un schéma numérique $\Phi^{\varepsilon}_{\Delta t}$, et elle peut s'écrire

$$u_{n+s} = u_{n+s-1} + \Delta t\, \Phi^{\varepsilon}_{\Delta t}(u_{n+s-1}, \dots, u_n).$$

Cette notation peut être pratique pour étudier le schéma, mais il faut garder à l'esprit qu'elle peut camoufler de nombreuses difficultés. Par exemple, le schéma d'Euler implicite s'écrit

$$u_{n+1} = u_n + \Delta t \left(-\frac{1}{\varepsilon} A u_{n+1} + f(u_{n+1}) \right).$$

Pour trouver $\Phi^{\varepsilon}_{\Delta t}$, il faut inverser cette relation ; on parle de schéma *implicite*, par opposition aux schémas *explicites*. Cette inversion peut s'avérer particulièrement coûteuse si f présente des non-linéarités et si le système est grand. Ainsi par la suite on se limite à des schémas *explicites* en f, et on fait l'hypothèse suivante :

Hypothèse. *On sait calculer $t \mapsto e^{-tA}$ et $t \mapsto (\mathrm{id} + tA)^{-1}$ de manière exacte.*

Calculer le semi-groupe $t \mapsto e^{-tA}$ peut sembler contraignant, mais rappelons nous qu'on se restreint ici au cas où A est une matrice diagonale.

Définition. *On dit qu'un schéma numérique $\Phi_{\Delta t}$ est d'ordre q s'il existe une constante $C > 0$ et un pas de temps maximal $\Delta t_0 > 0$ tel que pour toute subdivision de pas de temps $\Delta t \leq \Delta t_0$, l'erreur de schéma est bornée par $C\Delta t^q$.*

Dans le contexte d'un schéma qui dépend du paramètre ε, la constante d'erreur C et le pas de temps maximal Δt_0 dépendent généralement de ε, ainsi il faut distinguer l'ordre du schéma (calculé pour $\Delta t \ll \varepsilon$) de l'ordre de « convergence uniforme » du schéma, qui fait disparaître la dépendance en ε. Cette distinction sera clarifiée par les exemples à venir.

Méthodes numériques

On présente les résultats associés à trois méthodes d'ordre 2, qui traitent la partie raide différemment de la partie non-raide. Les méthodes sont bien définies dans la limite $\varepsilon \to 0$, et on s'intéresse au comportement de l'erreur en fonction de Δt *et* de ε.

On ne considèrera pas de méthodes complètement implicites parce qu'elles sont très coûteuses notamment dans un contexte d'EDP. Néanmoins la convergence de ces méthodes est souvent excellente et des implémentations très efficaces existent. La référence sur le sujet est [HW96], et toutes les bonnes boîtes à outils de résolution d'EDO contiennent la méthode RadauII.[3] On ne considère pas non plus de méthodes purement explicites demandant $\Delta t < \varepsilon$, ce qui est beaucoup trop coûteux en pratique.

Il est important de noter que cette introduction ne présente pas une étude des schémas présentés. Il s'agit d'une compilation non-exhaustive de résultats et d'observations sur les propriétés de ces méthodes appliquées à (1) afin de contextualiser les contributions du manuscrit par la suite. Néanmoins, les schémas sont présentés plus en détails dans l'Annexe B.

Splitting de Strang

Une approche courante consiste à séparer le problème (1) en deux parties, l'une raide

3. Attention cependant, la plupart de ces boîtes à outils masquent la difficulté associée aux méthodes implicites, à savoir la résolution d'un système et l'erreur associée. En outre, il est parfois difficile de désactiver le pas de temps adaptatif, ce qui est problématique pour une étude de convergence.

et l'autre non-raide. La manière naturelle de procéder fournit

$$\begin{cases} \partial_t u^{(1)} = -\dfrac{1}{\varepsilon} A u^{(1)}, \\ \partial_t u^{(2)} = f(u^{(2)}). \end{cases}$$

On note φ_t, $\varphi_t^{(1)}$ et $\varphi_t^{(2)}$ les t-flots associés aux problèmes en u, $u^{(1)}$ et $u^{(2)}$ espectivement. On remarque qu'il est simple de calculer $\varphi^{(1)}$ de manière exacte, et simple de calculer $\varphi^{(2)}$ de manière numérique. Cependant, ces deux dynamiques sont mélangées dans φ, ce qui rend le flot du problème d'origine difficile à calculer. Ainsi, on est en droit de se poser la question : est-il possible d'obtenir φ à partir de $\varphi^{(1)}$ et de $\varphi^{(2)}$?

La réponse est négative en général, mais on peut *approcher* φ à partir de $\varphi^{(1)}$ et de $\varphi^{(2)}$ à l'aide de compositions successives. C'est cette approche qu'on appelle *splitting*. Le plus couramment utilisé est le splitting de Strang, qui s'écrit

$$\varphi_t = \varphi_{t/2}^{(1)} \circ \varphi_t^{(2)} \circ \varphi_{t/2}^{(1)} + \mathcal{O}(t^3)$$

avec la constante d'erreur dans $\mathcal{O}(t^3)$ qui dépend du paramètre ε de manière raide. Pour la plupart des équations, l'ordre des opérations n'a pas d'importance, mais lorsque le système présente une partie de relaxation raide comme ici, il a été remarqué dans [Spo00 ; DM04] qu'il vaut mieux « terminer » par la relaxation.

Notons que le splitting de Strang peut être obtenu par « symétrisation » du splitting de Lie $\varphi_t^{(2)} \circ \varphi_t^{(1)}$, d'ordre 1. Le splitting est exact si et seulement si les champs A et f commutent, c'est-à-dire si on vérifie l'identité

$$Af - \partial_u f \cdot A = [A, f] = 0.$$

Dans ce cas, le splitting de Lie génère un flot qui coïncide avec φ. Évidemment, ce n'est pas le cas en général. En particulier dans le cas test (10), on a $[A, f] = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$, donc on s'attend à avoir une erreur qui décroit comme Δt^2 lors de la simulation. Cependant, on observe un résultat différent en Figure 3.

Dans cette figure, on observe que le comportement de la solution est celui attendu pour $\Delta t \ll \varepsilon$. Néanmoins, lorsqu'on trace l'erreur en fonction de ε, on voit qu'à Δt fixé, il y a toujours un seuil à partir duquel une réduction de ε entraîne une augmentation

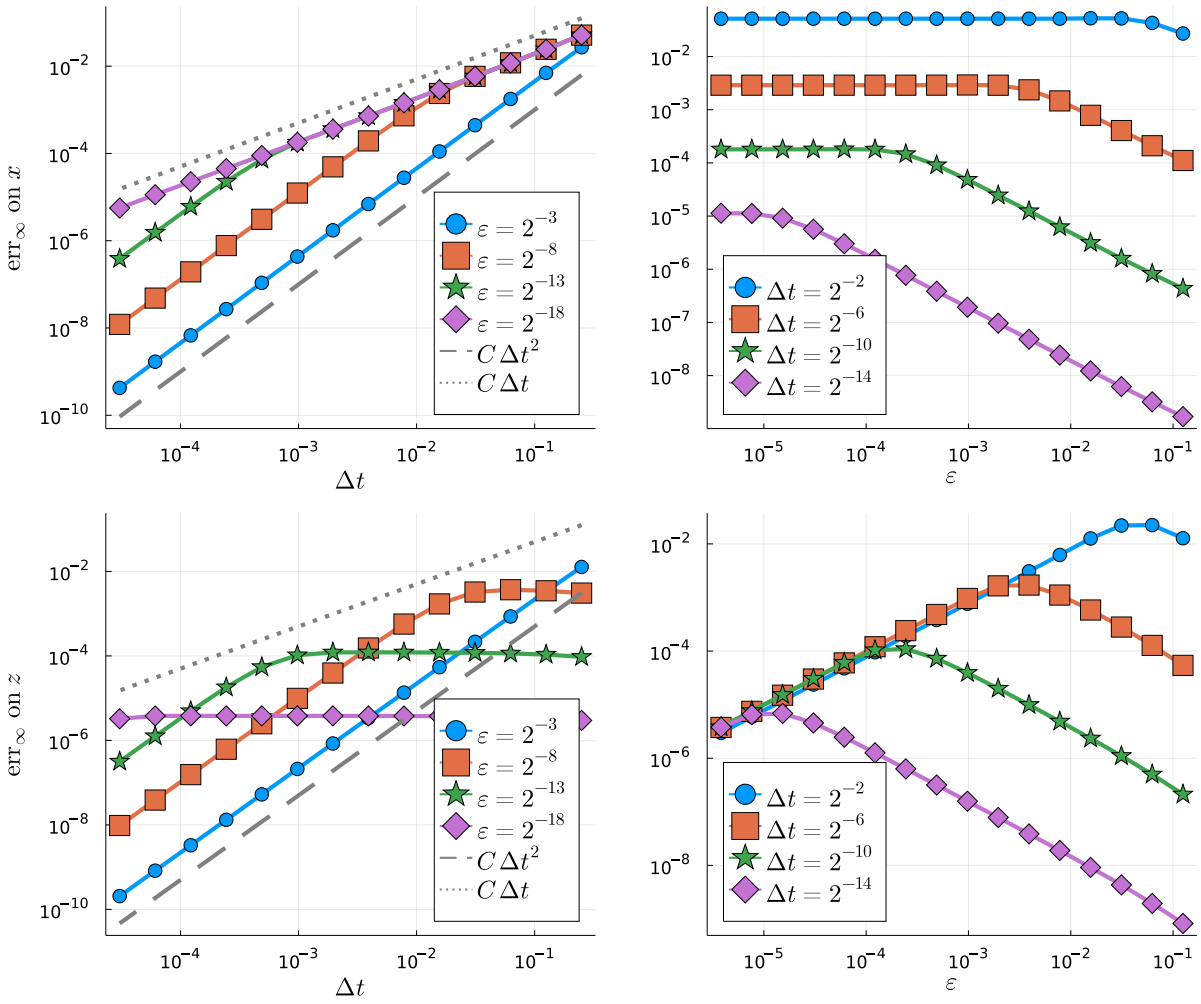


FIGURE 3 – Erreurs sur x (en haut) et z (en bas) en fonction de Δt (à gauche) et de ε (à droite) avec la méthode de Strang. Sur les tracés de l'erreur en fonction de Δt, les convergences théorique et uniforme sont tracées.

de l'erreur. Cette augmentation entraîne une *réduction d'ordre*, c'est-à-dire qu'on ne peut pas avoir

$$\sup_{0<\varepsilon\leq\varepsilon_0} \text{err} \leq C\Delta t^q,$$

avec $q = 2$, ce n'est possible qu'avec $q = 1$. Cette réduction d'ordre a été notée dans différents contextes [Spo00 ; FOS15], et différentes méthodes ont été développées pour dépasser cette limite [EO15 ; CR17 ; BV20], même si la méthode reste au plus d'ordre 2.

Méthodes exponentielles Runge-Kutta (expRK)

Ces méthodes exponentielles [4] proviennent de la formulation intégrale du problème (1),

$$u(t) = e^{-\frac{t}{\varepsilon}A}u_0 + \int_0^t e^{-\frac{t-\tau}{\varepsilon}A} f(u(s))\mathrm{d}s.$$

La partie du semi-groupe est gardée exacte tandis que la partie (possiblement) non-linéaire est approchée, ce qui conduit au schéma d'ordre 1 suivant,

$$u_{n+1} = e^{-\Delta t A/\varepsilon}u_n + \left(\int_0^{\Delta t} e^{(\tau-\Delta t)A/\varepsilon}\mathrm{d}\tau\right) f(u_n). \tag{13}$$

Pour passer à l'ordre supérieur, les parties raide et non-raide sont liées, donc l'approche demande plus de subtilité, mais on peut obtenir des méthodes exponentielles Runge-Kutta (i.e. des méthodes à un pas, pour obtenir $u_{n+1} \approx u(t_{n+1})$ à partir de seulement $u_n \approx u(t_n)$) d'ordre arbitraire. Une grande classe de schémas de ce type est compilée dans [HO05], et dans un autre article les mêmes auteurs obtiennent une convergence théorique.

Estimations d'erreur (Hochbruck, Ostermann - [HO04])

Avec un schéma expRK d'ordre $q \geq 1$, l'erreur vérifie la borne à une constante multiplicative près

$$\left|\left(I + \frac{1}{\varepsilon}A\right)\left(u_n - u(t_n)\right)\right| \leq \Delta t^q \left(\sup_{0\leq t\leq t_n}\left|\partial_t^{q-1}\mathcal{U}(t)\right| + \int_0^{t_n}\left|\partial_t^q\mathcal{U}(t)\right|\mathrm{d}t\right)$$

où $\mathcal{U} = \partial_t u + \frac{1}{\varepsilon}Au$.

Ce résultat est généralement évoqué dans un contexte fonctionnel de problème parabolique, mais cette version simplifiée est suffisante ici. Un intérêt remarquble de ces méthodes

4. À ne pas confondre avec les méthodes de Lawson (voir [Law67 ; HLO20]) qui procèdent en applicant des méthodes de Runge-Kutta standards sur la variable filtrée $v(t) = e^{tA/\varepsilon}u(t)$, puis en multipliant le résultat par $e^{-tA/\varepsilon}$. Les résultats théoriques sur la convergence de ces dernières sont encore très récents.

est que dans l'erreur, la composante z est renormalisée par ε. Dans notre cas, cela permet d'obtenir une sorte d'erreur relative puisque $z(t) = \varepsilon h^{\varepsilon}(x(t)) + \mathcal{O}(e^{-t/\varepsilon})$. D'ailleurs, le schéma (13) (parfois appelé « Euler exponentiel ») avait été proposé dans [VS98] pour obtenir une meilleure convergence que le splitting de Strang en erreur relative.

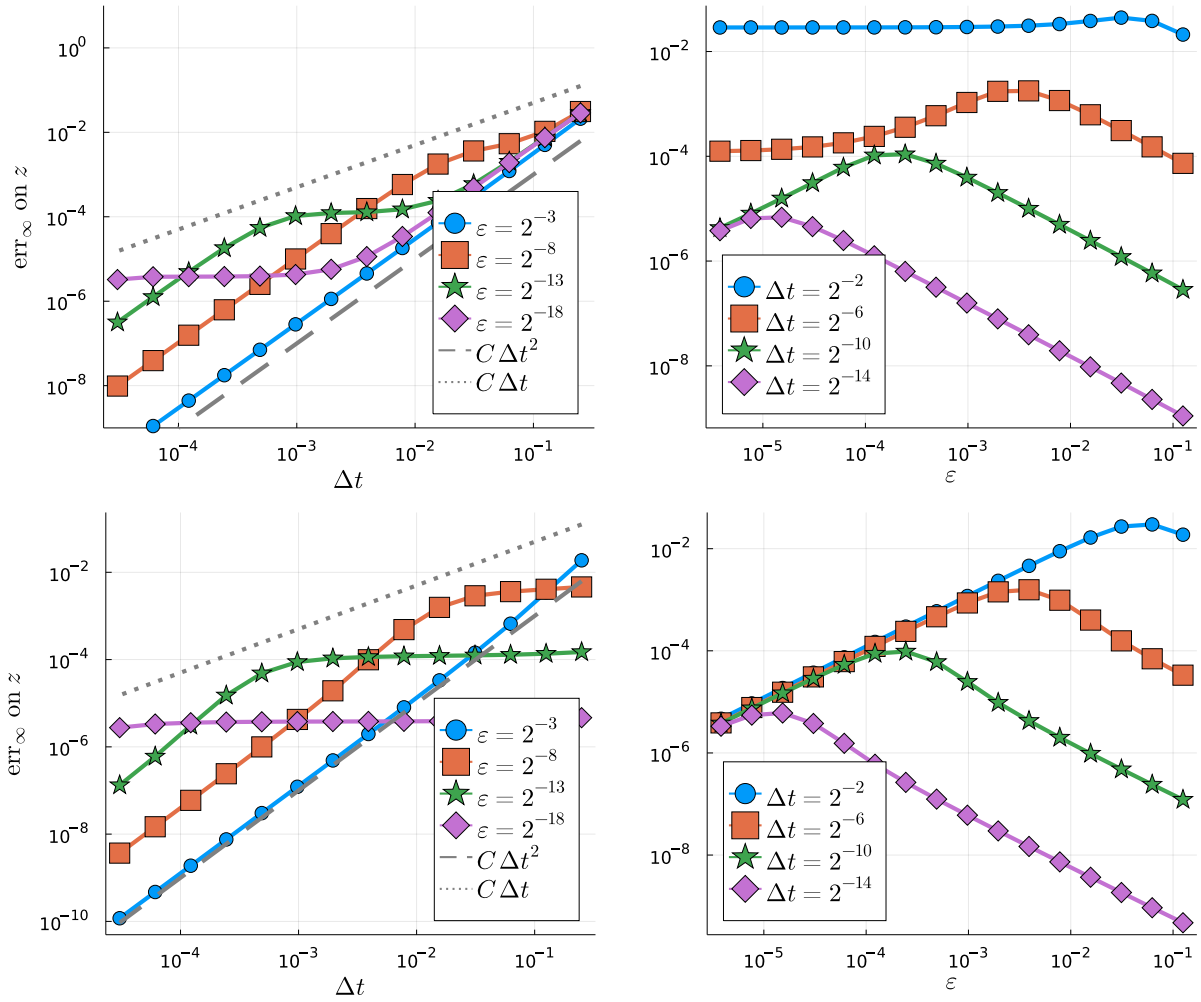


FIGURE 4 – Erreurs sur x (en haut) et z (en bas) en fonction de Δt (à gauche) et de ε (à droite) avec le schéma exponentiel RK2. Sur les tracés de l'erreur en fonction de Δt, les convergences théorique et uniforme sont tracées.

Cette renormalisation par ε de la composante z se voit nettement en Figure 4 sur l'erreur en fonction de ε est bornée par une fonction de la forme $C\varepsilon$. Sur x, à ε fixé, on observe trois phases dans les variations d'erreur en fonction de Δt :

— une décroissance en Δt^2 ;

— un plateau à partir de $\Delta t^2 \approx \varepsilon$ jusqu'à $\Delta t \approx \varepsilon$;

— de nouveau une décroissance en Δt^2.

Cette deuxième phase engendre une réduction d'ordre, où l'ordre « uniforme » est 1, et la troisième phase présente l'ordre « naturel » du schéma. Pour la composante z, il n'y a pas de première phase, mais la réduction d'ordre est la même. Malgré cela, la convergence est meilleure que pour le splitting de Strang : dans le paradigme $\varepsilon \to 0$, seule la première phase est observée et l'erreur décroit comme Δt^2. On dit que le schéma « préserve l'asymptote ». Cette description ne permet pas de décrire le comportement de l'erreur dans les phases 1 et 3, mais pas dans la phase 2.

Méthode IMEX-BDF

L'inconvénient des méthodes exponentielles est qu'elles demandent une intégration très précise du semi-groupe $t \mapsto e^{-tA/\varepsilon}$. On peut considérer des méthodes moins coûteuses, qui demandent seulement d'inverser un système linéaire. C'est le cas par exemple des méthodes implicite-explicites (IMEX), où la partie raide est implicite et la partie non-linéaire est explicite. Ainsi la méthode d'Euler implicite-explicite appliquée à (1) s'écrit

$$\frac{u_{n+1} - u_n}{\Delta t} = -\frac{1}{\varepsilon} A u_{n+1} + f(u_n).$$

On se restreint ici aux méthodes multipas dénommées IMEX-BDF (*backwards differentiation formula*), initialement développées dans [Cro80], puis dans [ARW95 ; ACM99 ; HR07 ; DP17]. On a là aussi une erreur théorique.

Estimations d'erreur (Crouzeix - [Cro80])

Avec une méthode IMEX-BDF d'ordre q à s pas, si on suppose qu'on a pu obtenir les $(s-1)$-ièmes premiers pas avec une manière quelconque, l'erreur sur les pas suivants peut être bornée par

$$\left|u_n - u(t_n)\right| \leq \sum_{i=0}^{s-1} \left|u_i - u(t_i)\right| + \Delta t^q \int_0^{t_n} \left(\left|\partial_t^{q+1} u(t)\right| + \frac{1}{\varepsilon} \left|A \partial_t^q u(t)\right| \right) \mathrm{d}t$$

à une constante multiplicative près.

La différence principale de cette erreur avec celle des méthodes expRK est que la composante z n'est pas « normalisée ». On voit ainsi en Figure 5 que l'erreur uniforme sur z dégénère à l'ordre *zéro*. Il est même possible d'augmenter l'erreur en diminuant le pas de temps Δt.

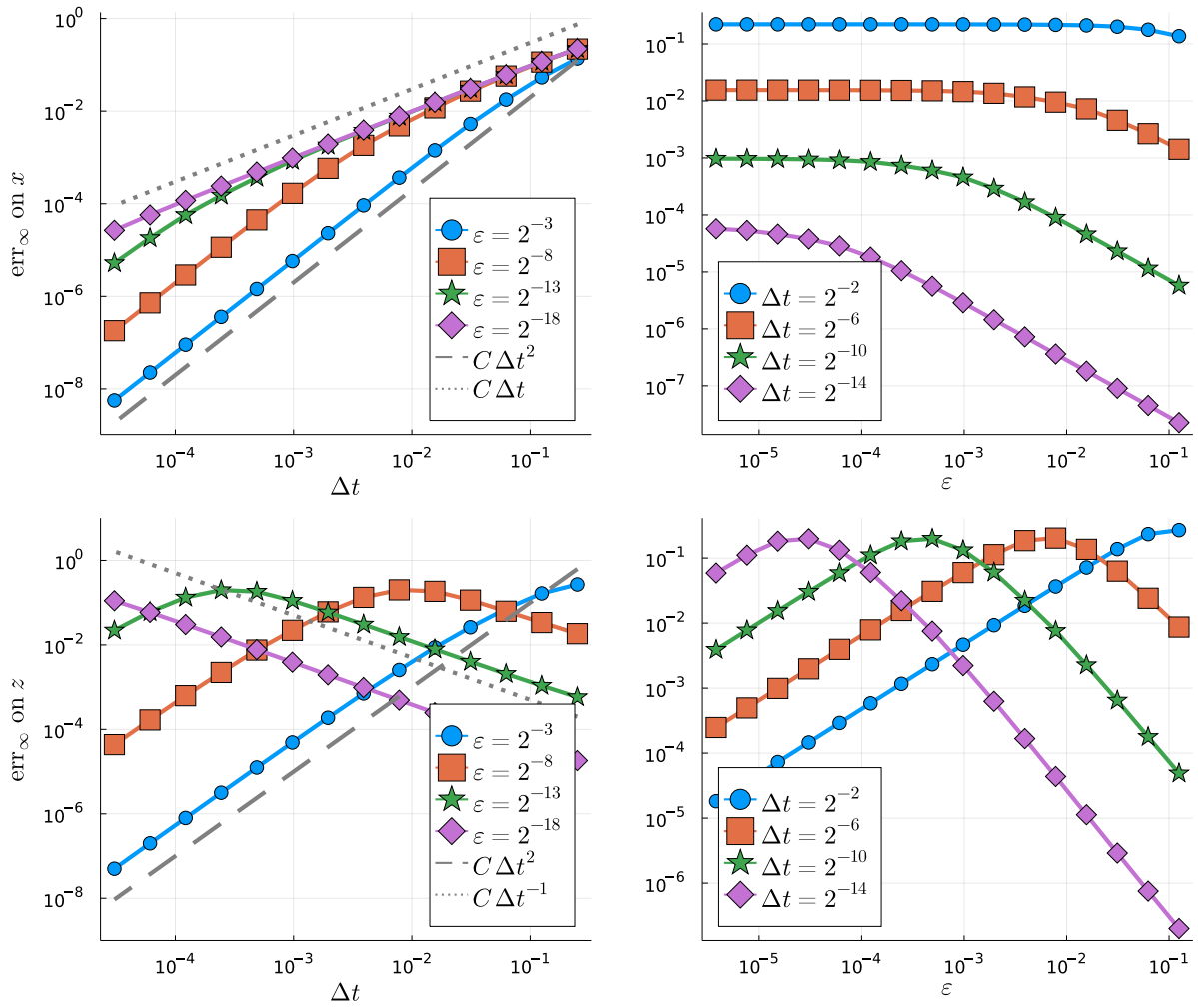


FIGURE 5 – Erreurs sur x (en haut) et z (en bas) en fonction de Δt (à gauche) et de ε (à droite) avec le schéma IMEX-BDF2. Sur les tracés de l'erreur en fonction de Δt, la convergence théorique est tracée, ainsi que pour x la convergence uniforme et pour z l'ordre négatif observé sur une portion de valeurs de Δt.

Ces méthodes sont néanmoins très utilisées, notamment dans le contexte de modèles cinétiques. Par exemple les méthodes IMEX-LM développées dans [LM08 ; BPR17 ; ADP20] sur un système

$$\partial_t \rho + \partial_x j = 0, \qquad \partial_t j + \frac{1}{\varepsilon}\partial_x \rho = -\frac{1}{\varepsilon} j$$

traitent de manière la partie en j en gardant explicite la partie en ρ. Cela permet d'avoir des schémas qui se comportent bien dans la limite $\varepsilon \to 0$ malgré la raideur sur le transport en ρ.

En outre, si la donnée initiale se situe proche de l'équilibre $z = \varepsilon h^\varepsilon(x)$ de sorte que

$\partial_t^{q+1} u$ reste bornée dans la limite $\varepsilon \to 0$, alors on peut obtenir une convergence uniforme. On peut voir une interprétation de ce résultat en Figure 6 où l'erreur sur z est améliorée en prenant une donnée initiale nulle. [5] Ainsi, il est fréquent de choisir une donnée initiale bien préparée et d'annoncer que ce type de schéma est « uniformément précis », par exemple dans [JPT00 ; HS21]. La même confusion est faite pour les schémas IMEX-RK dans [BR09 ; BPR17], par exemple.

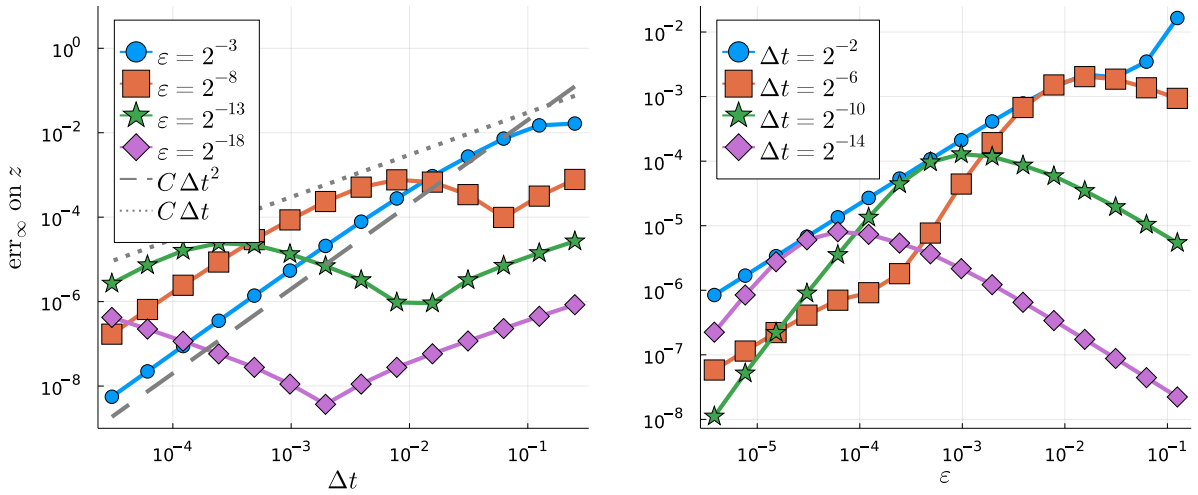


FIGURE 6 – Erreur sur z en fonction de Δt (à gauche) et de ε (à droite) avec le schéma IMEX-BDF2, avec une donnée initiale $z(0) = 0$. Sur les tracés de l'erreur en fonction de Δt, les convergences théorique et uniforme sont tracées.

D'autres notions de convergence

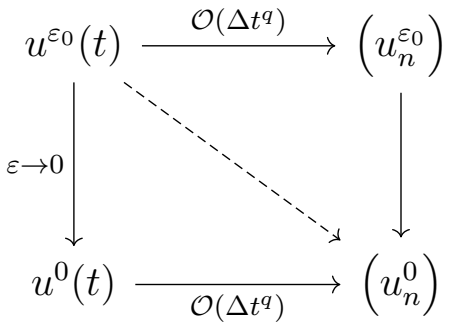
On voit que toutes ces méthodes subissent une réduction d'ordre : on passe d'une convergence d'ordre 2 à une convergence d'ordre 1, ou pire. Pourtant, les schémas sont bien *d'ordre 2* au sens asymptotique $\Delta t \to 0$, comme on l'observe pour $\varepsilon = 2^{-3}$ en Figures 3, 4 et 5. Il apparaît donc nécessaire d'introduire des concepts de convergence qui diffèrent des définitions usuelles.

Définition. *Considérons la solution $t \mapsto u^\varepsilon(t)$ du problème (1) avec une donnée initiale $u(0) = u_0$ indépendante de ε. On construit une solution approchée (u_n^ε) en appliquant un schéma $\Phi_{\Delta t}^\varepsilon$ d'ordre $q \geq 1$, et on suppose que $\Phi_{\Delta t}^\varepsilon$ admet une limite $\varepsilon \to 0$, qu'on note $\Phi_{\Delta t}^0$.*

5. Le même phénomène est observé pour le schéma expRK.

On dit que le schéma $\Phi^{\varepsilon}_{\Delta t}$ est asymptotic preserving *(AP) ou qu'il préserve l'asymptote si le schéma limite $\Phi^{0}_{\Delta t}$ existe et est du même ordre q que le schéma d'origine. On dit en outre que le schéma est* uniformly accurate *(UA) ou uniformément précis si l'erreur uniforme présente le même ordre de convergence que l'erreur « standard » du schéma.*

On peut résumer ces propriétés à l'aide diagramme de commutation suivant, où on distingue le régime non-raide $\varepsilon = \varepsilon_0 \sim 1$ et le régime asymptotique $\varepsilon \to 0$:



Les flèches verticales représentent le passage à la limite $\varepsilon \to 0$ tandis que les flèches horizontales représentent un calcul de solution numérique avec un ordre q. Un schéma UA permet d'emprunter la flèche en pointillés sans perte de précision (c'est-à-dire avec n'importe quelle valeur de ε), tandis qu'un schéma AP ne présente une bonne convergence que le long des flèches solides.

Comme mentionné précédemment, certains articles annoncent que des schémas sont UA mais en rajoutant une hypothèse de donnée initiale « bien préparée », i.e. proche de l'équilibre. Cette donnée initiale se traduit généralement en l'identité $z(0) = \varepsilon h^{\varepsilon}(x(0)) + \mathcal{O}(\varepsilon^q)$ grâce au théorème de variété centrale. On parlera alors de schéma UA « à l'équilibre ». Parmi les méthodes précédentes, on a les propriétés suivantes

	AP	UA éq.	UA
Strang			
IMEX-BDF2		✓	
expRK2	✓	✓	

La colonne UA étant vide pour ces schémas d'ordre 2, on cherche à développer de telles méthodes.

Contribution personnelle

Suite aux résultats de précision uniforme obtenus pour les problèmes hautement oscillants [CCLM15 ; CJL17 ; CLMV20], ce travail de thèse a cherché à développer des méthodes à précision uniforme dans le cadre des problèmes à relaxation rapide de type (1). Il existe deux grandes stratégies pour obtenir une convergence uniforme. La première est le développement double échelle, où on écrit la solution $t \mapsto u(t)$ comme une évaluation particulière d'une fonction à deux variable $(t, \theta) \mapsto U(t, \theta)$ en posant

$$u(t) = U(t, \theta)|_{\theta = t/\varepsilon}.$$

L'apparition de cette seconde variable permet de choisir une donnée initiale $U(0, \theta)$ qui réduit la raideur dans la direction t. La seconde stratégie est la décomposition micro-macro. L'idée est similaire à celle du double-échelle ; il s'agit de séparer la dynamique rapide oscillante (en $e^{it/\varepsilon}$) et la dynamique de *dérive* (en t). Ainsi on écrit

$$u(t) = \Omega^\varepsilon_{t/\varepsilon} \circ \Gamma^\varepsilon_t \circ \left(\Omega^\varepsilon_0\right)^{-1}(u_0)$$

où $\theta \mapsto \Omega^\varepsilon_\theta$ est périodique et Γ^ε_t est le t-flot d'un champ de vecteurs non-raide F^ε. C'est cette seconde approche que nous avons privilégié, puisqu'elle permet de ne pas avoir à considérer de seconde variable θ lors de la résolution numérique et d'utiliser des schémas standards. Néanmoins, on fournit en Annexe A des pistes pour adapter un développement double-échelle au cadre dissipatif.

L'idée de la méthode micro-macro est de chercher des approximations de Ω^ε et de F^ε à un certain ordre ε^{n+1} près, dans la veine des méthodes de moyennisation [Per69 ; LM88 ; SVM07 ; CMS10 ; CMS12a ; CCMM15]. La nouveauté ici est de s'intéresser en outre au reste de ce développement asymptotique, ainsi on obtient une décomposition exacte

$$u(t) = \Omega^{[n]}_{t/\varepsilon}\Big(v(t)\Big) + w(t)$$

avec $w(t) = \mathcal{O}(\varepsilon^{n+1})$ et $\partial_t v = F^{[n]}(v)$. Les applications $\Omega^{[n]}$ et $F^{[n]}$ sont des approximations de Ω^ε et F^ε respectivement. Dans [CLMV20], le problème sur (v, w) est moins raide et peut être résolu avec un schéma numérique d'ordre uniforme n.

L'utilisation de ces méthodes de moyennisation a permis, au cours de la troisième année de thèse, de dériver les résultats préexistants en moyennisation avec certaines preuves originales, notamment sur le caractère géométrique de la moyennisation dite *strobosco-*

pique. Précédemment, les démonstrations demandaient de construire le morphisme Ω^ε et le champ de vecteurs moyen F^ε, pour ensuite raisonner par récurrence ou avec des arbres. Ces nouvelles preuves ne font appel qu'à l'équation homologique (i.e. l'équation algébrique vérifiée par Ω^ε et F^ε) et sont plus élégantes. En outre, on met en évidence certains liens forts entre moyennisation et formes normales, qui sont déjà connus (voir [SVM07]) mais peu référencés. Cette synthèse fait l'objet du Chapitre I.

L'essentiel de ce travail de thèse consisté à construire des méthodes micro-macro adaptées aux problèmes à relaxation rapide de type (1). Formellement, au lieu de voir le changement de variable $\Omega^\varepsilon_\theta$ comme une série de Fourier en $\theta \in \mathbb{R}/\mathbb{Z}$, il est vu ici comme une série exponentielle $\Omega^\varepsilon_\tau = \sum_{k\geq 0} \omega_k e^{-k\tau}$ avec $\tau \in \mathbb{R}_+$. La nouvelle difficulté est alors de calculer l'équivalent formel de la « moyenne » de cette série exponentielle. Pour cette raison, on considère $\Omega^\varepsilon_{i\tau}$ et on adapte les résultats du cas périodique. En effectuant des développements à l'ordre n, on obtient ainsi un problème micro-macro en (v, w) qu'on peut résoudre avec une précision uniforme d'ordre $n+1$ (le caractère de relaxation permet un gain d'ordre par rapport aux problèmes hautement oscillants) en utilisant des schémas expRK. On peut aussi obtenir une précision uniforme d'ordre n avec des schémas IMEX-BDF. Il est intéressant de noter que ce résultat est étendu partiellement à des EDP bien connues : les problèmes hyperboliques relaxés

$$\begin{cases} \partial_t v_1 + \partial_x v_2 = 0, \\ \partial_t v_2 + \partial_x v_1 = \dfrac{1}{\varepsilon}\Big(g(v_1) - v_2\Big), \end{cases}$$

et l'équation de télégraphe (i.e. BGK à vitesses discrètes)

$$\begin{cases} \partial_t \rho + \partial_x j = 0, \\ \partial_t j + \dfrac{1}{\varepsilon}\partial_x \rho = -\dfrac{1}{\varepsilon} j. \end{cases}$$

De nombreuses méthodes AP ou UA à l'équilibre existent pour ces problèmes –on peut citer [Jin99 ; LM08 ; DP11 ; DP17 ; BPR17 ; ADP20]– mais le développement de méthodes UA est encore un sujet de recherche actif. On peut voir ce travail de thèse comme une étape préliminaire au développement de méthodes UA pour cette catégorie de problèmes. Ces résultats ont fait l'objet d'une publication,

> Philippe Chartier, Mohammed Lemou et Léopold Trémant, « A uniformly accurate numerical method for a class of dissipative systems », à paraître dans *Mathematics of Computation* (2021),

qui est présentée en Chapitre II.

En Chapitre III, on discute d'extensions directes des résultats de notre article, qui servent à rendre le résultat plus robuste, et on fournit quelques pistes pour traiter l'équation de télégraphe de manière complète.

Chapitre I

Les propriétés revisitées de la moyennisation

Ce chapitre propose une nouvelle approche sur des résultats connus de moyennisation. Dans la littérature les preuves de ces résultats font appel à des récurrences ou à des propriétés avancées sur les arbres, et requièrent de construire les applications de moyennisation, ce qui réduit les raisonnements à la construction en question. Dans cet article, on présente rapidement le cadre formel qui décrit ce qu'on entend par « moyennisation », puis on prouve quelques propriétés en supposant que les applications existent. En particulier, on s'intéresse aux propriétés géométriques de commutation, de conservation de volume et de structure hamiltonienne ou de Poisson. Dans une dernière partie, on adapte ces propriétés au cas d'applications *approchés*.

I.1 Introduction

This paper compiles results pertaining to *high-order averaging*, that is to say the problem of separating the slow and fast dynamics in a highly-oscillatory setting. The type of problem we consider may arise in many realistic physical models, such as molecular dynamics [GSS98] or charged-particle dynamics under a strong magnetic field [CCLMZ20 ; FSS09 ; FS00]. It may also arise in functional spaces ; two examples are the nonlinear Klein-Gordon equation in the nonrelativistic limit regime [BCZ14 ; BZ19 ; CLMV20] and the oscillatory nonlinear Schrödinger equation [CCLM15 ; CCMM15].

Mathematically speaking, we consider problems with forced oscillations of the form

$$\partial_t u(t) = f_{t/\varepsilon}\Big(u(t)\Big), \qquad u(0) = u_0 \in X, \qquad t \in [0,1] \tag{I.1.1}$$

where X is a Banach space of norm $|\cdot|$, the non-autonomous vector field $(\theta, u) \in \mathbb{T} \times X \mapsto f_\theta(u)$ is 1-periodic w.r.t. θ on the torus $\mathbb{T} := \mathbb{R}/\mathbb{Z}$. As mentioned, the space X may be

simply $\mathbb{R}^d$, in which case the problem is a simple ordinary differential equation in finite dimension, or it may be a functional space, such as the space of square-integrable function $L^2(\mathbb{R})$. Note that this type of equation can result from the *filtering* of an autonomous equation

$$\dot{v}^\varepsilon = \frac{1}{\varepsilon} G(v) + K(v), \quad v^\varepsilon(0) = v_0 \in X \tag{I.1.2}$$

if G generates a 1-periodic flow $(\theta, u) \mapsto \chi_\theta(u)$. It links to I.1.1 using the filtered variable $u^\varepsilon(t) = \chi_{-t/\varepsilon}\Big(v^\varepsilon(t)\Big)$ which follows an equation of the form (I.1.1) with $f_\theta(u) = (\partial_u \chi_{-\theta} \cdot K) \circ \chi_\theta(u)$.

The approach of averaging can be summarized as the decomposition of the solution $u(t)$ into a *near-identity, rapidly oscillating* change of variable $\Phi^\varepsilon_{t/\varepsilon}$ and the dynamics of an *average* autonomous vector field F^ε. This can be written

$$u(t) = \Phi^\varepsilon_{t/\varepsilon} \circ \Psi^\varepsilon_t \circ \Big(\Phi^\varepsilon_0\Big)^{-1}(u_0), \tag{I.1.3}$$

where $(\theta, u) \mapsto \Phi^\varepsilon_\theta(u)$ is 1-periodic w.r.t. θ and $(t, u) \mapsto \Psi^\varepsilon_t(u)$ is the t-flow associated to F^ε, i.e. for $(t, u) \in [0, 1] \times X$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Psi^\varepsilon_t(u) = F^\varepsilon\Big(\Psi^\varepsilon_t(u)\Big), \qquad \Psi^\varepsilon_0 = \mathrm{id}\,. \tag{I.1.4}$$

We refer to Lochak-Meunier [LM88] and Sanders-Verhulst-Murdock [SVM07] for textbooks on these issues. Since the goal is to separate the fast periodic part in θ and the slow drift in t, averaging can be seen as analogous to the two-scale expansion $u^\varepsilon(t) = U^\varepsilon(t, \theta)|_{\theta=t/\varepsilon}$ often found in the context of high-frequency PDEs. It is also similar to WKB expansions [Wen26 ; Kra26 ; Bri26], since in some sense Φ^ε captures the rapid phase dynamics and Ψ^ε the slow amplitude changes.

In this work, we shall not discuss specific methods to compute the periodic change of variable or the averaged vector fields, the traditional approach dating back to [Per69] consists in assuming the maps are power series in ε and injecting the ansatz $\Phi^\varepsilon_\theta = \mathrm{id} + \sum_{n\geq 1} \Phi^{[n]}_\theta$ in (I.1.3) and identifying like terms in ε. This formal series approach has been revisited using B-series or the Magnus expansion in [CMS10 ; CMS12a ; CCM19]. Another approach is that of "successive substitution" dating back to [Nei84] (albeit in a slightly different context), and more recently in [CCMM15 ; CLMV20]. This circumvents the ansatz and yields an *exponential* error, i.e. an error bounded by $Ce^{-\mu/\varepsilon}$ for some $C > 0$ and $\mu > 0$. Both approaches coincide formally. Our goal in this paper is to present known

results under a new light, and to offer original proofs without having to invoke any ansatz, formal series or construction process.

A particularly well-studied case is that of the autonomous problems with linearly-generated oscillations (i.e. linear G), for which the problem of averaging can often be reduced to finding some θ-independent change of variable $\left(\Phi_0^\varepsilon\right)^{-1}$, or some equivalent. It is then possible to consider the problem on this new variable $\left(\Phi_0^\varepsilon\right)^{-1}(u(t))$. As such, a link can be made with normal forms, and specifically Birkhoff's forms technique have been considered in this context by Bambusi [Bam03 ; BB05 ; Bam06 ; Bam08], Bourgain [Bou96], Colliander [CKSTT10 ; CKO12] and Grébert [Bam06 ; GV11 ; GT12], to mention only a few. We offer some insight on this approach in Section I.3.3. Note that many of these works consider the setting of multiple non-resonant frequencies, which is akin to considering f as a function of multiple phases $\theta_1, \theta_2, \dots$ in (I.1.1). This setting has also been studied with averaging using diophantine approximations in [CMTZ17] and with B-series in [CMS12b].

In Section I.2, we present some general properties of averaging, detailing the differences between standard and stroboscopic averaging. In Section I.3, we present some remarkable properties of averaging in the autonomous case. In Section I.4, we restrain ourselves to stroboscopic averaging, and present some of its geometric properties. Finally, in Section I.5, we discuss what is retained from the previous results in the case of a bounded domain.

I.2 A brief presentation of averaging

Differentiating (I.1.3) w.r.t. t generates

$$f_{t/\varepsilon} \circ \Phi^\varepsilon_{t/\varepsilon}\Big(v(t)\Big) = \frac{1}{\varepsilon}\partial_\theta \Phi^\varepsilon_{t/\varepsilon}\Big(v(t)\Big) + \partial_u \Phi^\varepsilon_{t/\varepsilon}\Big(v(t)\Big) \cdot F^\varepsilon\Big(v(t)\Big)$$

with $v(t) = \Psi^\varepsilon_t \circ \left(\Phi^\varepsilon_0\right)^{-1}(u_0)$ the average dynamics. By separating the rapid oscillations in t/ε and the slow drift in t, one obtains the homological equation, which is for $(\theta, u) \in \mathbb{T} \times X$,

$$\partial_\theta \Phi^\varepsilon_\theta(u) = \varepsilon \left(f_\theta \circ \Phi^\varepsilon_\theta(u) - \partial_u \Phi^\varepsilon_\theta(u)\, F^\varepsilon(u)\right). \tag{I.2.5}$$

Now taking the average, it appears that the change of variable Φ^ε alone stores the information of the averaged vector field. Indeed, for u in X, $F^\varepsilon(u)$ is given by

$$F^\varepsilon(u) = \Big(\partial_u \langle \Phi^\varepsilon \rangle(u)\Big)^{-1} \langle f \circ \Phi \rangle(u), \tag{I.2.6}$$

where $\langle \cdot \rangle$ denotes the average, defined for a periodic map $(\theta, u) \in \mathbb{T} \times X \mapsto \varphi_\theta(u)$ by

$$\langle \varphi \rangle(u) = \int_0^1 \varphi_\theta(u) \mathrm{d}\theta. \tag{I.2.7}$$

Up to a change of variable, Φ^ε is assumed to be near identity, i.e.

$$\Phi^\varepsilon = \mathrm{id} + \mathcal{O}(\varepsilon). \tag{I.2.8}$$

It is known that equation (I.2.5) generally has no rigorous solution, only solutions as a formal series in ε. An example where this divergence is observed can be found in [CMS10]. However the series converges in the case where f_θ is a linear and bounded operator, for ε small enough.

Hypothèses sur Φ^ε et F^ε, du style bornées sur E et smooth.

Perhaps the most straighforward approach to solve the homological equation is a fixed point method separating the right-hand side of the equation (of size ε) and the left (of size 1). It immediately appears that a closure condition on Φ^ε is needed to properly invert ∂_θ. Two choices are often considered.

Standard averaging : $\langle \Phi^\varepsilon \rangle = \mathrm{id}$,

also called the Chapmann-Enskog method in the context of kinetic theory (see [CCLM20]). This choice circumvents the computation of an inverse, as then $F^\varepsilon = \langle f \circ \Phi^\varepsilon \rangle$, therefore computations are not too costly. As highlighted in [CLMZ20], in numerical contexts the $\partial_u \Phi^\varepsilon \cdot F^\varepsilon$-term can be replaced by a finite-differences approximation up to some order in ε, thereby removing the need to compute an exact derivative and making automatic computations much simpler.

Stroboscopic averaging : $\Phi_0^\varepsilon = \mathrm{id}$,

for which the solution $u(t)$ coincides with the average $\Psi_t^\varepsilon(u_0)$ at "stroboscopic" times $t \in \varepsilon\mathbb{N}$. This produces more complex computations but renders fairly straighforward the conservation of geometric properties, such as energy preservation or symplectic structure.

We shall focus on the properties of stroboscopic averaging in the upcoming section, but it is important to keep in mind that these choices are conjugate. Indeed, the latter can be obtained from the former by setting

$$\Phi^{strob} = \Phi^{std} \circ \left(\Phi_0^{std}\right)^{-1} \quad \text{and} \quad \Psi^{strob} = \Phi_0^{std} \circ \Psi^{std} \circ \left(\Phi_0^{std}\right)^{-1},$$

i.e. $F^{strob} = \left(\partial_u \Phi_0^{std} \cdot F^{std}\right) \circ \left(\Phi_0^{std}\right)^{-1}$. Conversely, standard averaging can be obtained from stroboscopic averaging with the relations

$$\Phi^{std} = \Phi^{strob} \circ \left\langle \Phi^{strob} \right\rangle^{-1} \quad \text{and} \quad \Psi^{std} = \left\langle \Phi^{strob} \right\rangle \circ \Psi^{strob} \circ \left\langle \Phi^{strob} \right\rangle^{-1}. \tag{I.2.9}$$

Thus some properties of standard averaging will also be discussed.

I.3 Commutation of flows in the autonomous case

In this section we restrict ourselves to the case of an autonomous equation of the form

$$\dot{v}^\varepsilon = \frac{1}{\varepsilon} G(v^\varepsilon) + K(v^\varepsilon), \quad v^\varepsilon(0) = v_0 \in X \tag{I.3.10}$$

where G and K are smooth function from a Banach space X into itself and where G generates a 1-periodic flow $(\theta, u) \mapsto \chi_\theta(u)$. The approach is the same as for the non-autonomous problem, which is to say we search a solution under the form

$$v^\varepsilon(t) = \Omega^\varepsilon_{t/\varepsilon} \circ \Psi^\varepsilon_t \circ (\Omega^\varepsilon_0)^{-1}(v_0) \tag{I.3.11}$$

where $\theta \mapsto \Omega^\varepsilon_\theta$ is assumed to be 1-periodic and Ψ^ε_t is the t-flow associated to the averaged vector flow K^ε. The reasons why the notation of the change of variable changed but not that of the average flow will be made clear as this section progresses. The homological equation is now

$$\partial_\theta \Omega^\varepsilon_\theta(u) - G \circ \Omega^\varepsilon_\theta(u) = \varepsilon \Big(K \circ \Omega^\varepsilon_\theta(u) - \partial_u \Omega^\varepsilon_\theta(u)\, K^\varepsilon(u) \Big). \tag{I.3.12}$$

It appears that the closure condition of standard averaging must be reconsidered. Indeed, in the limit $\varepsilon \to 0$, the change of variable $\Omega^\varepsilon_\theta$ approaches $\chi_{\theta+\theta_0}$ for some initial phase θ_0. Consider for instance the case $G(u) = 2\pi \begin{pmatrix} -u_2 \\ u_1 \end{pmatrix} = 2\pi J u$, then clearly choosing

the standard closure condition $\langle \Omega^\varepsilon \rangle = \mathrm{id}$ cannot hold, as $\langle \chi \rangle = 0$. Rather than discarding standard averaging altogether, we may *filter* the equation, which is to say transform it into a forcibly-oscillating problem by left-multiplying it by $\partial_u \chi_{-\theta+\theta_1}\big(\Omega^\varepsilon_\theta\big)$ for some arbitrary phase θ_1. Define the filtered change of variable $\Phi^\varepsilon_{\theta,\theta_1} = \chi_{-\theta+\theta_1} \circ \Omega^\varepsilon_\theta$, it satisfies [1]

$$\partial_\theta \Phi^\varepsilon_{\theta,\theta_1}(u) = \varepsilon\Big(f_{\theta,\theta_1} \circ \Phi^\varepsilon_{\theta,\theta_1}(u) - \partial_u \Phi^\varepsilon_{\theta,\theta_1}(u)\, K^\varepsilon(u) \Big) \tag{I.3.13}$$

with $f_{\theta,\theta_1}(u) = (\partial_u \chi_{-\theta+\theta_1} \cdot K) \circ \chi_{\theta-\theta_1}(u)$. Note that we exploited the identity $\partial_\theta \chi_\theta = G \circ \chi_\theta = \partial_u \chi_\theta\, G$. Take now the average on θ of (I.3.13),

$$0 = \varepsilon\Big(\big\langle f_{\bullet,\theta_1} \circ \Phi_{\bullet,\theta_1} \big\rangle(u) - \partial_u \big\langle \Phi_{\bullet,\theta_1} \big\rangle(u)\, K^\varepsilon(u) \Big). \tag{I.3.14}$$

The standard choice of closure condition is therefore $\langle \Phi_{\bullet,\theta_1} \rangle = \mathrm{id}$, i.e. $\Omega^\varepsilon_\theta$ close to $\chi_{\theta-\theta_1}$. Remember however that the phase shift θ_1 is arbitrary, therefore there are an infinite number of standard closure conditions, the canonical one being $\langle \chi_{-\theta} \circ \Omega^\varepsilon_\theta \rangle = \mathrm{id}$.

Whatever the closure condition, it is possible to obtain K^ε from (I.3.14), since $\partial_u \langle \Phi^\varepsilon_{\bullet,\theta_1} \rangle$ is invertible. Indeed, assuming that $\Omega^\varepsilon_\theta$ is close to $\chi_{\theta+\theta_0}$, all filtered changes of variable satisfy

$$\Phi^\varepsilon_{\theta,\theta_1} = \chi_{-\theta+\theta_1} \circ \big(\chi_{\theta+\theta_0} + \mathcal{O}(\varepsilon) \big) = \chi_{\theta_1+\theta_0} + \mathcal{O}(\varepsilon).$$

This generates the identity

$$K^\varepsilon(u) = \Big(\partial_u \big\langle \chi_{-\theta+\theta_1} \circ \Omega^\varepsilon_\theta \big\rangle(u) \Big)^{-1} \Big\langle (\partial_u \chi_{-\theta+\theta_1} \cdot K) \circ \Omega^\varepsilon_\theta \Big\rangle(u). \tag{I.3.15}$$

Defining an operator extracting the average behaviour

$$\mathcal{A}^{\theta_1}[\varphi] := \Big(\partial_u \big\langle \chi_{-\theta+\theta_1} \circ \varphi_\theta \big\rangle \Big)^{-1} \big\langle (\partial_u \chi_{-\theta+\theta_1} \cdot K) \circ \varphi_\theta \big\rangle, \tag{I.3.16}$$

the change of variable Ω^ε may be defined as the unique solution to the homological equation

$$\partial_\theta \Omega^\varepsilon_\theta - G \circ \Omega^\varepsilon_\theta = \varepsilon\Big(K \circ \Omega^\varepsilon_\theta - \partial_u \Omega^\varepsilon_\theta \cdot \mathcal{A}^{\theta_1}[\Omega^\varepsilon] \Big) \tag{I.3.17}$$

that is 1-periodic and satisfies some closure condition. Note that the above equation is considered with fixed θ_1, but modifying this phase has no impact on the definition of Ω^ε.

1. This homological equation can also be obtained directly by considering the filtered problem of form (I.1.1) satisfied by $u^\varepsilon_{\theta_1}(t) = \chi_{-t/\varepsilon+\theta_1}(v^\varepsilon(t))$, which is $\partial_t u^\varepsilon_{\theta_1}(t) = f_{t/\varepsilon,\theta_1}\big(u^\varepsilon_{\theta_1}(t)\big)$.

Linking with (I.3.12), this may be restated as

$$\forall \theta_1 \in \mathbb{T}, \qquad K^\varepsilon = \mathcal{A}^{\theta_1}[\Omega^\varepsilon] = \mathcal{A}^0[\Omega^\varepsilon].$$

Thanks to this invariance, a group relation may be found in the case of stroboscopic averaging, summarized by the following proposition.

Proposition I.3.1. *When considering stroboscopic averaging, for all θ and all θ_0, the following group relation is satisfied*

$$\Omega^\varepsilon_\theta \circ \Omega^\varepsilon_{\theta_0} = \Omega^\varepsilon_{\theta+\theta_0}.$$

Equivalently, there exists a vector field G^ε such that

$$\forall \theta,\ \forall u, \qquad \frac{\mathrm{d}}{\mathrm{d}\theta}\Omega^\varepsilon_\theta(u) = G^\varepsilon \circ \Omega^\varepsilon_\theta(u).$$

Proof. Consider the θ-map

$$\widetilde{\Omega}^\varepsilon_\theta = \Omega^\varepsilon_{\theta+\theta_0} \circ (\Omega^\varepsilon_{\theta_0})^{-1}.$$

Writing equation (I.3.12) with θ replaced by $\theta+\theta_0$ and $(\Omega^\varepsilon_{\theta_0})^{-1}(u)$ in lieu of u, we obtain with all maps evaluated in u,

$$\partial_\theta \widetilde{\Omega}^\varepsilon_\theta - G \circ \widetilde{\Omega}^\varepsilon_\theta = \varepsilon \left(K \circ \widetilde{\Omega}^\varepsilon_\theta - \partial_u \widetilde{\Omega}^\varepsilon_\theta \cdot \left(\partial_u (\Omega^\varepsilon_{\theta_0})^{-1}\right)^{-1} \cdot K^\varepsilon \circ (\Omega^\varepsilon_{\theta_0})^{-1} \right). \tag{I.3.18}$$

The new averaged vector field $\widetilde{K}^\varepsilon = \left(\partial_u (\Omega^\varepsilon_{\theta_0})^{-1}\right)^{-1} \cdot K^\varepsilon \circ (\Omega^\varepsilon_{\theta_0})^{-1}$ can be written

$$\widetilde{K}^\varepsilon = \left(\left(\partial_u \left\langle \chi_{-\theta+\theta_0} \circ \Omega^\varepsilon_\theta \right\rangle \right) \circ \left(\Omega^\varepsilon_{\theta_0}\right)^{-1} \cdot \partial_u (\Omega^\varepsilon_{\theta_0})^{-1} \right)^{-1} \left\langle (\partial_u \chi_{-\theta+\theta_0} \cdot K) \circ \Omega^\varepsilon_\theta \circ (\Omega^\varepsilon_{\theta_0})^{-1} \right\rangle,$$

exploiting (I.3.15) with $\theta_1 = \theta_0$. The derivatives can be concatenated into $\partial_u \left\langle \chi_{-\theta+\theta_0} \circ \Omega^\varepsilon_\theta \circ (\Omega^\varepsilon_{\theta_0})^{-1} \right\rangle$. Exploiting then the phase invariance of the average, i.e. $\langle \varphi_\theta \rangle = \langle \varphi_{\theta+\theta_0} \rangle$, the identity becomes

$$\widetilde{K}^\varepsilon = \left(\partial_u \left\langle \chi_{-\theta} \circ \widetilde{\Omega}^\varepsilon_\theta \right\rangle \right)^{-1} \left\langle (\partial_u \chi_{-\theta} \cdot K) \circ \widetilde{\Omega}^\varepsilon_\theta \right\rangle = \mathcal{A}^0[\widetilde{\Omega}^\varepsilon].$$

Injecting this into (I.3.18), we find that $\widetilde{\Omega}^\varepsilon$ is a 1-periodic map which satisfies an equation of the form (I.3.17). As we only consider stroboscopic averaging, $\widetilde{\Omega}^\varepsilon$ also satisfies the same

closure condition as Ω^ε, which is to say $\widetilde{\Omega}^\varepsilon_0 = \Omega^\varepsilon_0 = \mathrm{id}$. Therefore, the two maps coincide and the proof is complete.

□

Proposition I.3.2. *The flows $\theta \mapsto \Omega^\varepsilon_\theta$ and $t \mapsto \Psi^\varepsilon_t$ commute with each other, i.e.*

$$\forall \theta, \quad \forall t, \qquad \Omega^\varepsilon_\theta \circ \Psi^\varepsilon_t = \Psi^\varepsilon_t \circ \Omega^\varepsilon_\theta.$$

Equivalently, the vector fields G^ε and K^ε commute with each other, i.e.

$$[G^\varepsilon, K^\varepsilon] = 0$$

where $[\cdot, \cdot]$ is the usual Lie-bracket.

Proof. The group law for $t \mapsto \Omega^\varepsilon_{t/\varepsilon} \circ \Psi^\varepsilon_t$ (recall that equation (I.3.10) is autonomous) gives for all s and t

$$\left(\Omega^\varepsilon_{s/\varepsilon} \circ \Psi^\varepsilon_s\right) \circ \left(\Omega^\varepsilon_{t/\varepsilon} \circ \Psi^\varepsilon_t\right) = \Omega^\varepsilon_{(s+t)/\varepsilon} \circ \Psi^\varepsilon_{s+t}. \tag{I.3.19}$$

The t-flow Ψ^ε_t satisfies a group-law by construction and owing to Proposition I.3.1, this is also the case for Ω^ε_τ. Hence, we can compose equation (I.3.19) from the left by $\Omega^\varepsilon_{-s/\varepsilon}$ and from the right by Ψ^ε_{-t} and obtain

$$\Psi^\varepsilon_s \circ \Omega^\varepsilon_{t/\varepsilon} = \Omega^\varepsilon_{t/\varepsilon} \circ \Psi^\varepsilon_s.$$

The commutation of the vector fields then follows in a standard way. □

Note that this result can also be obtained from the proof of Proposition I.3.1, since there we find

$$K^\varepsilon = \widetilde{K}^\varepsilon = \left(\partial_u (\Omega^\varepsilon_{\theta_0})^{-1}\right)^{-1} \cdot K^\varepsilon \circ (\Omega^\varepsilon_{\theta_0})^{-1},$$

i.e. K^ε is invariant when conjugated by $\Omega^\varepsilon_{\theta_0}$.

Remark I.3.3. *If G is linear, then differentiating $\Phi^\varepsilon_{\theta,\theta_0}$ w.r.t. θ and taking the average generates*

$$G^\varepsilon = \left\langle \partial_u \Phi^\varepsilon_{\theta,\theta_0} \right\rangle^{-1} G \left\langle \Phi^\varepsilon_{\theta,\theta_0} \right\rangle = \left(\partial_u \Omega^{std}_0 \cdot G\right) \circ \left(\Omega^{std}_0\right)^{-1}$$

if Ω^{std} is such that $\langle e^{-(\theta-\theta_0)G} \Omega^{std} \rangle = \mathrm{id}$ owing to (I.2.9)*. Furthermore the average vector*

field K^{std} commutes with G, thanks to the identity

$$[G, K^{std}] = [\mathbb{S}(G^\varepsilon), \mathbb{S}(K^\varepsilon)] = \mathbb{S}\Big([G^\varepsilon, K^\varepsilon]\Big) = 0$$

with $\mathbb{S}(F) = \left(\partial_u \Phi_0^{std} \cdot F\right) \circ \left(\Phi_0^{std}\right)^{-1}$. In other words, the change of variable $\left(\Omega_0^{std}\right)^{-1}$ transforms the perturbed vector field $G + \varepsilon K$ into $G + \varepsilon K^{std}$, where G and K^{std} commute. This links to the vision of normal forms as presented in [SVM07, Chap. IX].

I.4 Stroboscopic averaging and geometry

We start by introducing some geometric properties, then prove they are preserved by stroboscopic averaging.

I.4.1 Definitions of geometric properties

Definition I.4.1. *Define the matrix $J \in \mathcal{M}(\mathbb{R}^{2n})$ as the block matrix*

$$J = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix}.$$

A vector field function $f : \mathbb{R}^{2n} \mapsto \mathbb{R}^{2n}$ is said to be canonically Hamiltonian if there exists a scalar smooth function $H : \mathbb{R}^{2n} \mapsto \mathbb{R}$ such that

$$\forall u \in \mathbb{R}^{2n}, \quad f(u) = J^{-1} \nabla_u H(u).$$

A smooth map $(\tau, u) \in \mathbb{R} \times \mathbb{R}^{2n} \mapsto S_\tau(u) \in \mathbb{R}^{2n}$ is said to be symplectic iff

$$\forall (\tau, u) \in \mathbb{R} \times \mathbb{R}^{2n}, \quad \left(\partial_u S_\tau(u)\right)^T J \left(\partial_u S_\tau(u)\right) = J, \tag{I.4.20}$$

or equivalently

$$\forall (\tau, u) \in \mathbb{R} \times \mathbb{R}^{2n}, \quad \left(\partial_u S_\tau(u)\right) J^{-1} \left(\partial_u S_\tau(u)\right)^T = J^{-1}. \tag{I.4.21}$$

Remark I.4.2. *It is known that the τ-flow of a canonically Hamiltonian system is symplectic and that the converse is also true on connected sets. This is proved by differentiation and use of the integrability Lemma, which asserts that, on a connected set, a vector*

function derives from a gradient iff its jacobian is symmetric.

Definition I.4.3. *A vector field function $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ is said to be divergence free*

$$\forall u \in \mathbb{R}^{2n}, \quad \sum_{i=1}^{n} \partial_i f_i(u) = \mathrm{tr}(\partial_u f) = 0.$$

A smooth function $(\tau, u) \in \mathbb{R} \times \mathbb{R}^n \mapsto S_\tau(u) \in \mathbb{R}^n$ is said to be volume-preserving iff

$$\forall (\tau, u) \in \mathbb{R} \times \mathbb{R}^n, \quad \det\left(\partial_u S_\tau(u)\right) = 1.$$

Remark I.4.4. *By differentiation of the determinant, it is straightforward that the τ-flow of a divergence-free vector field is volume preserving. The converse is true as well.*

Definition I.4.5. *A matrix $B(u) \in \mathcal{M}(\mathbb{R}^n)$ is said to be a Poisson matrix if it is skew-symmetric and satisfies the Jacobi relation*

$$\forall i, j, k \in \{1, \ldots, n\}, \quad \sum_{l=1}^{n} (\partial_l b_{ij}) b_{lk} + (\partial_l b_{jk}) b_{li} + (\partial_l b_{ki}) b_{lj} = 0.$$

A vector field function $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ is said to be Poisson if there exists a scalar smooth function $H : \mathbb{R}^n \mapsto \mathbb{R}$ and a Poisson matrix $B(u)$ such that

$$\forall u \in \mathbb{R}^n, \quad f(u) = B(u) \nabla_u H(u).$$

A smooth function $(\tau, u) \in \mathbb{R} \times \mathbb{R}^n \mapsto S_\tau(u) \in \mathbb{R}^n$ is said to be a Poisson map iff

$$\forall u \in \mathbb{R}^n, \quad \left(\partial_u S_\tau(u)\right) B(u) \left(\partial_u S_\tau(u)\right)^T = B(S_\tau(u)).$$

Remark I.4.6. *The τ-flow of a Poisson system is a Poisson map, and the converse is locally true if in addition the Casimirs (spanning the null space of B) are preserved by the flow. This result is found for instance in [HLW06, Chap. VII, Thm. 4.5].*

I.4.2 The geometry of stroboscopic averaging

If both G and K satisfy a geometric property (i.e. are divergence-free or B-Poisson, or have an invariant) in the autonomous case, then this property is also satisfied by f_θ for all $\theta \in \mathbb{T}$ in the associated filtered problem with forced oscillations. Indeed, since G and K satisfy the same property, they belong to the same Lie algebra, and the flow of G

is in the generating group of the algebra. Because f_θ is obtained by conjugating K by the flow of G, i.e. by conjugating an element of the algebra by an element of the group, it still belongs to the algebra.

À clarifier

As the averaged vector field of the autonomous case coincides with the averaged vector field of this forced case, any geometric reasoning conducted on the forced case can be extended to the autonomous case. For this reason, we only state the following result in the case of forced oscillations.

Theorem I.4.7. *If $\varepsilon\partial_u F^\varepsilon$ is bounded, then stroboscopic averaging is a geometric procedure. More precisely, for ε small enough, if for all $\theta \in \mathbb{T}$,*

(i) f_θ is a divergence-free vector field and X is of dimension $d < \infty$, then F^ε is also divergence-free.

(ii) the functional I is preserved by the flow of f_θ, then it is an invariant of F^ε.

(iii) f_θ is a Hamiltonian vector field, then F^ε is Hamiltonian.

(iv) f_θ is a B-Poisson vector field, then F^ε is B-Poisson.

Proof. For the sake of the upcoming proofs, we shall denote for any map $(t,u) \mapsto \varphi_t(u) \in E$ with $(E, |\cdot|)$ a Banach space, at fixed t,

$$\|\varphi_t\| = \sup_{u\in X} |\varphi_t(u)| \qquad \text{and} \qquad \|\varphi\|_\varepsilon = \sup_{t\leq\varepsilon} \|\varphi_t\|.$$

We also set $C = \sup_\varepsilon \varepsilon\|\partial_u F^\varepsilon\|$.

(i) For $(t,u) \in [0,\varepsilon] \times X$, write $R_t(u)$ the deviation from volume preservation,

$$R_t(u) = \det\left(\partial_u \Psi_t^\varepsilon(u)\right) - 1.$$

Setting $L_t(u) = S_t(u) = \operatorname{tr}(\partial_u F^\varepsilon) \circ \Psi_t^\varepsilon(u)$, it satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} R_t = L_t\, R_t + S_t, \quad \text{i.e.} \quad R_t = \int_0^t L_\tau\, R_\tau \mathrm{d}\tau + \int_0^t S_\tau\, \mathrm{d}\tau. \tag{I.4.22}$$

Taylor's theorem with integral remainder generates the identity

$$R_\varepsilon = R_0 + \varepsilon S_0 + \int_0^\varepsilon (\varepsilon - t)\dot{R}_t \mathrm{d}t,$$

which can be simplified thanks the periodicity of Φ^ε, as then $R_\varepsilon = R_0 = 0$. Hence the bound

$$\|S_0\| \leq \frac{\varepsilon}{2}\|\dot{R}\|_\varepsilon \leq \frac{\varepsilon}{2}\left(\|L\|_\varepsilon \|R\|_\varepsilon + \|S\|_\varepsilon\right).$$

Now applying Gronwall's lemma to the integral form of R yields $\|R_t\| \leq t\|S\|_\varepsilon e^{t\|L\|_\varepsilon}$, which injects into the previous bound

$$\|S_0\| \leq \varepsilon\left(\varepsilon\|L\|_\varepsilon e^{\varepsilon\|L\|_\varepsilon} + 1\right)\|S\|_\varepsilon.$$

Since $S_t = S_0 \circ \Psi_t^\varepsilon$, it appears by one-to-one property of Ψ_t^ε that $\|S_0\| = \|S\|_\varepsilon$. The same can be said for L, thus $\varepsilon\|L\|_\varepsilon \leq dC$. We finally obtain

$$\|S\|_\varepsilon \leq \varepsilon\left(1 + dCe^{dC}\right)\|S\|_\varepsilon$$

therefore for $\varepsilon < \left(1 + dCe^{dC}\right)^{-1}$, the source term in (I.4.22) is zero, and the t-flow Ψ_t^ε is volume-preserving. Equivalently, the averaged vector field F^ε is divergence-free.

(ii) From the identity

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[I \circ \Psi_t^\varepsilon\right] = \left(\partial_u I \cdot F^\varepsilon\right) \circ \Psi_t^\varepsilon,$$

Taylor's theorem generates

$$I \circ \Psi_\varepsilon^\varepsilon = I + \varepsilon\partial_u I \cdot F^\varepsilon + \int_0^\varepsilon (\varepsilon - t)\left(\partial_u I \cdot F^\varepsilon\right) \circ \Psi_t^\varepsilon \mathrm{d}t.$$

By periodicity of Φ^ε, $I \circ \Psi_\varepsilon^\varepsilon = I$, therefore $\|\partial_u I \cdot F^\varepsilon\| \leq \frac{\varepsilon}{2}\|(\partial_u I \cdot F^\varepsilon) \circ \Psi^\varepsilon\|_\varepsilon$. By one-to-one property of Ψ_t^ε at all t, both norms are equal, therefore for $\varepsilon < 2$,

$$\|\partial_u I \cdot F^\varepsilon\| = 0,$$

i.e. I is an invariant of F^ε.

(iii) Writing R_t the deviation from simplecticity,

$$R_t = \partial_u \Psi_t^\varepsilon J^{-1}\left(\partial_u \Psi_t^\varepsilon\right)^T - J^{-1},$$

it satisfies an equation of the form (I.4.22) with $L_t M = \partial_u F^\varepsilon(\Psi_t^\varepsilon)M + M\left(\partial_u F^\varepsilon(\Psi_t^\varepsilon)\right)^T$ and $S_t = L_t J^{-1}$. Exactly the same reasoning can be made as in *(i)*, therefore the t-flow Ψ_t^ε is symplectic, i.e. the averaged vector field F^ε is Hamiltonian.

(iv) This follows from *(ii)* and *(iii)* thanks to Remark I.4.6.

□

Remark I.4.8. *It may be of interest to note that in the linear autonomous case* $\partial_t u = \frac{1}{\varepsilon}Gu + Ku$*, property (i) of volume-preservation does not involve the dimension. Indeed differentiating the filtered change of variable* $\Phi_\theta = e^{-\theta G}e^{\theta G^\varepsilon}$ *and taking the average yields*

$$G^\varepsilon = \langle\Phi\rangle^{-1} G \langle\Phi\rangle.$$

In the homological equation $\partial_\theta \Omega_\theta = \varepsilon(K\Omega_\theta - \Omega_\theta K^\varepsilon)$*, we obtain*

$$K^\varepsilon = \langle\Phi\rangle^{-1} K \langle\Phi\rangle.$$

The involvement of the dimension in our proof actually seems purely technical, since the averaged vector field F^ε *can be expressed as a power series in* ε *which converges for* ε *small enough. Our result shows that every term of the series must be divergence-free, but the radius of convergence of the series* ε_0 *may be independent of the dimension of the space.*

I.5 Approximations on bounded domains

In this section we discuss what becomes of the previous results in actual applications, which is to say when the maps Φ^ε and F^ε of averaging are not known exactly, but only up to error terms of size ε^{n+1} for some order $n \in \mathbb{N}$. We also get rid of the assumption that the vector field $(\theta, u) \mapsto f_\theta(u)$ is uniformly bounded on the entire space X, and conduct our study on a possibly-bounded open subset $\mathcal{K} \subset X$.

I.5.1 Assumptions

For technical purposes, define $\mathcal{K}_\rho$ this subset extended by a radius of $\rho \geq 0$, i.e.

$$\mathcal{K}_\rho = \{u \in X \quad \text{s.t.} \quad \exists\, v \in \mathcal{K}, |u - v| \leq \rho\}\,.$$

We also define, given a map φ from $\mathcal{K}_\rho$ to some Banach space $(E, |\cdot|)$, the norm

$$\|\varphi\|_\rho = \sup_{u \in \mathcal{K}_\rho} |\varphi(u)|.$$

In particular for the vector fields and morphisms $E = X$, and for their derivatives $E = \mathcal{L}(E, E)$.

Assumption I.5.1. *The vector field $(\theta, u) \mapsto f_\theta(u)$ and its derivative are bounded (uniformly w.r.t. θ) on K_{3R} for some radius $R > 0$. There exist positive constants ε_0 and C such that for any rank $n \in \mathbb{N}$, there is a continuous near-identity 1-periodic change of variable $\Phi^{[n]}$ and a near-averaged vector field $F^{[n]}$, both well-defined on $\mathcal{K}_{3R}$ for $\varepsilon \leq \varepsilon_n := \varepsilon_0/(n+1)$. Precisely,*

$$\sup_{\theta\in\mathbb{T}} \|\Phi_\theta^{[n]} - \mathrm{id}\,\|_{3R} \leq \frac{\varepsilon}{\varepsilon_n} R \qquad \textit{and} \qquad \|F^{[n]}\|_{3R} \leq C.$$

Furthermore, the error of approximation is of size ε^{n+1}, i.e. writing $\Psi_t^{[n]}$ the t-flow of $F^{[n]}$,

$$u(t) = \Phi_{t/\varepsilon}^{[n]} \circ \Psi_t^{[n]} \circ \left(\Phi_0^{[n]}\right)^{-1}(u_0) + \mathcal{O}(\varepsilon^{n+1}) \tag{I.5.23}$$

until some time $T_R > 0$.

The error of approximation is caracterised by the defect $\delta^{[n]}$ defined by

$$\delta_\theta^{[n]} = \frac{1}{\varepsilon}\partial_\theta \Phi_\theta^{[n]} - f_\theta \circ \Phi_\theta^{[n]} + \partial_u \Phi_\theta^{[n]} \cdot F^{[n]},$$

which corresponds to the error in the homological equation (I.2.5). The previous assumptions corresponds to the situation

$$\sup_{\theta\in\mathbb{T}} \|\delta^{[n]}\|_{3R} = \mathcal{O}(\varepsilon^n) \qquad \text{and} \qquad \langle \delta^{[n]} \rangle = \mathcal{O}(\varepsilon^{n+1}). \tag{I.5.24}$$

This assumption matches the behaviour generally observed with averaging, found for instance in [CCMM15] when assuming $(\theta, u) \mapsto f_\theta(u)$ analytic w.r.t. u. As noted in [CMS15], this is enough to ensure the historical optimal "exponential" error bound of [Nei84], which can be stated as such : There is a positive constant c such that for all $\varepsilon > 0$ there is an integer n such that for all t,

$$\left| u(t) - \Phi_{t/\varepsilon}^{[n]} \circ \Psi_t^{[n]} \circ \left(\Phi_0^{[n]}\right)^{-1}(u_0) \right| \leq c e^{-c/\varepsilon}.$$

This reflects the fact that the maps Φ^ε and F^ε can only be obtained as diverging power series in ε, therefore the error is *formal*, up to a flat function. Indeed, in order to increase the order of the approximation, ε must get smaller and smaller, such that an error $\mathcal{O}(\varepsilon^\infty)$

is impossible with $\varepsilon \neq 0$.

Note furthermore that this assumption is enough to ensure that $\Phi_0^{[n]}$ and $\langle \Phi^{[n]} \rangle$ are invertible from $\mathcal{K}_\rho$ to $\mathcal{K}_{\rho+R}$ for any $\rho \in [0, 3R]$. Indeed for $u \in \mathcal{K}_\rho$, the map $\varphi(v) = u + v - \Phi_0^{[n]}(u+v)$ maps the closed ball of radius R onto itself,[2] thus admits a fixed point by Brouwer's fixed point theorem. Therefore there exists $u^* = u + v^* \in \mathcal{K}_{\rho+R}$ such that $u = \Phi_0^{[n]}(u^*)$. The same reasoning holds for $\langle \Phi^{[n]} \rangle$.

I.5.2 Autonomous case

Consider the autonomous problem (I.3.10) of Section I.3,

$$\dot{v}^\varepsilon = \frac{1}{\varepsilon} G(v^\varepsilon) + K(v^\varepsilon), \qquad v^\varepsilon(0) = v_0.$$

The flow of G, denoted $(\theta, u) \mapsto \chi_\theta(u)$, is assumed 1-periodic w.r.t. θ, and we assume that for every radius ρ, the set $\mathcal{K}_\rho$ is invariant by the flow of G. Performing averaging on this problem is equivalent to performing it on the filtered problem

$$\dot{u}^\varepsilon(t) = \Big(\partial_u \chi_{-t/\varepsilon} \cdot K\Big) \circ \chi_{t/\varepsilon}(u^\varepsilon(t)), \qquad u^\varepsilon(0) = v_0.$$

The unfiltered variable is obtained as $v^\varepsilon(t) = \chi_{t/\varepsilon}(u^\varepsilon(t))$. Given an approximation $v^\varepsilon(t) = \Omega_{t/\varepsilon}^{[n]} \circ \Psi_t^{[n]} \circ \Big(\Omega_0^{[n]}\Big)^{-1} + \mathcal{O}(\varepsilon^{n+1})$, an approximation on u^ε of the form (I.5.23) is obtained by setting $\Phi_\theta^{[n]} = \chi_{-\theta} \circ \Omega_\theta^{[n]}$. Conversely, it is also possible to obtain $\Omega^{[n]}$ from working on the filtered problem, and in the case where $u \mapsto G(u)$ is non-linear, this latter approach is generally more straightforward. The defect associated to averaging on the autonomous problem is

$$\eta_\theta^{[n]} := \frac{1}{\varepsilon}\Big(\partial_\theta \Omega_\theta^{[n]} - G \circ \Omega_\theta^{[n]}\Big) - K \circ \Omega_\theta^{[n]} + \partial_u \Omega_\theta^{[n]} K^{[n]} \tag{I.5.25}$$

and the link is made with the filtered averaging with the formula

$$\eta_\theta^{[n]} = \partial_u \chi_\theta\Big(\Phi_\theta^{[n]}\Big) \cdot \delta_\theta^{[n]}.$$

Theorem I.5.2 (Adaptation of Propositions I.3.1 and I.3.2)**.**
Given averaging maps $\Phi^{[n]}$ and $K^{[n]}$ which satisfy Assumption I.5.1 (with $F^{[n]}$ replaced by $K^{[n]}$) and such that the associated defect $\delta^{[n]}$ satisfies (I.5.24)*, define the change of variable $(\theta, u) \mapsto \Omega_\theta^{[n]}(u) = \chi_{-\theta} \circ \Phi_\theta^{[n]}(u)$ for autonomous averaging. With this definition,*

2. If $\varepsilon \leq \alpha \varepsilon_n$ for $\alpha \in (0, 1]$, then this radius becomes αR, therefore $\Phi_0^{[n]}$ injects $\mathcal{K}_\rho$ into $\mathcal{K}_{\rho+\alpha R}$.

$\Omega_\theta^{[n]}$ is the θ-flow of a vector field $G^{[n]}$ defined on $\mathcal{K}_R$ up to $\mathcal{O}(\varepsilon^{n+2})$. Furthermore, $G^{[n]}$ and $K^{[n]}$ commute up to $\mathcal{O}(\varepsilon^{n+2})$ on $\mathcal{K}_R$.

Proof. The first step of the proof is to show

$$K^{[n]} = \mathcal{A}^{\theta_1}[\Omega^{[n]}] + \mathcal{O}(\varepsilon^{n+1})$$

for all phases $\theta_1 \in \mathbb{T}$, with $\mathcal{A}^{\theta_1}$ the operator defined in (I.3.16). This result stems from the identity on $\widetilde{\Phi}_\theta^{[n]} = \chi_{-\theta-\theta_1} \circ \Omega_\theta^{[n]}$,

$$\partial_\theta \widetilde{\Phi}_\theta^{[n]} = \varepsilon\Big(f_{\theta+\theta_1} \circ \widetilde{\Phi}_\theta^{[n]} - \partial_u \widetilde{\Phi}_\theta^{[n]} \cdot K^{[n]}\Big) - \varepsilon \partial_u \chi_{-\theta-\theta_1}(\Omega_\theta^{[n]}) \cdot \eta_\theta^{[n]}.$$

Before taking the average, compute

$$\begin{aligned}
\partial_u \chi_{-\theta-\theta_1}(\Omega_\theta^{[n]}) \cdot \eta_\theta^{[n]} &= \partial_u \chi_{-\theta-\theta_1}\Big(\chi_\theta \Phi_\theta^{[n]}\Big) \partial_u \chi_\theta(\Phi_\theta^{[n]}) \delta_\theta^{[n]} \\
&= \partial_u (\chi_{-\theta-\theta_1} \circ \chi_\theta \circ \Phi_\theta^{[n]}) \Big(\partial_u \Phi_\theta^{[n]}\Big)^{-1} \delta_\theta^{[n]} \\
&= \partial_u \chi_{-\theta_1}(\Phi_\theta^{[n]}) \delta_\theta^{[n]}
\end{aligned}$$

Hence this term can be written as $\partial_u \chi_{-\theta_1}\Big(\mathrm{id} + \mathcal{O}(\varepsilon)\Big)\delta_\theta^{[n]}$, and its average is of size $\mathcal{O}(\varepsilon^{n+1})$ thanks to the assumption on $\delta^{[n]}$. Taking the average of the previous identity, we finally obtain

$$K^{[n]} = \mathcal{A}^{\theta_1}[\Omega^{[n]}] + \mathcal{O}(\varepsilon^{n+1}).$$

We then proceed in the same manner as for the proof of Proposition I.3.1. For some phase $\theta_0 \in \mathbb{T}$, consider the map $\widetilde{\Omega}_\theta^{[n]} = \Omega_{\theta+\theta_0}^{[n]} \circ \Big(\Omega_{\theta_0}^{[n]}\Big)^{-1}$ defined on $\mathcal{K}_R$. By definition of the defect, this new map satisfies the equation,

$$\partial_\theta \widetilde{\Omega}_\theta^{[n]} - G \circ \widetilde{\Omega}_\theta^{[n]} = \varepsilon\left(K \circ \widetilde{\Omega}_\theta^{[n]} - \partial_u \widetilde{\Omega}_\theta^{[n]} \cdot \widetilde{K}^{[n]}\right) - \varepsilon \widetilde{\eta}_\theta^{[n]}.$$

with $\widetilde{K}^{[n]} = \Big(\partial_u (\Omega_{\theta_0}^{[n]})^{-1}\Big)^{-1} \cdot K^{[n]} \circ (\Omega_{\theta_0}^{[n]})^{-1}$ and $\widetilde{\eta}_\theta^{[n]} = \eta_{\theta+\theta_0}^{[n]} \circ (\Omega_{\theta_0}^{[n]})^{-1}$. From (I.5.25), it appears in particular that $K^{[n]} = \mathcal{A}^{\theta_0}[\Omega^{[n]}] + \mathcal{O}(\varepsilon^{n+1}\Big)$. Injected into $\widetilde{K}^{[n]}$, this generates

$$\widetilde{K}^{[n]} = \Big(\partial_u \Big\langle \chi_{-\theta} \circ \widetilde{\Omega}_\theta^\varepsilon \Big\rangle\Big)^{-1} \Big\langle (\partial_u \chi_{-\theta} \cdot K) \circ \widetilde{\Omega}_\theta^\varepsilon \Big\rangle + \mathcal{O}(\varepsilon^{n+1}) = \mathcal{A}^0[\widetilde{\Omega}^\varepsilon] + \mathcal{O}(\varepsilon^{n+1}).$$

Hence $\Omega^{[n]}$ and $\widetilde{\Omega}^{[n]}$ satisfy the same equation up to a modification of the defect while still

respecting (I.5.24). In other words, we can replace $\Omega^{[n]}$ by $\widetilde{\Omega}^{[n]}$ in the following equation

$$\partial_\theta \Omega_\theta^{[n]} - G \circ \Omega_\theta^{[n]} = \varepsilon \left(K \circ \Omega_\theta^{[n]} - \partial_u \Omega_\theta^{[n]} \cdot \mathcal{A}^0[\Omega^{[n]}] \right) + \mathcal{O}(\varepsilon^{n+1})$$

without impacting the result. Since these two maps satisfy the same closure condition $\Omega_0^{[n]} = \mathrm{id} + \mathcal{O}(\varepsilon^{n+1})$, they differ by only $\mathcal{O}(\varepsilon^{n+1})$ at any phase $\theta \in \mathbb{T}$. We can finally define

$$G^{[n]} = \partial_\theta \Omega_\theta^{[n]} \Big|_{\theta=0}.$$

The second part of the theorem stems from the identity $K^{[n]} = \widetilde{K}^{[n]} + \mathcal{O}(\varepsilon^{n+1})$ which becomes

$$K^{[n]} \circ \Omega_{\theta_0}^{[n]} = \partial_u \Omega_{\theta_0}^{[n]} \cdot K^{[n]} + \mathcal{O}(\varepsilon^{n+1}).$$

□

Note that the exact flow of $G^{[n]}$ may not be 1-periodic depending on its definition. Think for instance of the one-dimensional converging example $G^{[n]} = i(1-\varepsilon)\sum_{k=0}^{n} \varepsilon^k = i(1-\varepsilon^{n+1})$.

I.5.3 Geometric properties

Here is what the preservation of geometric properties presented in Section I.4 becomes.

Theorem I.5.3 (Adaptation of Theorem I.4.7)**.**
Consider Assumption I.5.1 met and denote φ_t^ε the t-flow associated to Problem (I.1.1)*. Up to a reduction of ε_0, the following properties are satisfied up to an error of size $\mathcal{O}(\varepsilon^{n+1})$: if for all $t \in [0, T_R]$,*

(i) $u \mapsto \varphi_t^\varepsilon(u)$ is volume-preserving on $\in \mathcal{R}$, then $\Psi_t^{[n]}$ is volume-preserving on $\mathcal{K}_R$;

(ii) the functional I is preserved by φ_t^ε on K_{2R}, then it is preserved by $\Psi_t^{[n]}$ on $\mathcal{K}_R$;

(iii) φ_t^ε is symplectic on $\mathcal{K}_{2R}$, then $\Psi_t^{[n]}$ is symplectic on $\mathcal{K}_R$;

(iv) φ_t^ε is B-symplectic and preserves Casimirs on $\mathcal{K}_{2R}$, then $\Psi_t^{[n]}$ also does on $\mathcal{K}_R$.

Note that since $\Phi_\theta^{[n]} = \varphi_{\varepsilon\theta}^\varepsilon \circ \Psi_{\varepsilon\theta}^{[n]} + \mathcal{O}(\varepsilon^{n+1})$, these properties are also true for $\Phi_\theta^{[n]}$. It is therefore possible to modify $\Phi^{[n]}$ and $F^{[n]}$ and have these properties met exactly, altough the impact of this process on the well-posedness of the maps is unclear.

Proof. As can be seen in the proof of Theorem I.4.7, every property can be proven in the same way. Therefore we will only describe how to prove *(iii)*, as it is probably the

most interesting property for the majority of readers. We refer to the other proof for the adaptation to other properties.

Set $(t,u) \mapsto \Delta_t(u)$ the deviation from symplecticity,

$$\Delta_t = \big(\partial_u \Psi_t^{[n]}\big) J^{-1} \big(\partial_u \Psi_t^{[n]}\big)^T - J^{-1},$$

defined and bounded for $u \in \mathcal{K}_{R_n}$ Thanks the periodicity of $\Phi^{[n]}$, $t \mapsto \Delta_t$ is almost zero at stroboscopic times, meaning that for all $k \in \mathbb{N}$ such that $\varepsilon k \leq T_R$, since $\Psi_{\varepsilon k}^{[n]} = \varphi_{\varepsilon k}^{\varepsilon} + \mathcal{O}(\varepsilon^{n+1})$,

$$\Delta_{\varepsilon k} = \big(\partial_u \varphi_{\varepsilon k}^{\varepsilon}\big) J^{-1} \big(\partial_u \varphi_{\varepsilon k}^{\varepsilon}\big)^T - J^{-1} + \mathcal{O}(\varepsilon^{n+1}) = \mathcal{O}(\varepsilon^{n+1}).$$

Setting $L_t M = \partial_u F^{[n]}(\Psi_t^{[n]}) M + M \big(\partial_u F^{[n]}(\Psi_t^{[n]})\big)^T$ and $S_t = L_t J^{-1}$, it satisfies

$$\partial_u \Delta_t = L_t\, \Delta_t + S_t, \quad \text{i.e.} \quad \Delta_t = \Delta_0 + \int_0^t L_\tau\, \Delta_\tau \mathrm{d}\tau + \int_0^t S_\tau\, \mathrm{d}\tau. \tag{I.5.26}$$

We want to prove $\sup_{0\leq t\leq \varepsilon} \|\Delta_t\|_R = \mathcal{O}(\varepsilon^{n+1})$. To that effect, introduce the norm $\|\cdot\|_{\varepsilon,\rho}$ and the radii R_k,

$$\|g\|_{\varepsilon,\rho} = \sup_{0\leq t\leq \varepsilon} \|g_t\|_\rho \qquad \text{and} \qquad R_k = R + \frac{k}{n+1} R,$$

and set $\alpha > 0$ such that $\|\Delta_0\|_{2R}, \|\Delta_\varepsilon\|_{2R} \leq \alpha \varepsilon^{n+1}$. Gronwall's lemma in the integral form of Δ_t yields

$$\|\Delta\|_{\varepsilon,R} \leq \big(\alpha \varepsilon^{n+1} + \varepsilon \|S\|_{\varepsilon,R}\big) e^{\varepsilon \|L\|_{\varepsilon,R}}, \tag{I.5.27}$$

therefore we want to show $\|S\|_{\varepsilon,R} = \mathcal{O}(\varepsilon^{n+1})$. Because S is transported by $F^{[n]}$, i.e. $S_t = S_0 \circ \Psi_t^{[n]}$, it is possible to bound S_t on some space $\mathcal{K}_\rho$ by the norm of S_0 on a larger space. In particular, assuming $\varepsilon_0 \leq R/C$,

$$\|S\|_{\varepsilon,R_k} \leq \|S_0\|_{R_{k+1}} \tag{I.5.28}$$

since $\|\Psi_t^{[n]} - \mathrm{id}\,\|_{R_n} \leq tC$. This bound is fairly useful, as Taylor's theorem with integral remainder generates the identity

$$\Delta_\varepsilon = \Delta_0 + \varepsilon S + \int_0^\varepsilon (\varepsilon - t)\, \partial_t \Delta_t \mathrm{d}t,$$

from which a Cauchy inequality yields

$$\|S_0\|_{R_1} \leq \frac{\varepsilon}{2}\|\partial_t \Delta\|_{\varepsilon,R_1} + 2\alpha\varepsilon^n \leq \frac{\varepsilon}{2}\left(\|L\|_{\varepsilon,R_1}\,\|\Delta\|_{\varepsilon,R_1} + \|S\|_{\varepsilon,R_1}\right) + 2\alpha\varepsilon^n.$$

Injecting (I.5.27) and (I.5.28) into the right-hand term, we obtain

$$\|S_0\|_{R_1} \leq \frac{\varepsilon}{2}(1+C_L)\|S_0\|_{R_2} + (2+\varepsilon C_L)\alpha\varepsilon^n.$$

where $C_L = \varepsilon\|L_0\|_{2R}e^{\varepsilon\|L_0\|_{2R}}$ (exploiting the fact that L is transported by $F^{[n]}$). We set $\kappa = \frac{1}{2}(1+C_L)$ and $q = \alpha(2+\varepsilon C_L)$ for brevity, and successive applications of this reasoning on $\|S_0\|_{R_k}$ generate

$$\|S_0\|_{R_1} \leq (\varepsilon\kappa)^n\|S_0\|_{2R} + \sum_{k=0}^{n-1}(\varepsilon\kappa)q\varepsilon^n \leq \left(\frac{\varepsilon}{2\varepsilon_0}\right)^n \|S_0\|_{2R} + 2q\varepsilon^n$$

assuming $\varepsilon_0 \leq 1/(2\kappa)$. Finally, $\|S_0\|_{R_1} = \mathcal{O}(\varepsilon^{n+1})$ thus $\|\Delta\|_{\varepsilon,R} = \mathcal{O}(\varepsilon^{n+1})$. This reasoning can be conducted on any time interval of the form $[\varepsilon k, \varepsilon(k+1)]$, proving that R is of size $\left(\varepsilon^{n+1}\right)$ at all times.

□

Chapitre II

Convergence uniforme pour un problème dissipatif

Ce chapitre reprend un article à paraître dans *Mathematics of Computation*, intitulé

A uniformly accurate numerical method for a class of dissipative systems,

co-écrit avec mes directeurs, Philippe Chartier et Mohammed Lemou. Dans cet article, on construit un problème micro-macro pour une classe de problèmes à relaxation rapide, ce qui permet de résoudre le problème avec une précision uniforme d'ordre arbitraire. Le caractère bien posé de ce problème est prouvé en dressant un lien original avec les problèmes hautement oscillant pour lesquels des résultats existaient déjà.

II.1 Introduction

We are interested in problems of the form, for $x^\varepsilon(t) \in \mathbb{R}^{d_x}$ and $z^\varepsilon(t) \in \mathbb{R}^{d_z}$,

$$\begin{cases} \dot{x}^\varepsilon = a(x^\varepsilon, z^\varepsilon), & x^\varepsilon(0) = x_0, \\ \dot{z}^\varepsilon = -\dfrac{1}{\varepsilon} A z^\varepsilon + b(x^\varepsilon, z^\varepsilon), & z^\varepsilon(0) = z_0, \end{cases} \tag{II.1.1}$$

with $\varepsilon \in (0, 1]$ a small parameter, Λ a diagonal positive matrix with integer coefficients, and where a, b are respectively the x-component and the z-component of an analytic map f which smoothly depends on ε. We look for a solution $x^\varepsilon(t)$, $z^\varepsilon(t)$, defined for $t \in [0, 1]$, irrespectively of the value of ε. The exact value of the right bound of the interval of definition of the solution, here 1, is somehow arbitrary, as it can be rescaled by changing the value of $\frac{1}{\varepsilon}A$. In the limit when ε goes to zero, the problem becomes stiff on the considered interval : in other words, the problem resorts to long-time integration as 1 becomes large compared to ε. In the sequel we shall more often write the equations in

compact form as

$$\dot{u}^\varepsilon = -\frac{1}{\varepsilon} A u^\varepsilon + f(u^\varepsilon), \quad u^\varepsilon(0) = u_0, \tag{II.1.2}$$

where $u = \begin{pmatrix} x \\ z \end{pmatrix}$, $A = \begin{pmatrix} 0 & 0 \\ 0 & \Lambda \end{pmatrix}$ and $f(u) = \begin{pmatrix} a(x,z) \\ b(x,z) \end{pmatrix}$. We set $d = d_x + d_z$ the dimension of u such that $u \in \mathbb{R}^d$. Note that x^ε may be zero-dimensional without impacting our results, or that it may include a component $\tilde{x}(t) = t$ such that f depends on t in a "hidden" manner. In contrast, it should be emphasized that we do not address the case where the map $u \mapsto f(u)$ is a differential operator and u lies in a functional space : the theory required for that situation is outside the scope of our theorems. Nonetheless, two of our examples are discretized hyperbolic partial differential equations (PDEs) for which the method is successfully applied, even though an additional specific treatment is required.

Problems of the form (II.1.2) recurrently appear in population dynamics (see [GHM94 ; AP96 ; SAAP00 ; CCS18]), where Λ accounts for migration (in space and/or age) and a and b account for both the demographic and inter-population dynamics. In this context, the factor $1/\varepsilon$ accounts for the fact that the migration dynamics is quantifiably faster than other dynamics involved.

When solving this kind of system numerically, problems arise due to the large range of values that ε can take. To be more specific, the error for standard methods of order $q > 1$ behave like

$$E_\varepsilon(\Delta t) \leq \min\left(C_q \frac{\Delta t^q}{\varepsilon^r}, C_s \Delta t^s \right),$$

for some positive constants C_q and C_s independent of ε and integers $s \leq q$ and $r \geq 0$. This forces very small values of Δt in order to achieve some accuracy and causes the computational cost of the simulation to increase greatly, often prohibitively so. Additionally, the order is reduced to s in the sense that [1]

$$\sup_{\varepsilon \in (0,1]} E_\varepsilon(\Delta t) \leq C \Delta t^s. \tag{II.1.3}$$

This behaviour is documented for instance in [HW96, Section IV.15] or in [HR07]. In order to ensure a given error bound, one must either accept this order reduction (if $s > 0$), as is done for asymptotic-preserving (AP) schemes [Jin99] by taking a modified time-step $\tilde{\Delta} t = \Delta t^{q/s}$, or use an ε-dependent time-step $\Delta t = \mathcal{O}(\varepsilon^{r/q})$.

1. In particular, the scheme cannot be any usual explicit scheme since it would require a stability condition of the form $\Delta t/\varepsilon < C$ with C independent of ε.

A common approach to circumvent this difficulty is to invoke the *center manifold theorem* (see [Vas63 ; Car82 ; Sak90]), which dictates the long-time behaviour of the system and presents useful characteristics for numerical simulations : the dimension of the system is reduced and the dynamics on the manifold is non-stiff. However, this approach does not allow to capture the *transient phase* of the solution, i.e. the solution in short time before it reaches the stable manifold. Insofar as one wishes to describe the system out of equilibrium, this is clearly unsatisfactory. Furthermore, even if the solution is exponentially (w.r.t. time) close to the manifold, the center manifold approximation is accurate up to a certain error $\mathcal{O}(\varepsilon^n)$, rendering it useless if ε is of the order of 1.

The strategy developed in this paper is based on a *micro-macro* decomposition of the problem in combination with the use of standard q^{th}-order *exponential Runge-Kutta* methods. It aims at deriving an overall scheme with an error $E_\varepsilon(\Delta t)$ that can be bounded from above independently of ε, that is to say

$$E_\varepsilon(\Delta t) \leq C\Delta t^q$$

for some positive constant C independent of ε. In order to construct the appropriate transformation of the original system, we first provide a systematic way to compute asymptotic models at any order in ε approaching the solution over the *whole interval of time*. We then use the defect of this approximation to compute the solution with usual explicit numerical schemes and *uniform* accuracy (i.e. the cost and error of the scheme must be independent of ε). This approach automatically overcomes the challenges posed by both extremes $\varepsilon \ll 1$ and $\varepsilon \sim 1$.

The aforementioned micro-macro decomposition is obtained by writing the solution u^ε of (II.1.2) as the following composition of maps

$$u^\varepsilon(t) = \Omega^\varepsilon_{t/\varepsilon} \circ \Gamma^\varepsilon_t \circ \left(\Omega^\varepsilon_0\right)^{-1}(u_0) \tag{II.1.4}$$

where $(\tau, u) \in \mathbb{R}_+ \times \mathbb{R}^d \mapsto \Omega^\varepsilon_\tau(u) \in \mathbb{R}^d$ is a change of variable ε-close to the map $(\tau, u) \mapsto e^{-\tau A}u$ and where $(t, u) \in [0, T] \times \mathbb{R}^d \mapsto \Gamma^\varepsilon_t(u)$ is the flow associated to a *non-stiff* autonomous vector field $u \mapsto F^\varepsilon(u)$, yet to be defined. The formal maps Ω^ε and F^ε are approached at an arbitrary order $n \in \mathbb{N}$ by $\Omega^{[n]}$ and $F^{[n]}$ respectively such that the equality

$$u^\varepsilon(t) = \Omega^{[n]}_{t/\varepsilon}\left(v^{[n]}(t)\right) + w^{[n]}(t) \tag{II.1.5}$$

holds true, where $v^{[n]}(t) = \Gamma_t^{[n]} \circ \left(\Omega_0^{[n]}\right)^{-1}(u_0)$ and $w^{[n]}$ are respectively called the *macro* component and the *micro* component. A crucial feature of this decomposition is that $w^{[n]}$ remains of size $\mathcal{O}(\varepsilon^{n+1})$.

Now, the main contribution of this work is to prove that, using explicit exponential Runge-Kutta (ERK) schemes of order $n+1$ (which can be found for instance in [HO05]), it is possible to approximate u^ε with *uniform accuracy* and at *uniform computational cost* with respect to ε. In other words, we prove that formula (II.1.3) holds with $s = q = n+1$ and $r = 0$. More precisely, if $(t_i)_{0\leq i\leq N}$ is a time-step grid of mesh-size Δt, and if (v_i) and (w_i) are computed numerically by applying the ERK method to the micro-macro decomposition, then there exists C *independent of* ε such that ($|\cdot|$ stands for the usual Euclidian norm)

$$\max_{0\leq i\leq N}\left\{\left|x^\varepsilon(t_i) - x_i\right| + \frac{1}{\varepsilon}\left|z^\varepsilon(t_i) - z_i\right|\right\} \leq C\Delta t^{n+1} \quad \text{with} \quad \begin{pmatrix} x_i \\ z_i \end{pmatrix} = \Omega^{[n]}_{t_i/\varepsilon}(v_i) + w_i.$$

We emphasize here the expected occurrence of the scaling factor $1/\varepsilon$ accounts for the fact that z becomes of size $\mathcal{O}(\varepsilon)$ after a time $\mathcal{O}(\varepsilon \log(1/\varepsilon))$. IMEX methods such as CNLF and SBDF (see [ARW95 ; ACM99 ; HS21]), which mix implicit and explicit parts are not the focus of the article, but their use is briefly discussed in Remark II.2.9.

The present work is related to the recent paper [CCS16], where asymptotic expansions of the solution of (II.1.1) are constructed for the special case where Λ is the identity matrix. The theory developed therein is however of no relevance for the construction of micro-macro decompositions as it relies heavily on trees and associated elementary differentials which can hardly be computed in practice. Our approach actually shares more similarities with the one introduced for highly-oscillatory problems in [CLMV20] and later modified to become amenable for actual computations at any order [CLMZ20]. As a matter of fact, the technical arguments that sustain decomposition (II.1.4) are essentially adapted from [CCMM15] in a way that will be fully explained in Section II.3.

The rest of the paper is organized as follows. In Section II.2, we show our method to construct a micro-macro problem up to any order, and state our main result, i.e that solving this micro-macro problem with ERK schemes generates uniform accuracy on u^ε. In Section II.3, we give proofs of all the results from Section II.2. In Section II.4, we present some techniques to adapt our method to discretized hyperbolic PDEs. Namely, we study a

relaxed conservation law and the telegraph equation, which can be respectively found for instance in [JX95] and [LM08]. In Section II.5, we verify our theoretical result of uniform accuracy by successfully obtaining uniform convergence when numerically solving micro-macro problems obtained from a toy ODE and from the two aforementioned discretized PDEs.

II.2 Uniform accuracy from a decomposition

We start by considering the solution u of

$$\partial_t u^\varepsilon = -\frac{1}{\varepsilon} A u^\varepsilon + f(u^\varepsilon), \qquad u^\varepsilon(0) = u_0 \in \mathbb{R}^d, \tag{II.2.6}$$

and write it as the composition of a *non-stiff* flow $(t,u) \mapsto \Gamma_t^\varepsilon(u)$ with a change of variable $(\tau,u) \mapsto \Omega_\tau^\varepsilon(u)$ with $\tau \in \mathbb{R}_+$,

$$u^\varepsilon(t) = \Omega_{t/\varepsilon}^\varepsilon \circ \Gamma_t^\varepsilon \circ \left(\Omega_0^\varepsilon\right)^{-1}(u_0). \tag{II.2.7}$$

In order for our approach to be rigorous, we start by introducing some definitions and assumptions in Subsection II.2.1. We then present a way to approach these maps at any rank $n \in \mathbb{N}$ by $\Gamma^{[n]}$ and $\Omega^{[n]}$ in Subsection II.2.2. This approximation is such that the error in (II.2.7) is of size $\mathcal{O}(\varepsilon^{n+1})$. In Subsection II.2.3, we use this approximation to construct a micro-macro problem which can be solved numerically using standard IMEX schemes. This leads to our main result : reconstructing the solution u^ε of (II.2.6) from the numerical solution of the micro-macro problem yields an error *independent of* ε on u^ε. All proofs are delayed until Section II.3.

II.2.1 Definitions and assumptions

Before proceeding, we must first state the assumptions on the vector field $u \mapsto f(u)$ and the operator A.

Assumption II.2.1. *The matrix A is diagonal with nonnegative integer eigenvalues, and these values are nondecreasing when following the diagonal. In other words, $A = \mathrm{Diag}(\lambda_1, \dots, \lambda_d)$ with $(\lambda_i)_{1\le i\le d} \in \mathbb{N}^d$ and $\lambda_1 \le \dots \le \lambda_d$.*

Thanks to this assumption, we write $u = \begin{pmatrix} x \\ z \end{pmatrix}$, with (x,z) such that $Au = \begin{pmatrix} 0 \\ \Lambda z \end{pmatrix}$

for some Λ positive definite. The dimension of z may be zero without making our results invalid.

Assumption II.2.2. *Let us set d_x and d_z the dimensions of x and z respectively. There exists a compact set $X_1 \subset \mathbb{R}^{d_x}$ and a radius $\check{\rho} > 0$ such that for every x in X_1, the map $u \in \mathbb{R}^d \mapsto f(u) \in \mathbb{R}^d$ can be developed as a Taylor series around $\begin{pmatrix} x \\ 0 \end{pmatrix}$, and the series converges with a radius not smaller than $\check{\rho}$.*

It is therefore possible to naturally extend f to compact subsets of $\mathbb{C}^d$ defined by

$$\mathcal{U}_\rho := \left\{ u \in \mathbb{C}^d \,;\, \exists x \in X_1,\, \left| u - \begin{pmatrix} x \\ 0_{d_z} \end{pmatrix} \right| \leq \rho \right\},$$

for all $0 \leq \rho < \check{\rho}$ as it is represented by a Taylor series in $u \in \mathbb{C}^d$ on these sets. Here $|\cdot|$ is the natural extension of the Euclidian norm on $\mathbb{R}^d$ to $\mathbb{C}^d$.

It may seem particularly restrictive to assume that the z-component of the solution u^ε of (II.1.2) stays in a neighborhood of 0, however this is somewhat ensured by the *center manifold theorem*. This theorem states that there exists a map $x \in \mathbb{R}^{d_x} \mapsto \varepsilon h^\varepsilon(x) \in \mathbb{R}^{d_x}$ smooth in ε and x, such that the manifold $\mathcal{M}$ defined by

$$\mathcal{M} = \left\{ (x, z) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_z} \;:\; z = \varepsilon h^\varepsilon(x) \right\}$$

is a stable invariant for (II.1.1). It also states that all solutions $(x^\varepsilon, z^\varepsilon)$ of (II.1.1) converge towards it exponentially quickly, i.e. there exists $\mu > 0$ independent of ε such that

$$|z^\varepsilon(t) - \varepsilon h^\varepsilon(x^\varepsilon(t))| \leq C e^{-\mu t/\varepsilon}. \tag{II.2.8}$$

This means that the growth of z^ε is bounded by that of x^ε, and that after a time $t \geq \varepsilon \log(1/\varepsilon)$, $z^\varepsilon(t)$ is of size $\mathcal{O}(\varepsilon)$. Therefore it is credible to assume that z^ε stays somewhat close to 0. This is translated into a final assumption.

Assumption II.2.3. *There exist two radii $0 < \rho_0 \leq \rho_1 < \check{\rho}$ and a closed subset $X_0 \subset X_1 \subset \mathbb{R}^{d_x}$ such that the initial condition $u_0 \in \mathbb{C}^d$ satisfies*

$$\min_{x \in X_0} \left| u_0 - \begin{pmatrix} x \\ 0_{d_z} \end{pmatrix} \right| \leq \rho_0,$$

and for all $\varepsilon \in (0, 1]$, Problem (II.2.6) *is well-posed on $[0, 1]$ with its solution u^ε in $\mathcal{U}_{\rho_1}$.*

Note that this is different to assuming that the initial data (x_0, z_0) is close to the center manifold. Indeed, the size of the initial condition is supposed independent of ε, therefore the distance from $z(0)$ to the center manifold is always $\mathcal{O}(1)$.

For $\rho \in [0, \check{\rho} - \rho_1)$, we define the sets

$$\mathcal{K}_\rho := \mathcal{U}_{\rho_1+\rho} = \left\{ u \in \mathbb{C}^d \,;\, \exists x \in X_1, \left| u - \begin{pmatrix} x \\ 0 \end{pmatrix} \right| \leq \rho_1 + \rho \right\} \tag{II.2.9}$$

which help quantify the distance to the solution u^ε. By Assumption II.2.3, the solution of (II.1.2) is in $\mathcal{K}_0$ at all time.

Definition II.2.4. *We introduce some technical constants :*

(i) A radius $0 < R < \frac{1}{2}(\check{\rho} - \rho_1)$

(ii) An arbitrary rank p *and a positive constant* M *such that for all* $0 \leq \alpha, \beta \leq p+2$ *and all* $\sigma \in [0, 6|||A|||]$,

$$\frac{\sigma^\beta}{\beta!} \left| \left| \left| (\rho_1 + 2R)^\alpha \partial_u^\alpha f \right| \right| \right| \leq M$$

Given a radius $0 \leq \rho \leq 2R$ and a map $(\tau, u) \in \mathbb{R}_+ \times \mathcal{K}_\rho \mapsto \psi_\tau(u)$, we define the norm,

$$\|\psi\|_\rho := \sup_{(\tau,u)\in\mathbb{R}_+\times\mathcal{K}_\rho} |\psi_\tau(u)|. \tag{II.2.10}$$

If the map is furthermore p-times continuously differentiable w.r.t. τ, then we define

$$\|\psi\|_{\rho,p} := \max_{0\leq\nu\leq p} \|\partial_\tau \psi\|_\rho. \tag{II.2.11}$$

II.2.2 Constructing the micro-macro problem

We assume that the vector field in (II.2.7) follows an autonomous vector field F^ε, i.e.

$$\frac{\mathrm{d}}{\mathrm{d}t} \Gamma_t^\varepsilon(u) = F^\varepsilon\Big(\Gamma_t^\varepsilon(u)\Big). \tag{II.2.12}$$

Injecting this and (II.2.7) into (II.2.6) and writing $v_0 = \Big(\Omega_0^\varepsilon\Big)^{-1}(u_0)$

$$\Big(\partial_\tau + A\Big)\Omega_{t/\varepsilon}^\varepsilon\Big(\Gamma_t^\varepsilon(v_0)\Big) = \varepsilon\Big(f \circ \Omega_{t/\varepsilon}^\varepsilon\Big(\Gamma_t^\varepsilon(v_0)\Big) - \partial_u \Omega_{t/\varepsilon}^\varepsilon\Big(\Gamma_t^\varepsilon(v_0)\Big) \cdot F^\varepsilon\Big(\Gamma_t^\varepsilon(v_0)\Big)\Big)$$

which by separation of scales t and t/ε generates the homological equation on Ω^ε, for all $(\tau, u) \in \mathbb{R}_+ \times K_\rho$,

$$\Big(\partial_\tau + A\Big)\Omega^\varepsilon_\tau(u) = \varepsilon\Big(f \circ \Omega^\varepsilon_\tau(u) - \partial_u \Omega^\varepsilon_\tau(u) \cdot F^\varepsilon(u)\Big). \tag{II.2.13}$$

It is furthermore possible to extract the vector field F^ε from this equation to get

$$F^\varepsilon = \Big\langle \partial_u \Omega^\varepsilon \Big\rangle^{-1} \langle f \circ \Omega^\varepsilon \rangle \tag{II.2.14}$$

where $\langle \cdot \rangle$ is defined by the following formula

$$\langle \psi \rangle := \frac{1}{2\pi} \int_0^{2\pi} e^{i\theta A} \psi_{i\theta} \, \mathrm{d}\theta, \tag{II.2.15}$$

with the canonical definition $\psi_{i\theta} = \sum_{k\geq 0} e^{-ik\theta} \hat{\psi}_k$. To see this, we first observe that for an exponential series $\tau \in \mathbb{R}_+ \mapsto \psi_\tau$ which converges absolutely for $\tau = 0$, i.e. $\psi_\tau = \sum_{k\geq 0} e^{-k\tau} \hat{\psi}_k$ with $\sum_k \hat{\psi}_k$ absolutely converging, we can extract the coefficient $\hat{\psi}_k$ as the Fourier coefficient of $\psi_{i\theta}$ according to

$$\hat{\psi}_k = \frac{1}{2\pi} \int_0^{2\pi} e^{ik\theta} \psi_{i\theta} \, \mathrm{d}\theta. \tag{II.2.16}$$

Therefore, we write equation (II.2.13) as follows

$$\partial_\tau \Big(e^{\tau A} \Omega^\varepsilon_\tau\Big)(u) = \varepsilon\Big(e^{\tau A} f \circ \Omega^\varepsilon_\tau(u) - e^{\tau A} \partial_u \Omega^\varepsilon_\tau(u) \cdot F^\varepsilon(u)\Big), \tag{II.2.17}$$

and apply the Fourier operator (II.2.16) to get

$$\widehat{\partial_\tau \Big(e^{\tau A} \Omega^\varepsilon_\tau\Big)(u)}_k = \varepsilon \left(\widehat{\Big(e^{\tau A} f \circ \Omega^\varepsilon_\tau(u)\Big)}_k - \widehat{\Big(\partial_u \Omega^\varepsilon_\tau(u) \cdot F^\varepsilon(u)\Big)}_k \right).$$

Taking now $k = 0$ and using definition (II.2.15) we get the expression (II.2.14). This framework of exponential series comes naturally thanks to Assumption II.2.1.

The homological equation (II.2.13) has no unique solution in general, however we can approximate a solution as a *formal* solution as a power series in ε. This is generally the idea behind *normal forms*, where different methods have been developed (see [Mur06] for instance). Here we only consider a basic method to compute approximations $\Omega^{[n]}$ and $F^{[n]}$

of Ω^ε and F^ε at any rank $n \in \mathbb{N}$ by setting

$$(\partial_\tau + A)\Omega_\tau^{[n+1]} = \varepsilon\Big(f \circ \Omega_\tau^{[n]} - \partial_u \Omega_\tau^{[n]} \cdot F^{[n]}\Big). \tag{II.2.18}$$

with initial condition $\Omega_\tau^{[0]} = e^{-\tau A}$. Because we want $\Omega^{[n+1]}$ to be an exponential series, it appears that necessarily,

$$F^{[n]} = \Big\langle \partial_u \Omega^{[n]} \Big\rangle^{-1} \langle f \circ \Omega^{[n]} \rangle. \tag{II.2.19}$$

However these equations alone are not enough to obtain $\Omega^{[n]}$ at any order. Indeed, from (II.2.18), one gets

$$\Omega_\tau^{[n+1]} = e^{-\tau A}\Omega_0^{[n+1]} + \varepsilon \int_0^\tau e^{(\sigma-\tau)A}\Big(f \circ \Omega_\sigma^{[n]} - \partial_u \Omega_\sigma^{[n]} \cdot F^{[n]}\Big)\mathrm{d}\sigma \tag{II.2.20}$$

meaning a choice of initial data $\Omega_0^{[n+1]}$ is needed. One could think that choosing $\Omega_0^{[n+1]} = \mathrm{id}$ is the easiest choice, but computing (II.2.19) requires an inversion of $\langle \partial_u \Omega^\varepsilon \rangle$. Therefore we choose $\Omega_0^{[n+1]}$ such that $\langle \Omega^{[n+1]} \rangle = \mathrm{id}$, i.e. for all $n \in \mathbb{N}$,

$$\Omega_0^{[n+1]} = \mathrm{id} - \varepsilon \left\langle \int_0^\bullet e^{(\sigma-\bullet)A}\Big(f \circ \Omega_\sigma^{[n]} - \partial_u \Omega_\sigma^{[n]} \cdot F^{[n]}\Big)\mathrm{d}\sigma \right\rangle \qquad \text{thus} \quad F^{[n]} = \langle f \circ \Omega^{[n]} \rangle. \tag{II.2.21}$$

Now that we have a way to compute an approximate solution of (II.2.13), we introduce the error of approximation

$$\eta_\tau^{[n]} = \frac{1}{\varepsilon}\left(\partial_\tau + A\right)\Omega_\tau^{[n]} + \partial_u \Omega_\tau^{[n]} \cdot F^{[n]} - f \circ \Omega^{[n]}. \tag{II.2.22}$$

With these definitions, the maps $(\tau, u) \mapsto \Omega_\tau^{[n]}(u)$, $u \mapsto F^{[n]}(u)$ and $(\tau, u) \mapsto \eta_\tau^{[n]}$ have the following properties.

Theorem II.2.5. *For n in $\mathbb{N}$, let us denote $r_n = R/(n+1)$ and $\varepsilon_n := r_n/16M$ with R and M from Definition II.2.4. For all $\varepsilon > 0$ such that $\varepsilon \leq \varepsilon_n$, the maps $(\tau, u) \mapsto \Omega_\tau^{[n]}(u)$, $u \mapsto F^{[n]}(u)$ and $(\tau, u) \mapsto \eta_\tau^{[n]}(u)$ given by* (II.2.20) *and* (II.2.21) *are well-defined on $\mathbb{R}_+ \times \mathcal{K}_R$ and are analytic w.r.t. u. The change of variable $\Omega^{[n]}$ and the residue $\eta^{[n]}$ are both $p+1$-times continuously differentiable w.r.t. τ. Moreover, with $\|\cdot\|_R$ and $\|\cdot\|_{R,p+1}$*

given by (II.2.10) *and* (II.2.11)*, the following bounds are satisfied for all* $0 \leq \nu \leq p+1$*,*

$$(i) \quad \left\|\Omega^{[n]} - e^{-\tau A}\right\|_R \leq 4\varepsilon M, \qquad (ii) \quad \left\|\partial_\theta^\nu \left[\Omega^{[n]} - e^{-\tau A}\right]\right\|_R \leq 8\Big(1 + |||A|||\Big)^\nu \varepsilon M\, \nu!$$
$$(iii) \quad \|F^{[n]}\|_R \leq 2M \qquad (iv) \quad \|\eta_\tau^{[n]}(u)\|_{R,p} \leq 2M\Big(1 + |||A|||\Big)^p \left(2\mathcal{Q}_p \frac{\varepsilon}{\varepsilon_n}\right)^n$$

where $|||\cdot|||$ *is the induced norm from* $\mathbb{R}^d$ *to* $\mathbb{R}^d$*, and* $\mathcal{Q}_p$ *is a* p*-dependent constant.*

The proof will be treated in Subsection II.3.1, and this results remains valid with the choice $\Omega_0^{[n]} = \mathrm{id}$.

II.2.3 A result of uniform accuracy

Given a rank $n \in \mathbb{N}$, we now denote $v^{[n]}(t) := \Gamma_t^{[n]} \circ \left(\Omega_0^{[n]}\right)^{-1}(u_0)$ and inject the decomposition

$$u^\varepsilon(t) = \Omega_{t/\varepsilon}^{[n]}\Big(v^{[n]}(t)\Big) + w^{[n]}(t) \tag{II.2.23}$$

into Problem (II.2.6) in order to find an equation on $w^{[n]}$. The main interests of this decomposition can be roughly summarized as follows. First, the change of variable $\Omega_{t/\varepsilon}^{[n]}$ is known explicitly and the macro solution $v^{[n]}$ is smooth in ε, in the sense that time derivatives of $v^{[n]}$ at any order are uniformly bounded with respect to $\varepsilon \in (0,1]$. Second, the micro part $w^{[n]}$ is less stiff than the original solution u^ε in the sense that its time derivatives, up to order $n+1$, are uniformly bounded in ε. These important properties naturally allow the construction of numerical schemes on $v^{[n]}$ and $w^{[n]}$ that enjoy the *uniform accuracy*, i.e. in which the order of the numerical methods is independent of ε and is not degraded by the stiffness generated by the possibly small values of ε.

From decomposition (II.2.23) we obtain the following system

$$\begin{cases} \partial_t v^{[n]}(t) = F^{[n]}(v^{[n]}), \\ \partial_t w^{[n]}(t) = -\dfrac{1}{\varepsilon} A\left(\Omega_{t/\varepsilon}^{[n]}(v^{[n]}) + w^{[n]}\right) + f\left(\Omega_{t/\varepsilon}^{[n]}(v^{[n]}) + w^{[n]}\right) - \dfrac{\mathrm{d}}{\mathrm{d}t}\Omega_{t/\varepsilon}^{[n]}(v^{[n]}), \end{cases}$$

with initial conditions $v^{[n]}(0) = \left(\Omega_0^{[n]}\right)^{-1}(u_0)$ and $w^{[n]}(0) = 0$. By definition of $v^{[n]}$ and

using (II.2.22),

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\Omega^{[n]}_{t/\varepsilon}(v^{[n]}(t)) &= \frac{1}{\varepsilon}\partial_\tau\Omega^{[n]}_{t/\varepsilon}(v^{[n]}) + \partial_u\Omega^{[n]}_{t/\varepsilon}(v^{[n]})\cdot F^{[n]}(v^{[n]}) \\ &= -\frac{1}{\varepsilon}A\Omega^{[n]}_{t/\varepsilon}(v^{[n]}) + \eta^{[n]}_{t/\varepsilon}(v^{[n]}) + f\Big(\Omega^{[n]}_{t/\varepsilon}(v^{[n]})\Big).\end{aligned}$$

We get the micro-macro problem

$$\begin{cases}\partial_t v^{[n]}(t) = F^{[n]}(v^{[n]}), & \text{(II.2.24a)}\\ \partial_t w^{[n]}(t) = -\dfrac{1}{\varepsilon}Aw^{[n]} + f\left(\Omega^{[n]}_{t/\varepsilon}(v^{[n]}) + w^{[n]}\right) - f\left(\Omega^{[n]}_{t/\varepsilon}(v^{[n]})\right) - \eta^{[n]}_{t/\varepsilon}(v^{[n]}). & \text{(II.2.24b)}\end{cases}$$

with initial conditions $v^{[n]}(0) = \left(\Omega^{[n]}_0\right)^{-1}(u_0)$, $w^{[n]}(0) = 0$. The properties of this micro-macro problem can be summed up as followed.

Theorem II.2.6. *For all $n \in \mathbb{N}^*$, let us define $r_n = R/n$ and $\varepsilon_n := r_n/16M$, with R and M from Definition II.2.4. For all $\varepsilon \leq \varepsilon_n$, Problem* (II.2.24) *is well-posed until some final time T_n independent of ε, and the following bounds are satisfied for all $t \in [0, T_n]$ and $0 \leq \nu \leq \min(n, p)$,*

$$\begin{aligned}&(i)\quad v^{[n]}(t) \in \mathcal{K}_R && (ii)\quad |w^{[n]}(t)| \leq \frac{R}{4}\left(\frac{\varepsilon}{\varepsilon_n}\right)^{n+1}\\ &(iii)\quad |\partial_t^\nu E^{[n]}(t)| = \mathcal{O}(\varepsilon^{n-\nu}) && (iv)\quad \|\partial_t^{\nu+1} E^{[n]}\|_{L^1} = \mathcal{O}(\varepsilon^{n-\nu})\end{aligned}$$

where $E^{[n]} = \partial_t w^{[n]} + \frac{1}{\varepsilon}Aw^{[n]}$.

Remark II.2.7. *The attentive reader may notice that, while we made the computation of $F^{[n]}$ easy with* (II.2.21)*, the initial condition of the macro part, $v^{[n]}(0) = \left(\Omega^{[n]}_0\right)^{-1}(u_0)$, is not explicit. However, this system must be solved only once, while $F^{[n]}$ is used at every time-step. Furthermore, it is possible to compute an approximation of $v^{[n]}(0)$ explicitely up to $\mathcal{O}(\varepsilon^{n+1})$ using*[2]

$$v^{[n+1]}(0) = u_0 - \left(\Omega^{[n+1]}_0 - \mathrm{id}\right)\left(v^{[n]}(0)\right) + \mathcal{O}(\varepsilon^{n+2}) \qquad \text{(II.2.25)}$$

with initialization $v^{[0]}(0) = u_0$. Because $\Omega^{[n+1]}_0$ is near-identity (up to $\mathcal{O}(\varepsilon)$), an error of

2. The above formula is a consequence of the behaviour of the error, $\Omega^{[n+1]} = \Omega^{[n]} + \mathcal{O}(\varepsilon^{n+1})$ (see [CLMV20]), therefore $v^{[n+1]}(0) = v^{[n]}(0) + \mathcal{O}(\varepsilon^{n+1})$. Injecting this last approximation in $v^{[n+1]}(0) = u_0 - \big(\Omega^{[n+1]} - \mathrm{id}\big)(v^{[n+1]}(0))$ generates the formula.

size ε^{n+1} *on* $v^{[n]}(0)$ *will only translate in an error of size* ε^{n+2} *on* $v^{[n+1]}(0)$.

We can now define approached initial conditions for the micro-macro problem iterating (II.2.25) *at each rank* n *and truncating the* $\mathcal{O}(\varepsilon^{n+2})$ *term. The initial condition of the micro part becomes*

$$w^{[n]}(0) = u_0 - \Omega_0^{[n]}\Big(v_n\Big) \tag{II.2.26}$$

which ensures $w^{[n]}(0) = \mathcal{O}(\varepsilon^{n+1})$*, meaning our results are not jeopardised.*

Using a standard explicit scheme to solve Problem (II.2.24) cannot work due to the term $\frac{1}{\varepsilon}Aw^{[n]}$. This is why we focus on exponential schemes, which render this term non-problematic in terms of stability (see [MZ09]). Of course, the only use of these exponential schemes does not solve the problem of non-uniform order of accuracy however, as these schemes all reduce to order 1 when taking the supremum of the error for $\varepsilon \in (0, \varepsilon^*]$. This is where our micr-macro formulation plays a crucial role since it allows standard numerical schemes (like exponential Runge-Kutta schemes for instance) to *keep their order uniformly* in $\varepsilon \in (0, 1]$. It should be noted that exponential schemes are well-established and the formulas to implement them can be found for example in [HO05] up to the fourth-order.

The first-order Euler method applied to (II.1.2) would yield

$$u_{i+1} = e^{-\frac{\Delta t}{\varepsilon}A}u_i + \Delta t\ \varphi\left(-\frac{\Delta t}{\varepsilon}A\right) f(u_i)$$

with $\varphi(-hA) = \frac{1}{h}\int_0^h e^{-sA}\mathrm{d}s$. Because A is diagonal, this type of integral is easy to compute. There is no computational drawback to exponential schemes in this case. Furthermore, for these schemes the error bound involves the "modified" norm

$$|u|_\varepsilon = \left|u + \frac{1}{\varepsilon}Au\right|. \tag{II.2.27}$$

This norm is interesting because after a short time $t \geq \varepsilon\log(1/\varepsilon)$, the z-component of the solution u^ε of (II.1.2) is of size ε, as evidenced by the center manifold theorem in (II.2.8). Using the norm $|\cdot|_\varepsilon$ somewhat rescales z^ε (but not x^ε) by ε^{-1} such that studying the error in this norm can be seen as a sort of "relative" error.

The following result asserts that, indeed, our micro-macro reformulation of the problem allows any numerical scheme of order p, namely exponentiel schemes, to enjoy the uniform accuracy property, with the same order p. A detailed presentation of exponential Runge-Kutta schemes can be found for instance in [HO05 ; HO04].

Theorem II.2.8. *Under the assumptions of Theorem II.2.6 and denoting $T_n \leq T$ a final time such that Problem* (II.2.24) *is well-posed on $[0, T_n]$. Given $(t_i)_{i\in[\![0,N]\!]}$ a discretisation of $[0, T_n]$ of time-step $\Delta t := \max_i |\, t_{i+1} - t_i \,|$. computing an approximate solution (v_i, w_i) of* (II.2.24) *using an exponential Runge-Kutta scheme of order $q := \min(n, p) + 1$ yields a uniform error of order q, i.e.*

$$\max_{0\leq i\leq N} \left| u^\varepsilon(t_i) - \Omega^{[n]}_{t_i/\varepsilon}(v_i) - w_i \right|_\varepsilon \leq C\Delta t^q \tag{II.2.28}$$

where C is independent of ε.

The left-hand side of this inequality involves $|\cdot|_\varepsilon$ and shall be called the modified error. It dominates the absolute error which uses $|\cdot|$.

Remark II.2.9. *Only exponential schemes are considered here rather than for instance IMEX-BDF schemes which are sometimes preferred (as in [HS21]). The reason for this is twofold.*

First, as was mentioned already, iterations are easy to compute because of the diagonal nature of A. Second, the error bounds are generally better for these schemes. Indeed, an IMEX-BDF scheme of order q involves the L^1 norm of $\partial_t^{q+1} w^{[n]}$, which is worse than the L^1 norm of $\partial_t^q E^{[n]}$. The former is of size $\mathcal{O}(\varepsilon^{n-q})$ while the latter is of size $\mathcal{O}(\varepsilon^{n+1-q})$. We made the choice to prioritize methods of order $n+1$ rather than n.

II.3 Proofs of theorems from Section II.2

II.3.1 Proof of Theorem II.2.5 : properties of the decomposition

For some rank $n \in \mathbb{N}$, consider the change of variable $(\tau, u) \mapsto \Omega^{[n]}_\tau(u)$ given by (II.2.20) and (II.2.21). From a straightforward induction using Assumptions II.2.1 and II.2.2, it appears that this change of variable can be written as a *formal* exponential series,

$$\Omega^{[n]}_\tau(u) = \sum_{k\in\mathbb{N}} e^{-k\tau} \widehat{\Omega^{[n]}}_k(u).$$

This can be associated to a power series $\Xi^{[n]}(\xi; u) = \sum_{k\in\mathbb{N}} \xi^k \widehat{\Omega^{[n]}}_k(u), \xi \in \mathbb{C}, |\xi| \leq 1$, which is entirely determined by its behaviour on the border, i.e. by the periodic map

$$\Phi^{[n]}_\theta(u) = \Xi^{[n]}(e^{i\theta}; u) = \Omega^{[n]}_{-i\theta}(u) = \sum_{k\in\mathbb{N}} e^{ik\theta} \widehat{\Omega^{[n]}}_k(u). \tag{II.3.29}$$

Differentiating $\Phi^{[n+1]}$ w.r.t. θ and identifying the coeffients in (II.2.18), we obtain a (still formal) homological equation on $\Phi^{[n]}$:

$$\Big(\partial_\theta - iA\Big)\Phi_\theta^{[n+1]} = -i\varepsilon\left(f\circ\Phi_\theta^{[n]} - \partial_u\Phi_\theta^{[n]}\cdot F^{[n]}\right). \tag{II.3.30}$$

The periodic defect $\delta_\theta^{[n]} = -i\eta_{-i\theta}^{[n]}$ satisfies

$$\delta_\theta^{[n]} = \frac{1}{\varepsilon}\Big(\partial_\theta - iA\Big)\Phi_\theta^{[n]} + if\circ\Phi_\theta^{[n]}. - i\partial_u\Phi_\theta^{[n]}\cdot F^{[n]} \tag{II.3.31}$$

Note that these relations both use the identity

$$\sum_{k\in\mathbb{N}}\xi^k\widehat{f\circ\Omega^{[n]}}_k = f\left(\sum_{k\in\mathbb{N}}\xi^k\widehat{\Omega^{[n]}}_k\right) \tag{II.3.32}$$

which seems fairly evident, but requires the right-hand side of the equation to be well-defined for all $|\xi|\leq 1$.

Setting the filtered map $\widetilde{\Phi}_\theta^{[n]} = e^{-i\theta A}\Phi_\theta^{[n]}$, it satisfies

$$\partial_\theta\widetilde{\Phi}_\theta^{[n+1]} = \varepsilon\left(g_\theta\circ\widetilde{\Phi}_\theta^{[n]} - \partial_u\widetilde{\Phi}_\theta^{[n]}\cdot G^{[n]}\right) \tag{II.3.33}$$

with $g_\theta(u) = e^{-i\theta A}f\Big(e^{i\theta A}u\Big)$ and $G^{[n]} = iF^{[n]}$.

Property II.3.1. *Assumptions II.2.2 and II.2.3 ensure the following properties, with R, M and p given in Definition II.2.4 :*

(i) For all $\varepsilon\in(0,1]$, the Cauchy problem $\partial_t y^\varepsilon = g_{t/\varepsilon}(y^\varepsilon), y^\varepsilon(0) = u_0$ is well-posed in $\mathcal{K}_0$ up to some final time independent of ε.

(ii) For all $\theta\in\mathbb{T}$, the function $u\mapsto g_\theta(u)$ is analytic from $\mathcal{K}_{2R}$ to $\mathbb{C}^d$.

(iii) For all $\sigma\in[0,3]$,

$$\forall\ 0\leq\nu\leq p+2,\quad \frac{\sigma^\nu}{\nu!}\left\|\partial_\theta^\nu g\right\|_{\mathbb{T},2R}\leq M, \tag{II.3.34}$$

Initial condition (II.2.21) means that the periodic change of variable would be defined by

$$\widetilde{\Phi}_\theta^{[n+1]} = \mathrm{id} + \varepsilon\Big(T_\theta^{[n]} - \Pi(T^{[n]})\Big)\quad\text{and}\quad \Phi_\theta^{[n+1]} = e^{i\theta A}\Phi_\theta^{[n+1]} \tag{II.3.35}$$

with Π the average [3] and $T_\theta^{[n]} = \int_0^\theta \Big(g_\sigma \circ \widetilde{\Phi}_\sigma^{[n]} - \partial_u \widetilde{\Phi}_\sigma^{[n]} \cdot G^{[n]}\Big) \mathrm{d}\sigma$. Because $\widetilde{\Phi}^{[n]}$ is periodic at all rank n, taking the average in (II.3.33) gives the vector field

$$G^{[n]} = \Pi\Big(g \circ \widetilde{\Phi}^{[n]}\Big). \tag{II.3.36}$$

This is known as *standard averaging*. We introduce norms on periodic maps akin to (II.2.10) and (II.2.11), namely for $0 \leq \rho \leq 2R$, given a periodic map $(\theta, u) \in \mathbb{T} \times \mathcal{K}_\rho \mapsto \varphi_\theta(u)$,

$$\|\varphi\|_{\mathbb{T},\rho} := \sup_{(\theta,u)\in\mathbb{T}\times\mathcal{K}_\rho} |\varphi_\theta(u)| \quad \text{and} \quad \|\varphi\|_{\mathbb{T},\rho,\nu} := \max_{0\leq\alpha\leq\nu} \|\varphi_\theta(u)\|_{\mathbb{T},\rho} \tag{II.3.37}$$

where the second norm assumes that φ is ν-times continuously differentiable w.r.t. θ. Then the following bounds are satisfied.

Theorem II.3.2 (from [CLMV20] and [CCMM15])**.** *For $n \in \mathbb{N}$, let us denote $r_n = R/(n+1)$ and $\varepsilon_n := r_n/16M$. For all $\varepsilon > 0$ such that $\varepsilon \leq \varepsilon_n$, the maps $\Phi^{[n]}$ and $G^{[n]}$ are well-defined by* (II.3.35) *and* (II.3.36)*. The change of variable $\Phi^{[n]}$ and the defect $\delta^{[n]}$ are both $(p+2)$-times continuously differentiable w.r.t. θ, and $\Phi_0^{[n]}$ is invertible with analytic inverse on $\mathcal{K}_{R/4}$. Moreover, the following bounds are satisfied for $1 \leq \nu \leq p+1$,*

$$(i) \quad \|\widetilde{\Phi}^{[n]} - \mathrm{id}\,\|_{\mathbb{T},R} \leq 4\varepsilon M \leq \frac{r_n}{4}, \qquad (ii) \quad \|\partial_\theta^\nu \widetilde{\Phi}^{[n]}\|_{\mathbb{T},R} \leq 8\varepsilon M \nu!$$

$$(iii) \quad \|G^{[n]}\|_{\mathbb{T},R} \leq 2M \qquad (iv) \quad \|\widetilde{\delta}^{[n]}\|_{\mathbb{T},R,p+1} \leq 2M \left(2\mathcal{Q}_p \frac{\varepsilon}{\varepsilon_n}\right)^n$$

where $\widetilde{\Phi}_\theta^{[n]} = e^{-i\theta A}\Phi_\theta^{[n]}$ and $\widetilde{\delta}^{[n]} = e^{-i\theta A}\delta_\theta^{[n]}$ correspond to the filtered equation (II.3.33)*, and $\mathcal{Q}_p$ is a p-dependent constant.*

In order to prove Theorem II.2.5, we show that the previous calculations of this section are rigorous rather than formal. Let us work by induction and assume that the negative modes of $\Phi^{[n]}$ vanish (this is true for $\Phi_\theta^{[0]} = e^{i\theta A}$ since A is positive semidefinite). Because $(\theta, u) \mapsto \Phi_\theta^{[n]}(u)$ is continuously differentiable w.r.t. θ, its Fourier series converges absolutely, thus $(\xi, u) \mapsto \Xi^{[n]}(\xi; u)$ is well-defined for all $|\xi| \leq 1$ and $u \in \mathcal{K}_R$. By maximum modulus principle,

$$\|\Omega^{[n]} - e^{-\tau A}\|_R \leq \sup_{|\xi|\leq 1,\, u\in\mathcal{K}_R} |\Xi^{[n]}(\xi; u) - \xi^A| \leq \|\Phi^{[n]} - e^{i\theta A}\|_{\mathbb{T},R} \leq \|\widetilde{\Phi}^{[n]} - \mathrm{id}\,\|_{\mathbb{T},R}$$

3. Explicitely, $\Pi(\varphi) = \frac{1}{2\pi}\int_0^{2\pi} \varphi_\sigma \mathrm{d}\sigma$

The reasoning also stands for all derivatives $1 \leq \nu \leq p+1$,

$$\left\|\partial_\tau^\nu[\Omega^{[n]} - e^{-\tau A}]\right\| \leq \sup_{\xi,u}\left|(\xi\partial_\xi)^\nu[\Xi^{[n]}(\xi;u) - \xi^A]\right| \leq \left\|\partial_\theta^\nu[\Phi^{[n]} - e^{i\theta A}]\right\|_{\mathbb{T},R}$$

and $\left\|\partial_\theta^\nu[\Phi^{[n]} - e^{i\theta A}]\right\|_{\mathbb{T},R} \leq \left(1 + |||A|||\right)^\nu \|\partial_\theta^\nu \widetilde{\Phi}^{[n]}\|_{\mathbb{T},R,\nu}$. Furthermore, for $u \in \mathcal{K}_R$, let $x \in X_1$ s.t. $\left|u - \begin{pmatrix} x \\ 0 \end{pmatrix}\right| \leq \rho_1 + R$. Then for all $|\xi| \leq 1$,

$$\left|\Xi^{[n]}(\xi;u) - \begin{pmatrix} x \\ 0 \end{pmatrix}\right| \leq \left|\Phi_\theta^{[n]}(u) - \begin{pmatrix} x \\ 0 \end{pmatrix}\right| \leq \left|\widetilde{\Phi}_\theta^{[n]}(u) - \begin{pmatrix} x \\ 0 \end{pmatrix}\right|,$$

since a multiplication by $e^{-i\theta A}$ has no influence on the norm, nor on $\begin{pmatrix} x \\ 0 \end{pmatrix}$. A triangle inequality yields

$$\left|\Xi^{[n]}(\xi;u) - \begin{pmatrix} x \\ 0 \end{pmatrix}\right| \leq |\widetilde{\Phi}^{[n]} - u| + \left|u - \begin{pmatrix} x \\ 0 \end{pmatrix}\right| < \rho_1 + 2R,$$

therefore $f\Big(\Xi^{[n]}(\xi;u)\Big)$ is well-defined for all $|\xi| \leq 1$ and $u \in \mathcal{K}_R$, by expanding it into an absolutely converging series around $\begin{pmatrix} x \\ 0 \end{pmatrix}$, thereby justifying relations (II.3.32) and (II.3.31). The maximum modulus principle can finally be applied to the couple ($\eta^{[n]}$, $\delta^{[n]}$) in order to obtain the last estimate of Theorem II.2.5.

□

II.3.2 Proof of Theorem II.2.6 : well-posedness of the micro-macro problem

This proof is in several parts : first we show that problem (II.2.24a) is well-posed, and use this result to show that the bound on $w^{[n]}$ is satisfied, thereby also proving that (II.2.24b) is well-posed. Finally we focus on the bounds on $E^{[n]}$.

Let us set $\varphi(v) = u_0 + v - \Omega_0^{[n]}(u_0 + v)$. Using Theorem II.2.5, if $|v| \leq R/4$ then $|\varphi(v)| \leq R/4$. By Brouwer fixed-point theorem, there exists v^* such that $\varphi(v^*) = v^*$, i.e. $u^* \in \mathcal{K}_{R/4}$ such that $\Omega_0^{[n]}(u^*) = u_0$. Therefore $v^{[n]}(0) := u^* \in \mathcal{K}_{R/4}$.

Given $t > 0$ and assuming $v^{[n]}(s) \in \mathcal{K}_R$ for all $s \in [0,t]$, one can bound $v^{[n]}(t)$ using Theorem II.2.5 :

$$\left|v^{[n]}(t) - v^{[n]}(0)\right| = \left|\int_0^t F^{[n]}\left(v^{[n]}(s)\right) \mathrm{d}s\right| \leq 2Mt.$$

Setting $T_v := \dfrac{3R}{8M}$ ensures $\left|v^{[n]}(t) - v^{[n]}(0)\right| \leq 3R/4$, meaning that for all $t \in [0, T_v]$, $v^{[n]}(t)$ exists and is in $\mathcal{K}_R$. Again from Theorem II.2.5, we deduce $\Omega_\tau^{[n]}\left(v^{[n]}(t)\right) \in \mathcal{K}_{5R/4}$.

Focusing now on $w^{[n]}$ and assuming for all $s \in [0,t]$, $|w^{[n]}(s)| \leq R/4$, the linear term $L^{[n]}\left(\tau, s, w^{[n]}(s)\right)$ is bounded using a Cauchy estimate :

$$\left|L^{[n]}\left(\tau, s, w^{[n]}(s)\right)\right| \leq \|\partial_u f\|_{3R/2} \leq \frac{\|f\|_{2R}}{2R - \frac{3}{2}R} \leq \frac{2M}{R}$$

using a Cauchy estimate. The integral form then gives the bounds

$$\begin{aligned}\left|w^{[n]}(t)\right| &\leq \left|\int_0^t e^{\frac{s-t}{\varepsilon}A} L^{[n]}\left(s/\varepsilon, s, w^{[n]}(s)\right) w^{[n]}(s) \mathrm{d}s + \int_0^t e^{\frac{s-t}{\varepsilon}A} S^{[n]}(s/\varepsilon, s) \mathrm{d}s\right| \\ &\leq \int_0^t \frac{2M}{R}\left|w^{[n]}(s)\right| \mathrm{d}s + \left|\int_0^t e^{\frac{s-t}{\varepsilon}A} S^{[n]}(s/\varepsilon, s) \mathrm{d}s\right| \end{aligned} \tag{II.3.38}$$

Using the notation of the previous subsection, $\widetilde{\delta}_\theta^{[n]} = -ie^{-i\theta A}\eta_{-i\theta}^{[n]}$, from which

$$\eta_\tau^{[n]}(u) = \sum_{k \in \mathbb{Z}} e^{-(k+A)\tau} c_k^{[n]}(u) \quad \text{with} \quad c_k^{[n]}(u) = \frac{1}{2\pi}\int_0^{2\pi} e^{-ik\theta}\widetilde{\delta}_\theta(u) \mathrm{d}\theta.$$

Since $\langle \eta^{[n]} \rangle = 0$, i.e. $c_0^{[n]} = 0$, it is possible to bound the source term in $w^{[n]}$ by

$$\begin{aligned}\left|\int_0^t e^{\frac{s-t}{\varepsilon}A} S^{[n]}(s/\varepsilon, s) \mathrm{d}s\right| &\leq \left|\!\left|\!\left| e^{-\frac{t}{\varepsilon}A} \right|\!\right|\!\right| \int_0^t \sum_{k \in \mathbb{Z}^*} \left(e^{-k\frac{s}{\varepsilon}} \|c_k^{[n]}\|_{\mathbb{T},R}\right) \mathrm{d}s \\ &\leq \sum_{k \in \mathbb{Z}^*} \frac{\varepsilon}{k} \|c_k^{[n]}\|_{\mathbb{T},R} \;\leq\; \varepsilon \left(\sum_{k \in \mathbb{Z}^*} \frac{1}{k^2}\right) \left\|\partial_\theta \widetilde{\delta}^{[n]}\right\|_{\mathbb{T},R}\end{aligned}$$

where $\|\cdot\|_{\mathbb{T},R}$ is given by (II.3.37). Using Theorem II.3.2, there exists a constant $M_n > 0$ such that for all $t \in [0, T_v]$,

$$\left|\int_0^t e^{\frac{s-t}{\varepsilon}A} S^{[n]}(s/\varepsilon, s) \mathrm{d}s\right| \leq M_n \left(\frac{\varepsilon}{\varepsilon_n}\right)^{n+1}. \tag{II.3.39}$$

Using Gronwall's lemma in (II.3.38) with this inequality yields

$$|w^{[n]}(t)| \leq M_n\, e^{\frac{2M}{R}t} \left(\frac{\varepsilon}{\varepsilon_n}\right)^{n+1} \leq M_n\, e^{\frac{2M}{R}t}.$$

We now set $T_w > 0$ such that $M_n\, e^{\frac{2M}{R}T_w} \leq R/4$ (T_w may therefore depend on n, but does not depend on ε) and

$$T_n = \min(T_v, T_w).$$

This ensures the well-posedness of (II.2.24) on $[0, T_n]$ as well as the size of $w^{[n]}$.

Finally, the results on $E^{[n]}$ are a direct consequence of the bounds on the linear term

$$\sup_{\alpha+\beta+\gamma\leq p+1} \|\partial_\tau^\alpha \partial_t^\beta \partial_u^\gamma L^{[n]}\| < +\infty$$

and on the source term

$$\sup_{0\leq\alpha+\beta\leq p} \|\partial_\tau^\alpha \partial_t^\beta S^{[n]}\|_{L^\infty} = \mathcal{O}(\varepsilon^n), \qquad \sup_{\substack{\beta\geq 1 \\ 1\leq\alpha+\beta\leq p+1}} \|\partial_\tau^\alpha \partial_t^\beta S^{[n]}\|_{L^1} = \mathcal{O}(\varepsilon^{n+1}).$$

This stems directly from Cauchy estimates and Theorem II.2.5.

□

II.3.3 Proof of Theorem II.2.8 : uniform accuracy

The idea in this proof is to bound the errors on the macro part and micro part separately, using

$$\left|u^\varepsilon(t_i) - \Omega^{[n]}_{t_i/\varepsilon}(v_i) - w_i\right|_\varepsilon \leq \left|\Omega^{[n]}_{t_i/\varepsilon}\left(v^{[n]}(t_i)\right) - \Omega^{[n]}_{t_i/\varepsilon}(v_i)\right|_\varepsilon + \left|w^{[n]}(t_i) - w_i\right|_\varepsilon.$$

As the macro part $v^{[n]}$ involves no linear term, the scheme acts like any RK scheme on this part. Since $v^{[n]}$ and $F^{[n]}$ are non-stiff, the scheme is necessarily *uniformly* of order q, i.e.

$$\left|v^{[n]}(t_i) - v_i\right| \leq \Delta t^q \cdot t_i \cdot \|\partial_t^{q+1} v^{[n]}\|_{L^\infty}$$

using usual error bounds on RK schemes. The reader may notice that the absolute error involving $|\cdot|$ was used, not the modified error involving $|\cdot|_\varepsilon$. The results in [HO04] state

that an exponential RK scheme of order q generates an error given by

$$\left|w^{[n]}(t_i) - w_i\right|_\varepsilon \leq C\Delta t^q \Big(\|\partial_t^{q-1} E^{[n]}\|_\infty + \|\partial_t^q E^{[n]}\|_{L^1} \Big). \tag{II.3.40}$$

The bounds on $E^{[n]} = \partial_t w^{[n]} + \frac{1}{\varepsilon} A w^{[n]}$ and its derivatives w.r.t. ε can be found in Theorem II.2.6, rendering the computation of bounds on the error of the micro part straightforward. From Theorem II.2.5.*(i)*, $\Omega_\tau^{[n]}(u) = e^{-\tau A}u + \mathcal{O}(\varepsilon)$, therefore the error on $\Omega_{t/\varepsilon}^{[n]}(v^{[n]})$ is of the form

$$\Omega_{t_i/\varepsilon}^{[n]}\Big(v^{[n]}(t_i)\Big) - \Omega^{[n]}(v_i) = e^{-t_i A/\varepsilon}\Big(v^{[n]}(t_i) - v_i\Big) + \varepsilon r_i$$

where $v^{[n]}(t_i) - v_i$ and r_i are of size $t_i \cdot \Delta t^q$. The error can therefore be bounded, denoting $|||\cdot|||$ the induced norm from $\mathbb{R}^d$ to $\mathbb{R}^d$,

$$\left|\Omega_{t_i/\varepsilon}^{[n]}\Big(v^{[n]}(t_i)\Big) - \Omega^{[n]}(v_i)\right|_\varepsilon \leq \left(1 + \left|\left|\left| \frac{t_i}{\varepsilon} A e^{-\frac{t_i}{\varepsilon}A} \right|\right|\right|\right) |v^{[n]}(t_i) - v_i| + (\varepsilon + |||A|||)\,|r_i|.$$

From this we get the desired result on u^ε.

□

II.4 Application to some ODEs derived from discretized PDEs

In this section, we construct micro-macro problems for two *discretized* hyperbolic relaxation systems of the form

$$\begin{cases} \partial_t u + \partial_x \tilde{u} = 0 \\ \partial_t \tilde{u} + \partial_x u = \dfrac{1}{\varepsilon}(g(u) - \tilde{u}) \end{cases}$$

where g acts either as a differential operator on u (telegraph equation, Subsection II.4.1), or as a scalar value function (relaxed conservation law, Subsection II.4.2). These two problems may seem similar in theory, and the latter actually serves as a stepping stone to treat the former in [JPT98 ; JPT00], but we will treat them quite differently in practice. Some recent AP schemes with promising convergence have been developed for this type of problems in [BPR17 ; ADP20].

Let us insist that we only consider these problems *after discretization* (using either Fourier modes or an upwind scheme), yet even in a discrete framework, it will be apparent

that a direct application of the method is impossible, often because of the apparition of a backwards heat equation. The goal of this section is precisely to present some possible workarounds to overcome the problems that appear. Should the reader wish to see a more detailed and direct application of our method, they can find one in Subsection II.5.1.

II.4.1 The telegraph equation

A commonly studied equation in kinetic theory is the one-dimensional Goldstein-Taylor model, also known as the telegraph equation (see [JPT98 ; LM08], for instance). It can be written, for $(t, x) \in [0, T] \times \mathbb{R}/2\pi\mathbb{Z}$

$$\begin{cases} \partial_t \rho + \partial_x j = 0, \\ \partial_t j + \dfrac{1}{\varepsilon} \partial_x \rho = -\dfrac{1}{\varepsilon} j, \end{cases} \tag{II.4.41}$$

where ρ and j represent the mass density and the flux respectively. Using Fourier transforms in x, it is possible to represent a function $v(t, x)$ by

$$v(t, x) = \sum_{k \in \mathbb{Z}} v_k(t) e^{ikx}.$$

Considering a given frequency $k \in \mathbb{Z}$ the problem can be reduced to

$$\begin{cases} \partial_t \rho_k = -ikj_k, \\ \partial_t j_k = -\dfrac{1}{\varepsilon} \left(j_k + ik\rho_k \right). \end{cases}$$

Treating this problem using our method directly leads to dead-ends, therefore we will guide the reader through our reasoning navigating some of these dead-ends. This will lead to micro-macro decompositions of orders 0 and 1. These struggles can be seen as limitations of our approach, however we show that with only slight tweaks, it is possible to obtain an error of uniform order 2 using a standard exponential RK scheme. This result is summed up at the end of this subsection as Proposition II.4.1.

In order to make a component $-\frac{1}{\varepsilon} z$ appear, it would be tempting to set $z_k = j_k + ik\rho_k$. This quantity would verify the following differential equation

$$\partial_t z_k = -\frac{1}{\varepsilon} z_k + k^2 z_k - ik^3 \rho_k.$$

Integrating this differential equation gives

$$z_k(t) = \exp\left(-\lambda\frac{t}{\varepsilon}\right) z_k(0) - ik^3 \int_0^t e^{(s-t)\lambda/\varepsilon}\rho_k(s)\mathrm{d}s. \tag{II.4.42}$$

where $\lambda = 1 - \varepsilon k^2$. Because $\varepsilon \in (0,1]$ and $k \in \mathbb{Z}$ should not be correlated, λ can take any value in $(-\infty, 1)$. For λ negative, this equation is unstable and cannot be solved numerically using standard tools. To overcome this, we consider the stabilized change of variable instead

$$z_k = j_k + \frac{ik}{1+\alpha\varepsilon k^2}\ \rho_k$$

where α is a positive constant which we shall calibrate as the study progresses. This is the same change of variable as before up to $\mathcal{O}(\varepsilon)$, but $ik\rho_k$ was regularized with an elliptic operator to help with high frequencies. The problem to solve becomes

$$\begin{cases} \partial_t \rho_k = -\dfrac{k^2}{1+\alpha\varepsilon k^2}\rho_k - ikz_k, \\ \partial_t z_k = -\dfrac{1}{\varepsilon}z_k + \dfrac{k^2}{1+\alpha\varepsilon k^2}z_k - \dfrac{ik^3}{1+\alpha\varepsilon k^2}\left(\alpha + \dfrac{1}{1+\alpha\varepsilon k^2}\right)\rho_k. \end{cases} \tag{II.4.43}$$

As in (II.4.42), the growth of z_k is given by $e^{-\lambda t/\varepsilon}$ if λ is defined by

$$\lambda = 1 - \frac{\varepsilon k^2}{1+\alpha\varepsilon k^2} \in \left(1 - \frac{1}{\alpha}, 1\right].$$

For stability reasons λ must be positive, therefore we shall choose $\alpha \geq 1$.

Let us set $u_k = (\rho_k, z_k)^T$ and $A = \mathrm{Diag}(0,1)$ such that $\partial_t u_k = -\frac{1}{\varepsilon}Au_k + f(u_k)$ with

$$f(u) = \begin{pmatrix} -\dfrac{k^2}{1+\alpha\varepsilon k^2}u_1 - iku_2 \\ \dfrac{k^2}{1+\alpha\varepsilon k^2}u_2 - \dfrac{ik^3}{1+\alpha\varepsilon k^2}\left(\alpha + \dfrac{1}{1+\alpha\varepsilon k^2}\right)u_1 \end{pmatrix}. \tag{II.4.44}$$

In the upcoming study, we usually prefer the notation $f(\rho, z)$ rather than $f(u)$ so as to keep the distinction between both coordinates clear. Assuming $|k| \leq k_{\max}$, it is possible to bound $f(\rho_k, z_k)$ independently of k and of ε, allowing us to apply the method developed in this paper in order to approximate every ρ_k and j_k, and eventually $\rho(x,t)$ and $j(x,t)$. Note that no rigorous aspects of convergence in functional spaces are considered here – this will be treated in a forthcoming work. We omit the index k going forward for the

sake of clarity.

The micro-macro method is initialized by setting the change of variable $\Omega_\tau^{[0]}(\rho, z) = (\rho, e^{-\tau} z)^T$. The vector field followed by the macro part $v^{[0]}$ is $F^{[0]}$ given by

$$F^{[0]}(\rho, z) = \hat{k}^2 \begin{pmatrix} -\rho \\ z \end{pmatrix} \qquad \text{with} \qquad \hat{k} = \frac{k}{\sqrt{1 + \alpha\varepsilon k^2}}. \tag{II.4.45}$$

This means that the macro variable $v^{[0]}(t)$ is given by

$$v^{[0]}(t) = \begin{pmatrix} e^{-\hat{k}^2 t} & 0 \\ 0 & e^{\hat{k}^2 t} \end{pmatrix} v^{[0]}(0).$$

Notice that the growth of $v_2^{[0]}(t)$ is in $e^{\hat{k}^2 t}$, which is akin to the heat equation in reverse time.This is problematic, as it is possible for $\hat{k}$ to be quite big. For example with $k = 10, \alpha = 2$ and $\varepsilon = 10^{-2}$, one gets $e^{\hat{k}^2} \approx 3 \cdot 10^{14}$. However in order to obtain the solution of (II.4.41), $u_k(t) = \Omega_{t/\varepsilon}^{[0]}\left(v^{[0]}(t)\right) + w^{[0]}(t)$, we are only interested in $\Omega_{t/\varepsilon}^{[0]}\left(v^{[0]}(t)\right)$ for the macro part, and $\eta_{t/\varepsilon}^{[0]}\left(v^{[0]}(t)\right)$ for the micro part, which only depend on $e^{-\frac{t}{\varepsilon}A} v^{[0]}(t)$ as can be seen in the upcoming expression of $\eta^{[0]}$ and using $\Omega_\tau^{[0]}(u) = e^{-\tau A} u$. This means that the interesting quantity is

$$e^{-\frac{t}{\varepsilon}A} v^{[0]}(t) = \begin{pmatrix} e^{-\hat{k}^2 t} & 0 \\ 0 & e^{-(1-\varepsilon\hat{k}^2)\frac{t}{\varepsilon}} \end{pmatrix} v^{[0]}(0). \tag{II.4.46}$$

Recognizing $1 - \varepsilon\hat{k}^2 = \lambda > 0$ in this expression, it follows that $v_2^{[0]}$ is a decreasing function of time, therefore it is bounded uniformly for all t, k and ε. Because the exact computation of $e^{-\frac{t}{\varepsilon}A} v^{[0]}(t)$ is readily available, it is used during implementation, leaving only $w^{[0]}$ to be computed numerically using ERK schemes. Should the reader wish to conduct their own implementation, they should use the defect

$$\eta_\tau^{[0]}(\rho, z) = \begin{pmatrix} ike^{-\tau} z \\ \hat{k}^2 \left(\alpha + \dfrac{1}{1 + \alpha\varepsilon k^2}\right) ik\rho \end{pmatrix} = \eta_0^{[0]}(\rho, e^{-\tau} z).$$

By linearity of f, the micro variable $w^{[0]}$ follows the differential equation

$$\partial_t w^{[0]} = -\frac{1}{\varepsilon} A w^{[0]} + f(w^{[0]}) - \eta_0^{[0]}\left(e^{-\frac{t}{\varepsilon}A} v^{[0]}(t)\right), \qquad w^{[0]}(0) = 0.$$

The rescaled macro variable $e^{-\frac{t}{\varepsilon}A}v^{[0]}(t)$ is given by relation (II.4.46) with initial condition $v^{[0]}(0) = u(0) = (\rho_k(0), z_k(0))^T$.

Extending our expansion to order 1 is not trivial either. Direct application of iterations (II.2.20) yields

$$\Omega_\tau^{[1]}(\rho, z) = \begin{pmatrix} \rho + \varepsilon ike^{-\tau} z \\ e^{-\tau} z - \varepsilon \hat{k}^2 \left(\alpha + \dfrac{1}{1+\alpha\varepsilon k^2}\right) ik\rho \end{pmatrix}$$

from which the vector field for the macro part is

$$F^{[1]}(\rho, z) = \hat{k}^2 \left(1 + \varepsilon k^2 \left(\alpha + \frac{1}{1+\alpha\varepsilon k^2}\right)\right) \begin{pmatrix} -\rho \\ z \end{pmatrix}.$$

Following the same reasoning as before, one should study the evolution of the z-component of the rescaled macro variable $e^{-\frac{t}{\varepsilon}A}v^{[1]}(t)$. This evolution is in $e^{-\widetilde{\lambda}t/\varepsilon}$ where $\widetilde{\lambda} = 1 - \varepsilon\hat{k}^2\left(1 + \varepsilon k^2\left(\alpha + \frac{1}{1+\alpha\varepsilon k^2}\right)\right)$. Studying $\widetilde{\lambda}$ as a function of εk^2 in $\mathbb{R}_+$ shows that it is negative for $\varepsilon k^2 > 1$, whatever the value of $\alpha \geq 1$.

To circumvent this, we replace ε by $\frac{\varepsilon}{1+\alpha\varepsilon k^2}$ in iterations (II.2.20). This adds terms of order ε^2 in the definition of $\Omega^{[1]}$ that do not modify any properties of the micro-macro decomposition but it regularises the problem. Specifically, we define

$$\Omega_0^{[1]}(\rho, z) = \begin{pmatrix} \rho + \dfrac{\varepsilon}{1+\alpha\varepsilon k^2} ikz \\ z - \dfrac{\varepsilon}{1+\alpha\varepsilon k^2} \hat{k}^2 \left(\alpha + \dfrac{1}{1+\alpha\varepsilon k^2}\right) ik\rho \end{pmatrix}, \tag{II.4.47}$$

from which we get the vector field

$$F^{[1]}(\rho, z) = \hat{k}^2 \left(1 + \varepsilon\hat{k}^2 \left(\alpha + \frac{1}{1+\alpha\varepsilon k^2}\right)\right) \begin{pmatrix} -\rho \\ z \end{pmatrix}.$$

This time also, the identities $\Omega_\tau^{[1]}(u) = \Omega_0^{[1]}(e^{-\tau A}u)$ and $\eta_\tau^{[1]}(u) = \eta_0^{[1]}(e^{-\tau A}u)$ are satisfied, therefore the interesting variable is $e^{-\frac{t}{\varepsilon}A}v^{[1]}(t)$. The quantity dictating its growth is

$$\tilde{\lambda} = 1 - \varepsilon\hat{k}^2 \left(1 + \varepsilon\hat{k}^2 \left(\alpha + \frac{1}{1+\alpha\varepsilon k^2}\right)\right)$$

which is positive for all $\varepsilon k^2 \in \mathbb{R}_+$ if and only if $\alpha \geq 2$. As with the expansion of order 0,

the macro variable should be rescaled and computed exactly. The micro part $w^{[1]}$ is given by the differential equation

$$\partial_t w^{[1]} = -\frac{1}{\varepsilon} A w^{[1]} + f(w^{[1]}) - \eta_0^{[1]}\left(e^{-\frac{t}{\varepsilon}A} v^{[1]}(t)\right), \quad w^{[1]}(0) = u_k(0) - \Omega_0^{[1]}\left(v^{[1]}(0)\right) \tag{II.4.48}$$

where, writing $\hat{I} = (1+\alpha\varepsilon k^2)^{-1}$,

$$\eta_\tau^{[1]}(\rho, z) = ik\cdot\varepsilon\hat{k}^2\left(\alpha + \hat{I}\left(2+\varepsilon\hat{k}^2(\alpha+\hat{I})\right)\right)\begin{pmatrix} e^{-\tau}z \\ \hat{k}^2(\alpha+\hat{I})\rho \end{pmatrix} \tag{II.4.49}$$

$$\text{and} \quad v^{[1]}(0) = \begin{pmatrix} \rho_k(0) - \varepsilon\hat{I}ikz_k(0) \\ z_k(0) + \varepsilon\hat{k}^2(\alpha+\hat{I})ik\rho_k(0) \end{pmatrix}. \tag{II.4.50}$$

We approached the initial condition using Remark II.2.7, but an exact computation of the exact initial condition $\left(\Omega_0^{[1]}\right)^{-1}(u_0)$ is possible, as the map $u \mapsto \Omega_0^{[1]}(u)$ is linear.

Proposition II.4.1. *Given a maximum frequency $k_{\max} > 0$ and a scalar $\alpha \geq 2$, and assuming $|k| \leq k_{\max}$, the solution u_k of problem* (II.4.43) *can be decomposed into*

$$u_k(t) = \Omega_0^{[1]}\left(e^{-\frac{t}{\varepsilon}A} v^{[1]}(t)\right) + w^{[1]}(t)$$

where $\Omega_0^{[1]}$ is given by (II.4.47) *and $w^{[1]}(t) = \mathcal{O}(\varepsilon^2)$. The macro component $v^{[1]}$ is given by*

$$e^{-\frac{t}{\varepsilon}A} v^{[1]}(t) = \begin{pmatrix} e^{-K^{[1]}t} & 0 \\ 0 & e^{-(1-\varepsilon K^{[1]})\frac{t}{\varepsilon}} \end{pmatrix} v^{[1]}(0)$$

with $K^{[1]} = \hat{k}^2\left(1+\varepsilon\hat{k}^2\left(\alpha + \frac{1}{1+\alpha\varepsilon k^2}\right)\right)$, $\hat{k} = \frac{k}{\sqrt{1+\alpha\varepsilon k^2}}$ and $v^{[1]}(0)$ is either $\left(\Omega_0^{[k]}\right)^{-1}\left(u_k(0)\right)$ or its approximation (II.4.50). *The micro component $w^{[1]}$ is the solution to*

$$\partial_t w^{[1]} = -\frac{1}{\varepsilon} A w^{[1]} + f\left(w^{[1]}\right) - \eta_0^{[1]}\left(e^{-\frac{t}{\varepsilon}A} v^{[1]}(t)\right), \qquad w^{[1]}(0) = u_k(0) - \Omega_0^{[k]}\left(v^{[1]}(0)\right)$$

with f and $\eta_0^{[1]}$ given respectively by (II.4.44) *and* (II.4.49). *With these definitions, $w^{[1]}$ can be computed with a uniform error of order 2, therefore u_k can be computed with a uniform error of order 2.*

The reader may notice that only a finite number of modes is considered. This is

required so that there exists a bound uniform w.r.t. k and ε on the micro part of the problem (II.4.48) in order to apply our method. This is amenable to a CFL condition, i.e. some stiffness still exists due to the nature of the problem, but this stiffness is independent of ε. This is what we mean by uniform accuracy.

II.4.2 Relaxed conservation law

Our second test case is a hyperbolic problem for $(t, x) \in [0, T] \times \mathbb{R}/2\pi\mathbb{Z}$,

$$\begin{cases} \partial_t u + \partial_x \tilde{u} = 0, \\ \partial_t \tilde{u} + \partial_x u = \dfrac{1}{\varepsilon}(g(u) - \tilde{u}), \end{cases} \tag{II.4.51}$$

with smooth initial conditions $u(0, x)$ and $\tilde{u}(0, x)$. This is a stiffly relaxed conservation law, as presented in [JX95].

Remark II.4.2. *Note that the assumption that A has integer coefficients is restrictive in this case. One may want to consider the equation on the second coordinate to be*

$$\partial_t \tilde{u} + \partial_x u = \frac{\sigma(x)}{\varepsilon}\Big(b(x)u - \tilde{u}\Big)$$

as is done in [HS21], however this is not possible with our method. Perhaps a link can be made with the highly-oscillatory study [CJL17] where the "phase" t/ε in (II.1.4) *is replaced by a space-dependent function $\varphi(t, x)/\varepsilon$.*

Without the space-derivatives, this problem is straightforward : One can simply set $x = u$ and $z = g(u) - \tilde{u}$. Here new difficulties appear. For instance, in order to proceed, we require the following condition to be met :

$$|g'(u)| < 1 \tag{II.4.52}$$

This is a known stability condition when deriving asymptotic expansions for this kind of problem.

We start by discretising this system in space with $N > 0$ points. Going forward, $(x_j)_{j \in \mathbb{Z}/N\mathbb{Z}}$ denotes a fixed uniform discretisation of $\mathbb{R}/2\pi\mathbb{Z}$, of mesh size $\Delta x := 2\pi/N$. We define the vectors $U = (u_j)_j, \tilde{U} = (\tilde{u}_j)_j$ and, given a vector $V = (v_j)_j$ of size N, $g(V) = (g(v_j))_j$. For simplicity, $u_j(t)$ is the approximation of $u(t, x_j)$, and the same goes

for $\tilde{u}$. We denote D the matrix of centered finite differences and L the standard discrete Laplace operator, which is to say

$$DV = \Big(\frac{1}{2\Delta x}(v_{j+1} - v_{j-1})\Big)_j \quad \text{and} \quad LV = \Big(\frac{1}{\Delta x^2}(v_{j+1} - 2v_j + v_{j-1})\Big)_j$$

Using an upwind scheme after diagonalising problem (II.4.51) yields

$$\begin{cases} \partial_t U + D\tilde{U} - \dfrac{\Delta x}{2} LU = 0, \\ \partial_t \tilde{U} + DU - \dfrac{\Delta x}{2} L\tilde{U} = \dfrac{1}{\varepsilon}(g(U) - \tilde{U}). \end{cases} \tag{II.4.53}$$

Setting $U_1 = U$ and $U_2 := \tilde{U} - g(U_1)$, and neglecting the terms involving L for clarity, this problem becomes

$$\begin{cases} \partial_t U_1 = -D\Big(U_2 + g(U_1)\Big), \\ \partial_t U_2 = -\dfrac{1}{\varepsilon} U_2 + g'(U_1) DU_2 - T(U_1) \end{cases} \tag{II.4.54}$$

where we defined $T(U_1) := DU_1 - g'(U_1)Dg(U_1)$. From this, our method can be applied, but precautions must be taken in order to avoid having to solve the heat equation in backwards time. Therefore we set

$$\Omega_\tau^{[1]}(U_1, U_2) = \begin{pmatrix} U_1 + \varepsilon(1 - 2\varepsilon D^2)^{-1} DU_2 \\ e^{-\tau} U_2 - \varepsilon T(U_1) \end{pmatrix}.$$

Similarly to the manipulations for the telegraph equation, we multiplied ε by $(I_N - 2\varepsilon D^2)^{-1}$, but this time only for the first component. Writing $\widetilde{D} = (I_N - 2\varepsilon D^2)^{-1} D$, the associated vector field is

$$F^{[1]}(U_1, U_2) = \begin{pmatrix} -Dg(U_1) + \varepsilon DT(U_1) \\ g'(U_1)DU_2 - \varepsilon T'(U_1)\widetilde{D}U_2 - \varepsilon^2 g''(U_1)\Big(T(U_1), \widetilde{D}U_2\Big) \end{pmatrix}.$$

It is possible to obtain $\Omega^{[0]}$ and $F^{[0]}$ by neglecting the terms of order ε and above in the expressions above.

Remark II.4.3. *Remember that for the telegraph equation, the macro variable $v^{[1]}(t)$ needed to be rescaled by $e^{-tA/\varepsilon}$. This is not the case here : In the limit $\Delta x \to 0$, the macro variable $v^{[1]} = (\overline{u}_1, \overline{u}_2)^T$ is given by*

$$\begin{cases} \partial_t \overline{u}_1 = -\partial_x \left[g(\overline{u}_1) - \varepsilon \Big(1 - g'(\overline{u}_1)^2\Big) \partial_x \overline{u}_1 \right], \\ \partial_t \overline{u}_2 = g'(\overline{u}_1)\partial_x \overline{u}_2 - \Big(1 - g'(\overline{u}_1)^2\Big) \cdot (1 - 2\varepsilon\partial_x^2)^{-1} \varepsilon \partial_x^2 \overline{u}_2 + \varepsilon \phi^\varepsilon(\overline{u}_1, \widetilde{D}\overline{u}_2) \end{cases}$$

with $\widetilde{D} = (1 - 2\varepsilon\partial_x^2)^{-1}\partial_x$ and $\phi^\varepsilon(u_1, u_2) = g''(u_1)\,(2g'(u_1) - \varepsilon(1 - g'(u_1)^2)\partial_x u_1)\,u_2$. The operator $(1 - 2\varepsilon\partial_x^2)^{-1}\varepsilon\partial_x^2$ is bounded, therefore $\overline{u}_2$ is well-defined. The equation on $\overline{u}_1$ is a well-known result. If ε was also relaxed in the U_2-component of $\Omega^{[1]}$, there might be no need for condition (II.4.52) *but the result would be different.*

Because D^2 is sparse, it is not too costly to compute $\Big(I_N - \varepsilon D^2\Big)^{-1}$, however the conditioning may depend on the ratio between ε and Δx. Indeed, studying the eigenvalues of D reveals that the eigenvalues $(\mu_k)_{k \in \mathbb{Z}/N\mathbb{Z}}$ of $I_N - \varepsilon D^2$ are

$$\mu_k = 1 + \frac{\varepsilon}{\Delta x^2} \sin^2\left(2\pi \frac{k}{N}\right) \tag{II.4.55}$$

meaning that for N big, the conditioning is approximately $1 + \varepsilon/\Delta x^2$. Therefore, for ε big and Δx small, this inversion can become very costly, even though the cost remains bounded independently of ε.

Obtaining the defects of order 0 and 1 from these expressions presents no difficulty. For $\eta^{[1]}$, we separate here the U_1-component and the U_2-component for clarity.

$$\eta_\tau^{[0]}(U_1, U_2) = \begin{pmatrix} e^{-\tau} D U_2 \\ T(U_1) \end{pmatrix},$$

$$\begin{aligned} \eta_0^{[1]}(U_1, U_2)_{U_1} = \; & D\Big(g(U_1 + \varepsilon\widetilde{D}W) - g(U_1)\Big) + (D - \widetilde{D})U_2 \\ & + \varepsilon\widetilde{D}\Big(g'(U_1)DW - \varepsilon T'(U_1)\widetilde{D}W - \varepsilon^2 g''(U_1)\Big(T(U_1), \widetilde{D}W\Big)\Big), \end{aligned} \tag{II.4.56a}$$

$$\begin{aligned}\eta_0^{[1]}(U_1,U_2)_{U_2} = & -\Big(g'(U_1+\varepsilon\widetilde{D}U_2)-g'(U_1)\Big)DU_2\\ & + T(U_1+\varepsilon\widetilde{D}U_2)-T(U_1)-\varepsilon T'(U_1)\widetilde{D}U_2\\ & + \varepsilon g'(U_1+\varepsilon\widetilde{D}U_2)DT(U_1)-\varepsilon^2 g''(U_1)\Big(\widetilde{D}U_2,T(U_1)\Big)\\ & + \varepsilon T'(U_1)\left(Dg(U_1)-\varepsilon T(U_1)\right).\end{aligned} \tag{II.4.56b}$$

The values of $\eta_\tau^{[1]}(U_1,U_2)$ can be recovered using the identity

$$\eta_\tau^{[1]}(U_1,U_2)=\eta_0^{[1]}(U_1,e^{-\tau}U_2).$$

II.5 Numerical simulations

In this section we shall demonstrate our results by confirming the theoretical convergence rates of exponential Runge-Kutta (ERK) schemes from [HO05]. We also use these schemes on the original problem (II.1.1), thereby exhibiting the problem of order reduction.

In Subsection II.5.1 we study a toy model and a PDE-inspired problem with some non-linearity, for which we compute the micro-macro expansion up to order 2. In Subsection II.5.2, we showcase the results of uniform convergence for the partial differential equations of Section II.4. For these, the exact solution shall not take into account the error in space, i.e. it will be the solution to the discretized problem. Finally in Subsection II.5.3, we share our thoughts on some remaining avenues of research following this paper.

II.5.1 Application to some ODEs

Slowly oscillating toy problem

We first study an "oscillating" problem presented in [CCS16] which demonstrates a possible use of the method when studying non-linear problems :

$$\begin{cases}\dot{x}=(1-z)\begin{pmatrix}0&-1\\1&0\end{pmatrix}x\\ \dot{z}=-\dfrac{1}{\varepsilon}z+x_1^2x_2^2\end{cases} \tag{II.5.57}$$

with initial conditions $x_0=(0.1,0.7)^T$ and $z_0=0.05$, and final time $T=1$. This is of the

form $\partial_t u = -\frac{1}{\varepsilon}Au + f(u)$ when setting

$$u = \begin{pmatrix} x \\ z \end{pmatrix}, \qquad A = \mathrm{Diag}(0,0,1) \qquad \text{and} \qquad f(u) = \begin{pmatrix} -(1-u_3)u_2 \\ (1-u_3)u_1 \\ (u_1u_2)^2 \end{pmatrix}.$$

The macro part of our micro-macro decomposition is built by solving iterations on the homological equation

$$\Big(\partial_\tau + A\Big)\Omega_\tau^{[n+1]} = \varepsilon\left(f \circ \Omega_\tau^{[n]} - \partial_u \Omega_\tau^{[n]} F^{[n]}\right) \tag{II.5.58}$$

where $F^{[n]} = \left\langle f \circ \Omega^{[n]} \right\rangle$ with $\langle\,\cdot\,\rangle$ the projector on the $e^{-\tau A}$-component parallel to the other components of the exponential series. We choose the initial condition $\Omega_\tau^{[0]} = e^{-\tau A}$ and closure condition $\langle \Omega^{[0]} \rangle = e^{-\tau A}$. The first iteration yields

$$\Omega_\tau^{[1]}(x,z) = \begin{pmatrix} x_1 - \varepsilon e^{-\tau} x_2 z \\ x_2 + \varepsilon e^{-\tau} x_1 z \\ e^{-\tau} z + \varepsilon\,(x_1x_2)^2 \end{pmatrix} \quad \text{and} \quad F^{[1]}(x,z) = \begin{pmatrix} -\Big(1-\varepsilon(x_1x_2)^2\Big)x_2 \\ \Big(1-\varepsilon(x_1x_2)^2\Big)x_1 \\ 2\varepsilon x_1x_2z\Big({x_1}^2 - {x_2}^2\Big) \end{pmatrix}.$$

In order to compute the second order decomposition, one must compute the difference $T^{[1]} = f\circ\Omega^{[1]} - \partial_u\Omega^{[1]}F^{[1]}$, which is also used to compute the defect $\delta^{[1]} = \frac{1}{\varepsilon}(\partial_\tau + A)\Omega^{[1]} - T^{[1]}$. From a direct calculation this writes,

$$T_\tau^{[1]}(x,z) = \begin{pmatrix} e^{-\tau} z\,(x_2 + \varepsilon e^{-\tau} x_1 z + 2\varepsilon^2 x_1 {x_2}^2({x_1}^2 - {x_2}^2)) \\ -e^{-\tau} z\,(x_1 - \varepsilon e^{-\tau} x_2 z - 2\varepsilon^2 {u_1}^2 u_2({x_1}^2 - {x_2}^2)) \\ Z_0 + \varepsilon Z_1 + \varepsilon^2 Z_2 \end{pmatrix}$$

where for clarity we defined

$$Z_0 = \Big({x_1}^2 + \varepsilon^2 e^{-2\tau}(x_2 z)^2\Big)\Big(x_2 + \varepsilon^2 e^{-2\tau}(x_1 z)^2\Big),$$

$$Z_1 = -2x_1x_2({x_1}^2 - {x_2}^2)\left(1 - \varepsilon(x_1x_2)^2 + \varepsilon e^{-3\tau} z^{\,3}\right) \quad \text{and} \quad Z_2 = -e^{-2\tau}(2u_1u_2u_3)^2.$$

To compute the expansion of order 2, we truncate terms of order ε^2 and above in $T^{[1]}$

(which will not impact results of uniform accuracy) and solve (II.5.58). This yields [4]

$$\Omega_\tau^{[2]}(x,z) = \begin{pmatrix} x_1 - \varepsilon e^{-\tau} x_2 z - \frac{1}{2}\varepsilon^2 e^{-2\tau} z^2 x_1 \\ x_2 + \varepsilon e^{-\tau} x_1 z - \frac{1}{2}\varepsilon^2 e^{-2\tau} z^2 x_2 \\ z + \varepsilon (x_1 x_2)^2 - 2\varepsilon^2 x_1 x_2 (x_1^2 - x_2^2) \end{pmatrix},$$

$$F^{[2]}(x,z) = \begin{pmatrix} x_2(-1 + \varepsilon(x_1x_2)^2 - 2\varepsilon^2 x_1 x_2 (x_1^2 - x_2^2)) \\ x_1(1 - \varepsilon(x_1x_2)^2 + 2\varepsilon^2 x_1 x_2 (x_1^2 - x_2^2)) \\ 2\varepsilon z x_1 x_2 (x_1^2 - x_2^2) \end{pmatrix}.$$

The defect $\eta^{[2]}$ is obtained using relation (II.2.22) or by computing $\delta^{[2]}$ and identifying the Fourier coefficients.

Remark II.5.1. *It is possible to find an approximation of the center manifold $x \mapsto \varepsilon h^\varepsilon(x)$ by taking the limit $\tau \to \infty$ of the z-component of $\Omega^{[k]}$. For example here*

$$\varepsilon h^\varepsilon(x) = \varepsilon (x_1 x_2)^2 - 2\varepsilon^2 x_1 x_2 (x_1^2 - x_2^2) + \mathcal{O}(\varepsilon^3).$$

This coincides with the results in [CCS16].

We remind the reader that the problem that is solved at times $(t_i)_{0 \le i \le N}$ is

$$\begin{cases} \partial_t v^{[k]}(t) = F^{[k]}(v^{[k]}), \\ \partial_t w^{[k]}(t) = -\dfrac{1}{\varepsilon} A w^{[k]} + f\left(\Omega_{t/\varepsilon}^{[k]}(v^{[k]}) + w^{[k]}\right) - f\left(\Omega_{t/\varepsilon}^{[k]}(v^{[k]})\right) - \eta_{t/\varepsilon}^{[k]}(v^{[k]}), \end{cases}$$

with $k = 1, 2$. This yields vectors $(v_i) \approx (v^{[k]}(t_i))$ and $(w_i) \approx (w^{[k]}(t_i))$, from which an approximation $u_i \approx u^\varepsilon(t_i)$ is then obtained by setting $u_i = \Omega_{t_i/\varepsilon}^{[k]}(v_i) + w_i$. Initial conditions $v^{[k]}(0)$ and $w^{[k]}(0)$ are computed using Remark II.2.7.

4. It has been pointed out to the authors that the same result is obtained using nonlinear coordinate transforms described in [Rob14]. Some normal form methods compiled in [Mur06] also yield this result.

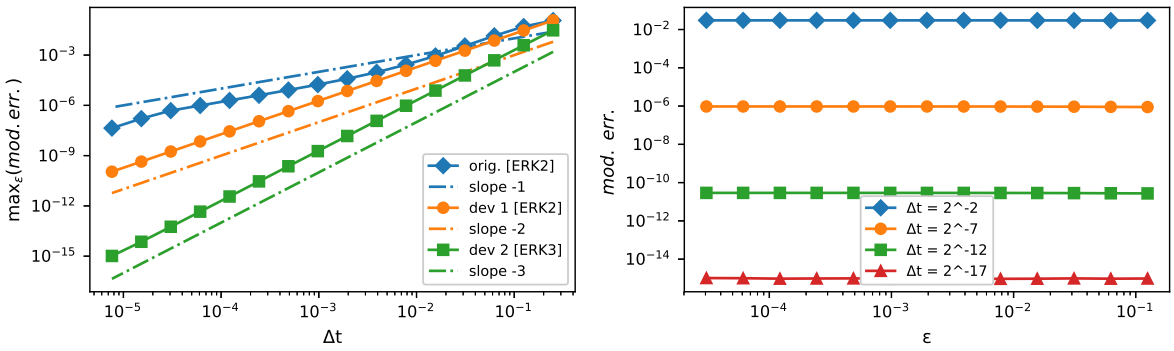


FIGURE II.1 – Oscillating case : On the left, maximum error on ε (for $\varepsilon = 2^{-k}$ with k spanning $\{3, \dots, 15\}$) as a function of Δt when using exponential RK schemes (abbr. ERK) of different orders. On the right, the error as a function of ε when solving the micro-macro problem of order 2 using ERK3.

The difference $f\left(\Omega^{[2]}_{t/\varepsilon}(v^{[2]}) + w^{[2]}\right) - f\left(\Omega^{[2]}_{t/\varepsilon}(v^{[2]})\right)$ is computed using

$$f(x+\tilde{x}, z+\tilde{z}) - f(x,z) = \begin{pmatrix} -(1-z)\tilde{x}_2 + (x_2+\tilde{x}_2)\tilde{z} \\ (1-z)\tilde{x}_1 - (x_1+\tilde{x}_1)\tilde{z} \\ \Big(x_1x_2 + (x_1+\tilde{x}_1)(x_2+\tilde{x}_2)\Big)\left(x_1\tilde{x}_2 + \tilde{x}_1x_2 + \tilde{x}_1\tilde{x}_2\right) \end{pmatrix}$$

in order to avoid rounding errors due to the size difference between u and $\tilde{u}$.

A PDE-inspired problem

Consider a problem similar to a relaxed conservation law (as in the next subsection) but without transport, written

$$\begin{cases} \dot{u} = \tilde{u}, & u(0) \in \mathbb{R}^d, \\ \dot{\tilde{u}} = \dfrac{1}{\varepsilon}\Big(g(u) - \tilde{u}\Big), & \tilde{u}(0) \in \mathbb{R}^d \end{cases}$$

for some smooth map $g : \mathbb{R}^d \mapsto \mathbb{R}^d$. This can be transformed in a system of the form (II.1.1) by setting $x = u$ and $z = g(u) - \tilde{u}$, yielding the problem

$$\begin{cases} \dot{x} = g(x) - z, & x(0) = u(0), \\ \dot{z} = -\dfrac{1}{\varepsilon}z + g'(x)\Big(g(x) - z\Big), & z(0) = g(u(0)) - \tilde{u}(0). \end{cases} \tag{II.5.59}$$

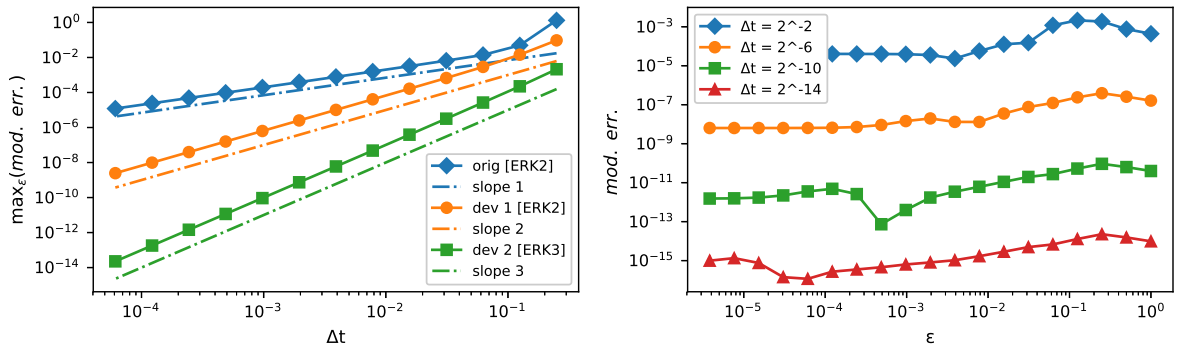


FIGURE II.2 – PDE-inspired problem : On the left, maximum error on ε (for $\varepsilon = 2^{-k}$ with k spanning $\{1, \dots, 18\}$) as a function of Δt when using exponential RK schemes (abbr. ERK) of different orders. On the right, the error as a function of ε when solving the micro-macro problem of order 2 using ERK3.

The change of variable and vector field can be computed by hand up to order 1,

$$\Omega_\tau^{[1]}(x,z) = \begin{pmatrix} x + \varepsilon e^{-\tau} z \\ e^{-\tau} z + \varepsilon g'(x) g(x) \end{pmatrix},$$

$$F^{[1]}(x,z) = \begin{pmatrix} g(x) - \varepsilon g'(x) g(x) \\ -g'(x) z + \varepsilon \Big(g'(x)^2 + g''(x) g(x) - \varepsilon g''(x) g'(x) g(x) \Big) z \end{pmatrix}.$$

Going to a higher order requires specific computations, as the expression of $\frac{1}{\varepsilon}(\partial_t + A)\Omega_\tau^{[2]}(x,z)$ is verbose and involves for instance $g(x + \varepsilon e^{-\tau} z) - g(x)$. It can be checked by hand that this expression involves no $e^{-\tau A}$-term with the above expressions of $\Omega^{[1]}$ and $F^{[1]}$. For numerical testing, we chose $g(x) = -x^3/3$, $u(0) = 1$ and $\tilde{u}(0) = 0$. The micro-macro problem was computed up to order 2.

Results

Figures II.1 and II.2 showcase the phenomenon of order reduction when solving the original problem (II.5.57) : Despite using a scheme of order 2, the error depends of ε in such a way that there exists no constant C such that the error is bounded by $C\Delta t^2$ for all ε. However there exists C such that the error is bounded by $C\Delta t$. In that case, we cannot say that the error is of *uniform* order 2, as this would require the error to be independent of ε.

This order reduction disappears when solving the micro-macro problem, as can be seen on the right-hand side of the figures for a decomposition of order 2. Furthermore, the theoretical orders of convergence from Theorem II.2.8 are confirmed. Indeed, using a scheme of order 2 (resp. 3) on the micro-macro problem of order 1 (resp. 2) generates a

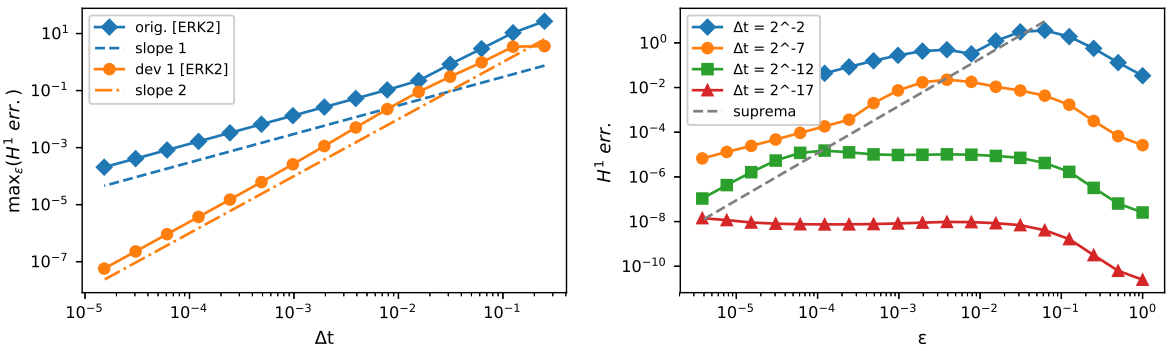


FIGURE II.3 – Telegraph equation : Absolute H^1 error on the solution of (II.4.41) computed by an ERK3 scheme. Supremum on ε as a function of Δt (left) and evolution of this error as a function of ε for the 1st-order decomposition (right).

uniform error of the expected order of convergence, with no order reduction.

II.5.2 Discretized hyperbolic partial differential equations

The telegraph equation

Using a spectral decomposition, we solve the problem, for $(t, x) \in [0, T] \times \mathbb{R}/2\pi\mathbb{Z}$,

$$\begin{cases} \partial_t \rho + \partial_x j = 0, \\ \partial_t j + \dfrac{1}{\varepsilon}\partial_x \rho = -\dfrac{1}{\varepsilon} j, \end{cases}$$

by setting $z = j + (1 - \alpha\varepsilon\Delta)^{-1}\partial_x z$, yielding problem (II.4.43). The micro-macro decomposition of order 1 is summarized in Property II.4.1, and its construction is detailed in Subsection II.4.1. Implementations are conducted using $\alpha = 2$, space frequencies are bounded by $k_{\max} := 12$, and initial data is $\rho(0, x) = e^{\cos(x)}$, $j(0, x) = \frac{1}{2}\cos^3(x)$.

Results can be seen in Figure II.3 when using a scheme of order 2. When solving the original problem, the uniform order degenerates from 2 to 1. When considering the micro-macro problem, the order of convergence is not reduced and stays of order 2. Although it varies with ε when considering a fixed Δt, when considering the supremum on ε, there is no order reduction. The dashed slope on the right plot interpolates the position of the supremum of the error for each fixed Δt. While the error seems to improve for $\varepsilon \ll \Delta t$, this does not cause any order reduction. This is stronger than the property of preservation of asymptotes (which ERK schemes have, see [DP11]), since AP schemes only need to be

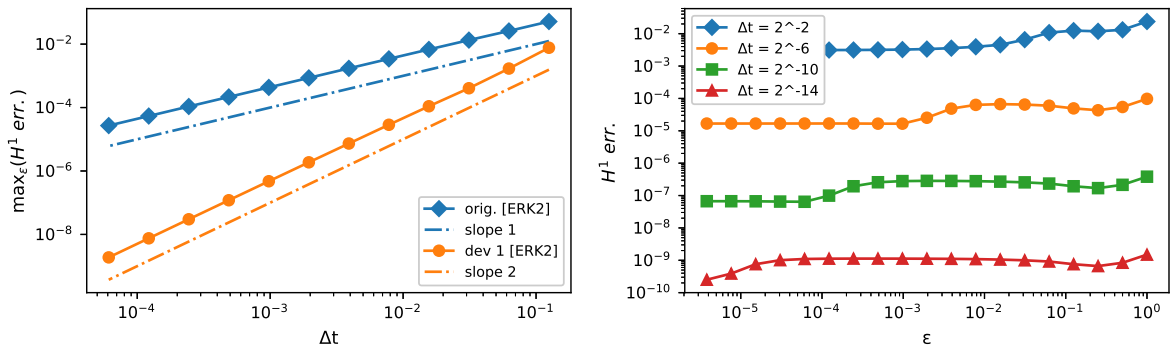


FIGURE II.4 – Relaxed Burgers-type problem : Maximum modified H^1 error (for ε spanning 1 to 2^{-18} using an ERK3 scheme as a function of Δt (left), and H^1 error as a function of ε for the micro-macro problem of order 1 (right).

well-defined in the limit $\varepsilon \to 0$. For them, this supremum does not need to be bounded. It appears that the relationship between the error bound and the stiffness of the linear operator is rather complex when using exponential RK schemes (again, see [HO05] for details).

Relaxed conservation law

Our second test case is a hyperbolic problem for $(t, x) \in [0, T] \times \mathbb{R}/2\pi\mathbb{Z}$,

$$\begin{cases} \partial_t u + \partial_x \tilde{u} = 0, \\ \partial_t \tilde{u} + \partial_x u = \dfrac{1}{\varepsilon}(g(u) - \tilde{u}), \end{cases}$$

discretized with finite volumes and written in the form of (II.1.1) by setting $u_1 = u$ and $u_2 = \tilde{u} - g(u)$ the x^ε- and the z^ε-component respectively. The micro-macro expansion is computed to order 1 using the strategy detailed in Subsection II.4.2.

For our tests, following [HS21], we consider $g(u) = bu^2$ with $b = 0.2$. Simulations run to a final time $T = 0.25$ and the mesh size is fixed : $N = 16$. Initial data is $u(0, x) = \frac{1}{2}e^{\sin(x)}$ and $\tilde{u}(0, x) = \cos(x)$. The reference solution was computed up to a precision 10^{-12} using an ERK2 scheme. Convergence results are presented in Figure II.4, confirming theoretical results once more.

It should be said again that our approach does not study the error in space, only in time. For instance, the relationship between the error bound and the grid size is not considered. Further studies will be conducted, especially considering CFL conditions, L^2 and H^1 norms, and computational costs.

II.5.3 Perspectives

Computing cost

Note that when using a given scheme, solving a single step is much more costly for the micro-macro problem than for the direct problem : Not only is the system size doubled, but the functions implicated require more computing power to obtain a single value (especially the defect, see (II.4.56) for instance). It is therefore plausible to think that our method is best for computing values during the transient phase, after which it is possible to solve the original problem with uniform accuracy.

The regularized derivation $\left(I_N - 2\varepsilon D^2\right)^{-1} D$ which appears in the micro-macro problem of the relaxed hyperbolic system may be prohibitively costly to compute for some schemes such as WENO, for which the derivation operator is non-linear. However we may be able to work around this, as the goal of the relaxation term is only to dampen high-frequencies, and as such inverting any discrete Laplace operator should suffice, independently of the scheme used to discretize the transport. Clearly, the subject of utilizing such regularizations for numerical purposes is complex and beyond the scope of this paper.

Near-equilibrium convergence

If one chooses an initial condition $z^\varepsilon(0) = 0$ in (II.1.1), then it is close to the center manifold up to $\mathcal{O}(\varepsilon)$, and Problem (II.1.2) can be solved with uniform accuracy of order 2 but only when considering the absolute error $|\cdot|$, not the modified error $|\cdot|_\varepsilon$ from (II.2.27). The same behaviour is observed for the telegraph equation when setting $j(0,x) = -\partial_x \rho(0,x)$, meaning $z = \mathcal{O}(\varepsilon)$. This would theoretically mean that we need to push the micro-macro decompositions up to order 2 if we want to improve the order of convergence. However, this is not the case : uniform accuracy of order 3 is obtained from an expansion of order 1 for all test cases. This "order gain" also propagates to our micro-macro decomposition of order 2 for the oscillating toy problem. These results can be seen in Figure II.5 and will be studied in future works.

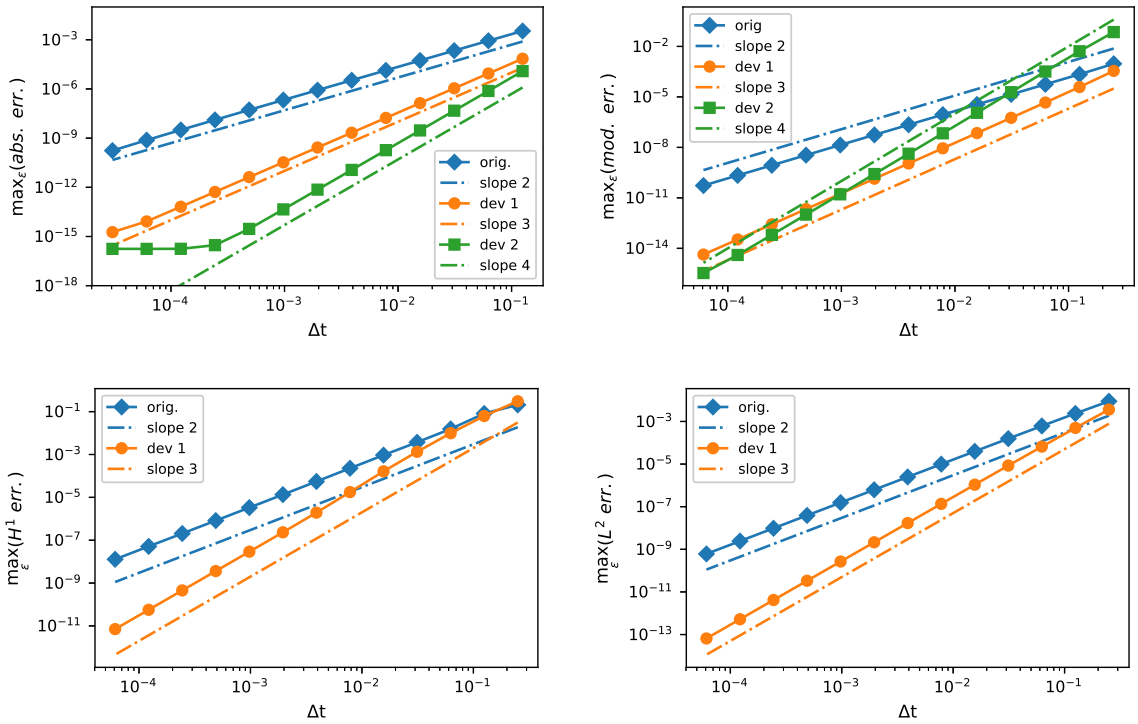


FIGURE II.5 – In reading order, errors when solving the oscillating toy problem, the PDE-inspired problem, the telegraph equation and the relaxed conservation law. All systems start near equilibrium and are solved with exponential Runge-Kutta schemes of the observed order of convergence.

Chapitre III

Discussion d'extension des résultats

> Closure. I keep hearing that word. [...] As soon as a show has a sense of closure, it gives you an excuse to forget you've seen the damn thing.
>
> David Lynch, 1990 [1]

Ce chapitre sert d'ouverture, il présente un éventail de pistes de réflexions, d'extensions des résultats sur les problèmes à relaxation rapide du chapitre précédent. Rappelons d'abord le contexte : on considère la solution $t \mapsto u^\varepsilon(t)$ du problème

$$\partial_t u^\varepsilon = -\frac{1}{\varepsilon} A u^\varepsilon + f(u^\varepsilon), \qquad u^\varepsilon(0) = u_0 \in \mathbb{R}^d \tag{III.0.1}$$

avec A une matrice diagonale positive à valeurs propres entières, et $u \mapsto f(u)$ un champ de vecteurs régulier. Pour tout ordre $n \in \mathbb{N}$, on construit d'un changement de variable $(\tau, u) \mapsto \Omega_\tau^{[n]}(u)$ proche de $e^{-\tau A}$ et d'un champ de vecteurs non-raide $u \mapsto F^{[n]}(u)$. Cela permet de décomposer la solution $t \mapsto u^\varepsilon(t)$ en une partie *macro* $v^{[n]}(t)$ et une partie *micro* $w^{[n]}(t) = \mathcal{O}(\varepsilon^{n+1})$, de sorte que

$$u^\varepsilon(t) = \Omega_{t/\varepsilon}^{[n]}\Big(v^{[n]}(t)\Big) + w^{[n]}(t). \tag{III.0.2}$$

Le problème micro-macro sur $(v^{[n]}, w^{[n]})$ s'écrit

$$\begin{cases} \partial_t v^{[n]}(t) = F^{[n]}(v^{[n]}), & \text{(III.0.3a)} \\ \partial_t w^{[n]}(t) = -\dfrac{1}{\varepsilon} A w^{[n]} + f\left(\Omega_{t/\varepsilon}^{[n]}(v^{[n]}) + w^{[n]}\right) - f\left(\Omega_{t/\varepsilon}^{[n]}(v^{[n]})\right) - \eta_{t/\varepsilon}^{[n]}(v^{[n]}), & \text{(III.0.3b)} \end{cases}$$

1. https://web.archive.org/web/20200805032143/https://www.latimes.com/archives/la-xpm-1990-02-18-ca-1500-story.html

avec $\eta^{[n]}$ le défaut définit par

$$\eta_\tau^{[n]} = \frac{1}{\varepsilon}\left(\partial_\tau + A\right)\Omega_\tau^{[n]} + \partial_u \Omega_\tau^{[n]} \cdot F^{[n]} - f \circ \Omega^{[n]}. \tag{III.0.4}$$

La donnée initiale est $v^{[n]}(0) = \left(\Omega_0^{[n]}\right)^{-1}(u_0)$, $w^{[n]}(0) = 0$. Ce problème est bien posé et on peut le résoudre avec une *précision uniforme* à l'ordre $n+1$ en utilisant un schéma exponentiel Runge-Kutta.

Dans ce chapitre, on discute de nombreux sujets ouverts qui pourraient faire le sujet d'études futures. En particulier, en Section III.1 on discute de problématiques d'implémentation, comme le coût de calcul, l'accumulation d'erreurs, et le besoin d'une dérivée exacte. En Section III.2 on définit un problème micro-macro autonome, ce qui permet d'expliquer le gain d'ordre remarqué en fin de Chapitre II et de définir une approche « pullback » comme ce qui avait pu être fait dans le cadre hautement oscillant. Enfin en Section III.3, on présente une piste basée sur des développements de Padé pour traiter l'équation du télégraphe de manière rigoureuse à tout ordre.

III.1 Extensions directes du micro-macro

On présente ici des problématiques d'implémentation et de coût numérique, ainsi qu'une piste d'extension possible pour rendre le développement micro-macro automatique.

III.1.1 Problématiques numériques

Avec notre construction micro-macro, les non-linéarités du problème sont exacerbées. Par exemple sur le problème

$$\begin{cases} \dot{x} = g(x) - z, & x(0) = u(0), \\ \dot{z} = -\dfrac{1}{\varepsilon} z + g'(x)\Big(g(x) - z\Big), & z(0) = g(u(0)) - \tilde{u}(0). \end{cases}$$

avec $g(x) = x^3/3$, on a implémenté (en Julia) le défaut à l'ordre 2 de la manière suivante :

```
function eta2(x,z,epsil)
    etaXCoeffs = [
        64*x^12*z/9 + 8*x^6*z^3/3 ;
        88*x^10*z/9 + 4*x^4*z^3 ;
```

```
        68*x^8*z/9 + 4*x^5*z^2 + 2*x^2*z^3 ;
        6*x^6*z + 4*x^3*z^2 + z^3/3 ;
        13*x^4*z/3 - x*z^2 ;
        0 ;
        0
    ]
    etaZCoeffs = [
        -32*x^10*z^5/3 ;
        -80*x^8*z^5/3 ;
        -80*x^9*z^4/3 - 80*x^6*z^5/3 ;
        -160*x^7*z^4/3 - 40*x^4*z^5/3 ;
        -160*x^11*z^2/9 - 80*x^8*z^3/3 - 32*x^5*z^4 - 10*x^2*z^5/3 ;
        -52*x^9*z^2/3 - 40*x^6*z^3 - 16*x^3*z^4/3 - z^5/3 ;
        -448*x^13/81 - 268*x^7*z^2/9 - 16*x^4*z^3 + x*z^4/3 ;
        -32*x^11/9 - 59*x^5*z^2/3 - 10*x^2*z^3/3 ;
        -95*x^9/27 - 26*x^3*z^2/3 - z^3 ;
        0 ;
        0
    ]
    etaX, etaZ = BigFloat(0), BigFloat(0)
    for xCoeff in etaXCoeffs
        etaX = epsil*etaX + xCoeff
    end
    for zCoeff in etaZCoeffs
        etaZ = epsil*etaZ + zCoeff
    end
    return [etaX;etaZ]
end
```

Cette implémentation voit $\eta^{[n]}(x,z)$ comme un polynôme en ε et applique la méthode d'Horner, ce qui permet d'avoir un résultat pertinent dans le cas où ε^2 est proche de l'erreur machine.

On voit cependant apparaître des termes jusqu'à x^{13}, ce qui peut vite être problématique du point de vue de la stabilité numérique, dès lors que x prend des valeurs élevées. Sur le problème d'origine, la non-linéarité est cubique seulement, et $g(5) \leq 100$, donc il ne serait pas surprenant que x prenne des valeurs de cette taille. Cependant, on a alors

$x^{13} \sim 10^9$, ce qui peut engendrer des erreurs importantes dans le calcul du défaut, surtout si ε est grand.

En outre, une évaluation de $\eta^{[n]}(x,z)$ avec l'implémentation ci-dessus demande 24.5 μs, ce qui est 35 fois supérieur aux 699 ns que demande une évaluation du champ de vecteurs d'origine $f(x,z) = \left(\frac{1}{3}x^3 - z\right)\begin{pmatrix}1\\x^2\end{pmatrix}$. Il paraît donc évident qu'une fois la phase transitoire passée, il vaut mieux utiliser un schéma « UA à l'équilibre » sur le système d'origine pour éviter les calculs coûteux.

III.1.2 Dépasser les développements formels

Dans le Chapitre II, la décomposition micro-macro est construite *via* des résolutions successives de l'équation homologique

$$\partial_\tau \Omega_\tau^{[n+1]} + A\Omega_\tau^{[n+1]} = \varepsilon\left(f \circ \Omega_\tau^{[n]} - \partial_u \Omega_\tau^{[n]} \cdot F^{[n]}\right). \tag{III.1.5}$$

On utilise donc des outils de calcul formel pour construire les applications $(\tau, u) \mapsto \Omega_\tau^{[n]}(u)$ et $u \mapsto F^{[n]}(u)$ *a priori* de l'implémentation du problème micro-macro. Cette approche présente une difficulté : la traduction entre le calcul formel et l'implémentation est difficile à automatiser car les expressions obtenues sont trop longues, et la transcription manuelle peut être source d'erreur. On peut donc souhaiter calculer le problème micro-macro « à la volée » sans calcul formel.

Étant donné que notre construction micro-macro est équivalente à celle pour les problèmes hautement oscillants de [CLMV20], on peut prendre se tourner vers cette catégorie de problèmes pour réaliser ce souhait. Dans [CLMZ20], les auteurs remplacent les dérivées $\partial_u \Omega_\tau^{[n]}$ par des différences finies

$$\partial_u \Omega_\tau^{[n]}(u) \cdot v = \frac{1}{\varepsilon^{n+1}}\left(\Omega_\tau^{[n]}(u + \varepsilon^{n+1}v) - \Omega_\tau^{[n]}(u)\right) + \mathcal{O}(\varepsilon^{n+1}),$$

de sorte que l'erreur engendrée sur $\Omega^{[n+1]}$ est de taille $\mathcal{O}(\varepsilon^{n+2})$. La partie périodique est traitée avec des séries de Fourier et la partie non-raide avec une méthode multipas de type Adams-Bashforth.

L'adaptation de cette approche aux problèmes de relaxation n'est cependant pas directe. En effet, ce résultat concerne les problèmes à oscillations forcées, pour lesquels le problème micro-macro ne présente pas la raideur $\frac{-1}{\varepsilon}Aw^{[n]}$. Ce terme impose l'utilisation

de schémas adaptés dont il faut vérifier la bonne convergence dans ce nouveau contexte.

III.2 Micro-macro autonome

On a remarqué qu'avec notre construction, on a les identités

$$\Omega_\tau^{[n]} = \Omega_0^{[n]} \circ e^{-\tau A}, \qquad F^{[n]} \circ e^{-\tau A} = e^{-\tau A} F^{[n]}, \qquad \eta_\tau^{[n]} = \eta_0^{[n]} \circ e^{-\tau A}. \tag{III.2.6}$$

Ainsi, on pourrait décider de modifier l'identité (III.0.2) en incluant la relaxation $\tau \mapsto e^{-\tau A}$ dans la variable macro. [2] Par commutation de $F^{[n]}$ et $e^{-\tau A}$, en posant $\tilde{v}(t) = e^{-tA/\varepsilon} v^{[n]}(t)$, on se ramène à un problème autonome,

$$\begin{cases} \partial_t \tilde{v} = -\dfrac{1}{\varepsilon} A\tilde{v} + F^{[n]}(\tilde{v}), & \text{(III.2.7a)} \\ \partial_t w^{[n]}(t) = -\dfrac{1}{\varepsilon} A w^{[n]} + f\left(\Omega_0^{[n]}(\tilde{v}) + w^{[n]}\right) - f\left(\Omega_0^{[n]}(\tilde{v})\right) - \eta_0^{[n]}(\tilde{v}). & \text{(III.2.7b)} \end{cases}$$

De prime abord, la résolution de (III.2.7a) semble présenter les mêmes difficultés que la résolution de (III.0.1), mais ici $F^{[n]}$ et A commutent, donc on peut résoudre ce problème par splitting de Lie sans perte de précision. En outre on peut calculer le défaut sans faire apparaître la variable artificielle τ grâce à l'identité

$$\eta_0^{[n]} = \frac{1}{\varepsilon}\left(A\Omega_0^{[n]} - \partial_u \Omega_0^{[n]} \cdot A\right) + \partial_u \Omega_0^{[n]} \cdot F^{[n]} - f \circ \Omega_0^{[n]}. \tag{III.2.8}$$

Remarque III.2.1. *Comme remarqué lors de la construction de la décomposition micro-macro, on a choisi que la composante constante de $(\tau, u) \mapsto e^{\tau A}\Omega_\tau^{[n]}(u)$ comme étant l'identité. Ainsi, notre construction est formellement équivalente à la construction « naturelle » des formes normales telles que présentées dans [Mur06, Sec. 3.2 & 4.3]. On conjecture donc qu'on peut construire le changement de variable Ω_0^ε avec d'autres méthodes de forme normale (e.g. Birkhoff) et obtenir des résultats similaires.*

Il est à noter que cette formulation autonome est incompatible avec l'application directe de méthodes exponentielles Runge-Kutta (expRK), puisque la variable $\tilde{v}$ est raide. Ceci met en évidence une limite importante de ces méthodes. L'approche avec le problème non-autonome revient à appliquer une méthode de Lawson [3] sur $v^{[n]}$ combinée à une mé-

2. C'est d'ailleurs ce qu'on fait dans nos développements pour l'équation du télégraphe.
3. Pour rappel, les méthodes de Lawson consistent à appliquer un schéma Runge-Kutta standard sur

thode expRK sur $w^{[n]}$. Une comparaison de ces méthodes est faite dans [CEM20] dans le cadre de champs qui commutent, bien que le contexte surjacent soit différent.

III.2.1 Interprétation du gain d'ordre

Comme le champ de vecteurs $F^{[n]}$ commute avec $e^{-\tau A}$, on peut séparer les composantes en x et en z en écrivant $F^{[n]} = \begin{pmatrix} a^{[n]} \\ b^{[n]} \end{pmatrix}$, et par commutation

$$\begin{pmatrix} a^{[n]}(u) \\ 0 \end{pmatrix} = \lim_{\tau \to 0} e^{-\tau A} F^{[n]}(u) = \lim_{\tau \to 0} F^{[n]}\Big(e^{-\tau A} u\Big) = F^{[n]}(x, 0)$$

La composante $a^{[n]}$ contient une réduction de dimension, ce qui est cohérent avec le théorème de variété centrale. L'autre aspect de cette identité est la partie $b^{[n]}(x, 0) = 0$. Ainsi, le problème sur la composante en z de $v^{[n]}$ est essentiellement linéaire, donc la donnée initiale est de taille $\mathcal{O}(\varepsilon^n)$, cette composante reste de la même taille pour tout t.

Or, si on considère une donnée initiale au problème (III.0.1) dont la composante en z est nulle, alors la donnée initiale $v^{[n]}(0)$ aura une composante z de taille $\mathcal{O}(\varepsilon)$. On voit alors que le problème (III.0.3) est mieux posé, au sens où ses dérivées sont bornées jusqu'à un ordre supérieur au cas général.

III.2.2 Approche *pullback*

Dans la veine des formes normales, on peut s'interroger sur le comportement de la nouvelle variable

$$y^{[n]}(t) := \Big(\Omega_0^{[n]}\Big)^{-1}(u^\varepsilon(t)).$$

Par dérivation et avec l'identité $\partial_u\Big(\Omega^{-1}\Big)(\Omega(y)) = \Big(\partial_u \Omega(y)\Big)^{-1}$, on obtient

$$\partial_t y^{[n]} = -\frac{1}{\varepsilon}\Big(\partial_u \Omega_0^{[n]}(y^{[n]})\Big)^{-1} \cdot A\Omega_0^{[n]}(y^{[n]}) + \Big(\partial_u \Omega_0^{[n]}(y^{[n]})\Big)^{-1} \cdot f \circ \Omega_0^{[n]}(y^{[n]}). \tag{III.2.9}$$

À l'aide de l'identité (III.2.8), on en déduit

$$\partial_t y^{[n]} = -\frac{1}{\varepsilon} A y^{[n]} + F^{[n]}\Big(y^{[n]}\Big) - \Big(\partial_u \Omega_0^{[n]}\Big(y^{[n]}\Big)\Big)^{-1} \eta_0^{[n]}(y^{[n]}). \tag{III.2.10}$$

la variable filtrée $e^{tA/\varepsilon}u^\varepsilon(t)$ et d'exprimer ce schéma en fonction de la variable non-filtrée $u^\varepsilon(t)$. Des estimations d'erreur sont présentées dans [HLO20].

Il apparaît alors une nouvelle difficulté, qui est le calcul du défaut modifié. Il faut trouver un algorithme de point fixe qui permet de faire ce calcul sans avoir de calculer un inverse $\partial_u \Omega_0^{[n]}$, comme cela a été fait dans [CLMV20].

Remarque III.2.2. *On peut aussi interpréter* (III.2.9) *comme une équation de la forme*

$$\partial_t y^{[n]} = -\frac{1}{\varepsilon} A^{[n]}(y^{[n]}) + f^{[n]}(y^{[n]})$$

avec $A^{[n]}$ et $f^{[n]}$ qui commutent à $\mathcal{O}(\varepsilon^n)$ près. Cette formulation rappelle la Proposition I.3.2 et le Théorème I.5.2 de moyennisation.

III.3 Autour de l'équation du télégraphe

On considère l'équation du télégraphe dans le domaine de Fourier en espace. On fixe s'intéresse à une fréquence fixe $\xi \in \mathbb{R}_+$ et on cherche $t \mapsto (\hat{\rho}(t), \hat{\jmath}(t))$ qui satisfait

$$\begin{cases} \partial_t \hat{\rho} = -i\xi \hat{\jmath}, \\ \partial_t \hat{\jmath} = -\dfrac{1}{\varepsilon^2} \left(\hat{\jmath} + i\xi \hat{\rho} \right). \end{cases}$$

On peut calculer les valeurs propre de ce système pour tout (ε, ξ), et on choisit de les écrire

$$\lambda_\pm = \frac{-1 \pm \sqrt{1 - 4\varepsilon^2 \xi^2}}{2\varepsilon^2},$$

quitte à étendre cette définition pour $2\varepsilon\xi > 1$ avec $\sqrt{-1} = i$. La construction micro-macro « standard » telle que développée précédemment permet de calculer un développement limité de ces valeurs propres en la limite $\varepsilon \to 0$ et de trouver le changement de variable qui diagonalise le problème. Par exemple les valeurs propres se développent en

$$\begin{aligned} \lambda_+ &= -\xi^2 - \varepsilon^2 \xi^4 - 2\varepsilon^4 \xi^6 - 5\varepsilon^6 \xi^8 + \mathcal{O}(\varepsilon^{10}), \\ \lambda_- &= -\frac{1}{\varepsilon^2} + \xi^2 + \varepsilon^2 \xi^4 + 2\varepsilon^4 \xi^6 + 5\varepsilon^6 \xi^8 + \mathcal{O}(\varepsilon^{10}). \end{aligned}$$

Malheureusement, la singularité en $\varepsilon = 1/(2\xi)$ limite le domaine de validité de ces développements asymptotiques aux valeurs $0 < \varepsilon < 1/(2\xi)$, ce qui est problématique étant donné que ξ peut être quelconque. L'extension de ces expressions à un domaine indépendant de ξ demande un changement d'approche, notamment à cause de λ_- qui se traduit en chaleur rétrograde pour $\xi > 1/\varepsilon$.

Quitte à faire un changement de variable $t \leftarrow t/\varepsilon^2$ et $\hat{\jmath} \leftarrow \varepsilon\hat{\jmath}$ pour l'étude formelle, on peut se ramener à un unique paramètre $\varepsilon\xi$. En effet, on a alors le système

$$\begin{cases} \partial_t \hat{\rho} = -i\varepsilon\xi\hat{\jmath}, \\ \partial_t \hat{\jmath} = -\left(\hat{\jmath} + i\varepsilon\xi\hat{\rho}\right), \end{cases} \tag{III.3.11}$$

et, puisqu'on souhaite avoir ε et ξ indépendant, on considère $\varepsilon\xi \in \mathbb{R}_+$. L'approche standard revient à faire une étude du problème autour de $|\varepsilon\xi| \ll 1$, et on souhaite « dépasser » la singularité en $\varepsilon\xi = 1/2$.

À la manière des formes normales, on cherche un changement de variable qui diagonalise la matrice « perturbée » $R + \varepsilon\xi A$ avec

$$R = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} \qquad \text{et} \qquad A = -i \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

En d'autres termes, on cherche $\Omega_{\varepsilon\xi}$ de sorte à avoir l'identité

$$R + \varepsilon\xi A = \Omega_{\varepsilon\xi} \left(R + \varepsilon\xi F_{\varepsilon\xi}\right) \Omega_{\varepsilon\xi}^{-1} \tag{III.3.12}$$

où R et $F_{\varepsilon\xi}$ commutent. Bien sûr, on sait faire cette construction pour $\varepsilon\xi \ll 1$, mais on cherche désormais à imposer en outre un comportement en $\varepsilon\xi \to \infty$ et à s'assurer que la décomposition existe pour tout $\varepsilon\xi \in \mathbb{R}_+$. Pour chaque coefficient $(\omega_{ij}(\varepsilon\xi))_{1\leq i,j\leq 2}$ de la matrice $\Omega_{\varepsilon\xi}$, les méthodes de formes normales permettent de calculer les coefficients des séries entières

$$\omega_{ij}(\varepsilon\xi) = \sum_{n\geq 0} \alpha_{ij}^{(n)} (\varepsilon\xi)^n \qquad \text{et} \qquad \omega_{ij}(\varepsilon\xi) = \sum_{n\geq 0} \beta_{ij}^{(n)} (\varepsilon\xi)^{-n}$$

en $\varepsilon\xi \to 0$ et $\varepsilon\xi \to \infty$ respectivement. La même chose est vraie pour les coefficients de $F_{\varepsilon\xi}$. Graphiquement, le principe est présenté en Figure III.1. Les séries (tronquées) sont tracées en pointillées et on les compare aux valeurs propres exactes. On remarque en outre avec une implémentation que les changements de variable « approchés » sont de déterminant positif sur leur domaine de validité.

Le principe de reconstruire une fonction $x \mapsto \widetilde{\omega}_{ij}(x)$ à partir de séries connues en $x \to 0$ et en $x \to \infty$ n'est pas nouveau. Une méthode de référence à cet effet est d'utiliser un approximant de Padé en deux points, dont la littérature est expensive [JM80 ; JNT83 ; BG96 ; DF01 ; TLH12]. Il faut néanmoins noter que ces constructions concernent seule-

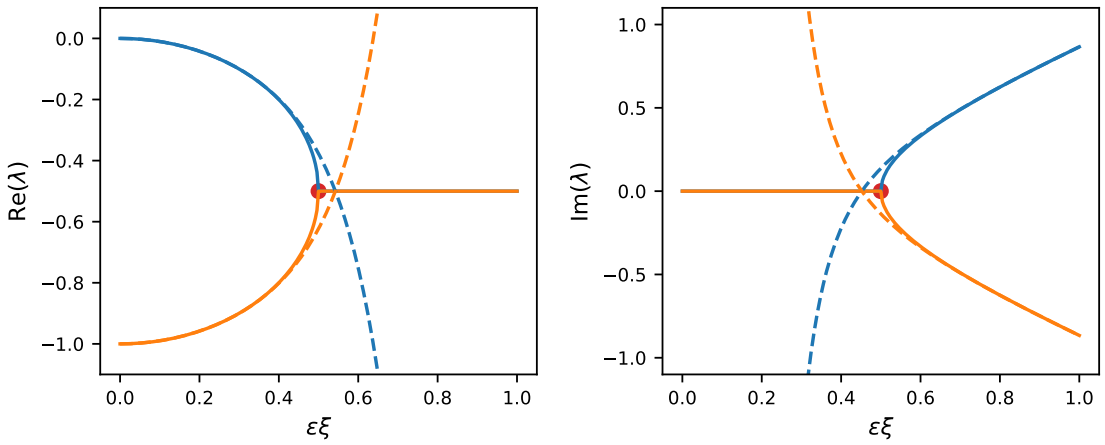


FIGURE III.1 – Tracé des valeurs propres du problème du télégraphe à un paramètre (III.3.11) en fonction de $\varepsilon\xi$, avec en bleu λ_+ et en orange λ_- et leurs approximations en pointillées.

ment des fonctions à valeurs scalaires, et que la littérature pour les approximations de Padé de matrices semble se limiter à quelques articles d'André Draux, [Dra84 ; Dra88]. Il n'est pas immédiat que l'approximation coefficient par coefficient présente des propriétés idéales (e.g. le changement de variable bien déterminé pour tout ε, ξ). On espère néanmoins pouvoir utiliser ces méthodes pour construire des multiplicateurs de Fourier adaptés au problème du télégraphe.

Annexe A

Autour d'un développement double-échelle

Dans ce chapitre, on discute rapidement d'une démarche de développement double-échelle pour les problèmes à relaxation rapide de la forme

$$\partial_t u = -\frac{1}{\varepsilon} A u + f(u), \qquad u(0) = u_0 \in \mathbb{R}^d \tag{A.1}$$

avec A un opérateur de relaxation et $u \mapsto f(u)$ un champ de vecteurs régulier. L'idée du développement double-échelle est d'écrire $t \mapsto u(t)$ comme une évaluation particulière d'une fonction à deux variables

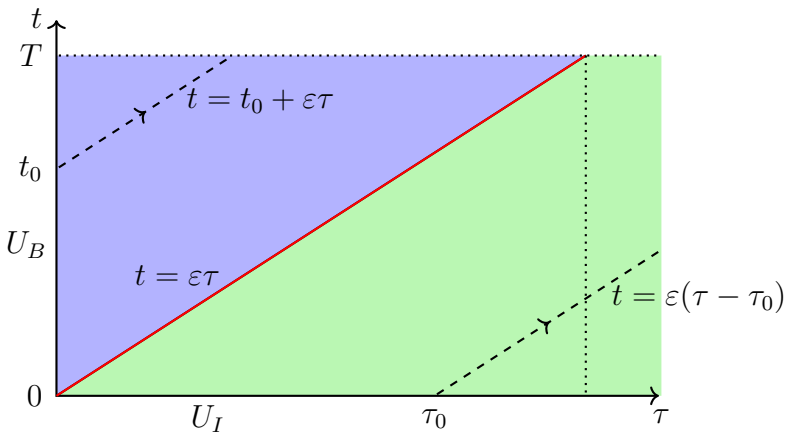
$$u(t) = U(t, \tau)|_{\tau = t/\varepsilon}. \tag{A.2}$$

Cette approche a été un succès dans le contexte de problèmes hautement oscillants (voir [CCLM15 ; CLMZ20]), et est à la base des méthodes d'homogénéisation [All92]. Dans ce contexte, on suppose que la nouvelle quantité $(t, \tau) \mapsto U(t, \tau)$ est périodique par rapport à la seconde variable τ.

Dans le cadre de problèmes à relaxation rapide, on suppose $\tau \in \mathbb{R}_+$, et on injecte (A.2) dans (A.1) pour obtenir le problème de transport pour $(t, \tau) \in [0, T] \times \mathbb{R}_+$

$$\partial_t U + \frac{1}{\varepsilon} (\partial_\tau + A) U = f(U). \tag{A.3}$$

De part la structure du problème, il faut définir une donnée initiale $\tau \mapsto U_I(\tau) := U(0, \tau)$ et une donnée au bord $t \mapsto U_B(t) := U(t, 0)$, tel que sur le diagramme suivant :



Les valeurs qui nous intéressent sont sur la ligne caractéristique $t = \varepsilon\tau$, en rouge. La donnée initiale U_I permet de déterminer les valeurs de U dans le domaine vert sous cette caractéristique, tandis que la donnée au bord U_B détermine les valeurs dans le domaine bleu au-dessus de la caractéristique.

Caractère bien posé du problème

Discutons d'abord du caractère bien posé d'un problème de la forme (A.3). Le caractère bien posé de ce problème se fait naturellement le long des caractéristiques, qui s'écrivent $t = t_0 + \varepsilon\tau$ au-dessus de la diagonale $t = \varepsilon\tau$, et $t = \varepsilon(\tau - \tau_0)$ en-dessous, avec $(t_0, \tau_0) \in [0, T] \times \mathbb{R}_+$. Ainsi, si les données U_I et U_B sont bornées, quitte à réduire le temps d'existence T, le problème est bien posé. Cependant, avec cette approche, on ne précise pas la régularité de la solution obtenue, le problème est bien posé dans $L^\infty([0, T] \times \mathbb{R}_+)$. Si en outre U_B et U_I sont continues avec $U_B(0) = U_I(0) = u_0$, le problème est bien posé dans $C^0([0, T] \times \mathbb{R}_+)$.

Si on veut que le problème soit bien posé dans C^1, il faut que U_B et U_I soient continues, mais aussi qu'à la limite $t \to 0$ et $\tau \to 0$, ces données soient compatibles. En effet, dans (A.3), on a en $(t, \tau) = (0, 0)$,

$$\partial_t U_B(0) + \frac{1}{\varepsilon}(\partial_\tau + A)U_I(0) = f(u_0).$$

Si cette condition est respectée, on remarque que la dérivée $V = \partial_t U$ vérifie le problème de transport

$$\partial_t V + \frac{1}{\varepsilon}(\partial_\tau + A)V = f'(U)V.$$

Les conditions initiales et au bord associé à ce problème sont continues, donc cette dérivée

est continue. Le même raisonnement peut être tenu pour $\partial_\tau U$, si bien que la fonction $(t,\tau) \mapsto U(t,\tau)$ est de classe C^1. À l'ordre 2, on trouve

$$\partial_t^2 U_B(0) - \frac{1}{\varepsilon^2}(\partial_\tau + A)^2 U_I(0) = f'(u_0)\partial_t U_B(0) - \frac{1}{\varepsilon}(\partial_\tau + A) f(U_B(\tau))|_{\tau=0},$$

et on peut dériver des conditions similaires aux ordres supérieurs. Si les données initiale et au bord sont de régularité C^n avec des dérivées bornées à tout ordre (jusqu'à n), et que les conditions de compatibilité sont respectées jusqu'à l'ordre n, alors, quitte à modifier le temps T d'existence de la solution, le problème (A.3) est bien posé dans $C^n([0,T]\times\mathbb{R}_+)$.

Motivation et calcul numérique

Évidemment, on ne peut pas résoudre le problème directement le long d'une caractéristique, puisque cela revient à résoudre un problème de la forme (A.1). En revanche, on peut choisir la donnée initiale et la donnée au bord de sorte que le problème soit bien posé au sens où un certain nombre de dérivées $\partial_t U$ soient uniformément bornées. Par exemple en choisissant

$$U(0,\tau) = e^{-\tau A} u_0,$$

on a $\partial_t U(0,\tau) = f(e^{-\tau A}u_0)$. Cependant à l'ordre supérieur on peut calculer

$$\partial_t^2 U(0,\tau) = \partial_u f(e^{-\tau A}u_0) f(e^{-\tau A}u_0) + \frac{1}{\varepsilon}[f,A](e^{-\tau A}u_0)$$

avec $[f,A] = \partial_u f \cdot A - Af$. En supposant que A et f ne commutent pas (sinon on pourrait simplement faire du splitting sur le problème d'origine), cette donnée est raide, et quelle que soit la donnée au bord qu'on choisira, le problème ne sera pas uniformément (i.e. indépendamment de ε) bien posé dans $C^2([0,T]\times\mathbb{R}_+)$.

Supposons maintenant que l'on dispose d'une donnée initiale $\tau \mapsto U_I(\tau)$ et d'une donnée au bord $t \mapsto U_B(t)$ compatibles à l'ordre 2 et telles que ces données et leurs dérivées jusqu'à l'ordre 2 soient uniformément bornées. Alors en supposant qu'on connait de manière exacte [1] le comportement par rapport à τ d'une approximation de la solution à un temps t_n, $U_n(\tau) \approx U(t_n,\tau)$, alors on peut calculer la solution au temps $t_{n+1} = t_n + \Delta t$

1. Si le comportement selon τ n'est pas connu parfaitement, le problème devient plus complexe à résoudre car il faut discrétiser l'espace de cette variable, et les schémas numériques doivent prendre cette discrétisation en compte dans la donnée au bord. Voir par exemple [BNSTC19].

avec la formule

$$\frac{U_{n+1}(\tau) - U_n(\tau)}{\Delta t} + \frac{1}{\varepsilon}(\partial_\tau + A)U_{n+1}(\tau) = f(U_n(\tau)),$$

soit de manière explicite en posant $\mu = \varepsilon/\Delta t$,

$$U_{n+1}(\tau) = e^{-\tau(\mu I + A)} U_B(t_{n+1}) + \mu \int_0^\tau e^{(\sigma-\tau)(\mu I + A)} \Big(U_n(\sigma) + \Delta t \, f\Big(U_n(\sigma)\Big)\Big) \, \mathrm{d}\sigma.$$

En s'inspirant des preuves de [CCLM15], on peut alors prouver que si la dérivée seconde $\partial_t^2 U$ peut être bornée indépendamment de ε, alors il existe C et Δt_0 indépendants de ε, tels que pour toute discrétisation $(t_n)_{0\leq n\leq N}$ de pas de temps $\Delta t < \Delta t_0$, l'erreur du schéma prend la forme

$$\sup_{\tau\in\mathbb{R}_+} |U_n - U(t_n, \cdot)| \leq C\Delta t \sup_{(t,\tau)\in[0,T]\times\mathbb{R}_+} |\partial_t^2 U(t,\tau)|$$

pour tout $0 \leq n \leq N$.

De la même manière, si le problème est uniformément bien posé dans C^3, on peut construire un schéma à deux étages d'ordre 2, comme cela a pu être proposé dans un cadre hautement oscillant chez [CCLM15]. On peut certainement construire des schémas d'ordre plus élevé en s'inspirant des méthodes de type Adams-Bashforth utilisées dans [CLMZ20]. Néanmoins, il faut pour cela être capable de construire des données initiale et au bord pour que le problème soit uniformément bien posé dans C^2, C^3 ou dans un espace encore plus régulier.

Choix de la donnée initiale

Un moyen de trouver un problème uniformément bien posé est d'écrire $U(t,\tau)$ comme une série en puissances de ε,

$$U(t,\tau) = \sum_{n\geq 0} \varepsilon^n U_n(t,\tau).$$

On effectue ensuite une séparation d'échelles afin de trouver, pour les premiers termes,

$$\begin{cases} \partial_\tau U_0 + AU_0 = 0 & (\varepsilon^{-1}) \\ \partial_\tau U_1 + AU_1 = f(U_0) - \partial_t U_0 & (\varepsilon^0) \\ \partial_\tau U_2 + AU_2 = f'(U_0)U_1 - \partial_t U_1 & (\varepsilon^1) \\ \quad \vdots \end{cases}$$

Ainsi on retrouve le premier terme,

$$U_0(t,\tau) = e^{-\tau A}\widetilde{u}_0(t)$$

et si on néglige les termes d'ordre ε et au-delà dans la série, on peut choisir $t \mapsto u_0(t)$ librement. Si on cherche à étendre le développement, en revanche, on trouve

$$U_1(t,\tau) = e^{-\tau A}\widetilde{u}_1(t) + \int_0^\tau e^{(\sigma-\tau)A}\left(f(e^{-\tau A}\widetilde{u}_0(t)) - e^{-\tau A}\partial_t\widetilde{u}_0(t)\right)\mathrm{d}\sigma.$$

Si on veut que cette quantité soit sous forme d'une série exponentielle (cette forme serait avantageuse pour l'implémentation, car on peut alors discrétiser τ de manière adaptée), il faut choisir $\partial_t\widetilde{u}_0$ comme étant la composante en $e^{-\tau A}$ de $f(e^{-\tau A}\widetilde{u}_0)$. La même chose se produit à l'ordre suivant, où $\partial_t\widetilde{u}_1$ est déterminé comme étant la composante en $e^{-\tau A}$ de $f'(U_0)U_1$, etc. Les valeurs initiales $\widetilde{u}_k(0)$ sont encore libres.

Enfin, on peut tronquer la série afin de choisir une donnée initiale

$$U_I^{[n]}(\tau) = \sum_{k=0}^{n}\varepsilon^k U_n(0,\tau).$$

Il faut ensuite déterminer les données initiales $\widetilde{u}_k(0)$ de sorte à obtenir un problème bien posé dans $C^{n+1}([0,T]\times\mathbb{R}_+)$, et résoudre les problèmes associés à chacune de ces variables pour obtenir une donnée au bord

$$U_B(t) = \sum_{k=0}^{n}\varepsilon^k\widetilde{u}_k(t).$$

Nous n'avons cependant pas de méthode pour trouver ces données au bord, et nous n'avons pas non plus étudié les conséquences d'avoir des données compatibles « à ε^n près. »

Remarque A.1. *On serait tenté de penser que la définition de $U_B(t)$ a peu d'importance tant que les conditions de compatibilité sont vérifiées. Cependant, cette définition est certainement cruciale pour pouvoir implémenter la méthode. En effet, en ne considérant que les conditions de compatibilité, on pourrait définir pour un problème dans $C^2([0,T]\times\mathbb{R}_+)$*

$$U_B(t) = u_0 + t\ \partial_t U_B(0) + \frac{t^2}{2}\partial_t{}^2 U_B(0),$$

mais alors à un temps $t > 0$ fixé, on observe que la solution $\tau \mapsto U(t,\tau)$ comporte deux parties de comportement distinct de part et d'autre de t/ε. La partie de gauche ressemble à

une parabole, de la forme $\alpha(\varepsilon\tau)^2+\beta\varepsilon\tau+\gamma$, tandis que la partie de droite est exponentielle. La discrétisation de la variable τ doit alors capturer ces deux comportements, tandis qu'un choix judicieux de $U_B(t)$ permet de conserver une forme exponentielle pour tout (t,τ).

On retrouve ainsi formellement la même construction que dans le Chapitre II. La donnée initiale U_I peut en effet s'écrire

$$U_I(\tau) = \Omega_\tau^\varepsilon \circ \left(\Omega_0^\varepsilon\right)^{-1}.$$

Ce lien entre décomposition micro-macro et décomposition double-échelle est également soulevé dans le cadre périodique chez [CLMZ20]. La nouveauté par rapport à ce cadre est la donnée au bord $t \mapsto U_B(t)$, et on a vu que la bonne détermination de cette donnée est problématique. Tracer un lien entre cette donnée au bord et la construction micro-macro donnerait certainement des pistes pour déterminer les valeurs initiales $\widetilde{u}_k(0)$ associées au problèmes au bord.

Annexe B

Présentation de schémas exponentiels

On présente ici les schémas exponentiel Runge-Kutta utilisés en Chapitre II. Il ne s'agit pas de discuter des preuves de convergence des schémas, mais de fournir les outils nécessaires pour l'implémentation de ceux-ci. Tous les schémas décrits ici sont disponibles dans [HO05], on adapte seulement les notations de cet article au problème qui nous concerne, c'est-à-dire un problème de la forme

$$\partial_t u = -\frac{1}{\varepsilon} A u + f(u), \qquad u(0) = u_0 \in \mathbb{R}^d \tag{B.1}$$

avec $(-A)$ un qui engendre un semi-groupe $t \mapsto e^{-tA}$ uniformément borné pour $t \in \mathbb{R}_+$ et $u \mapsto f(u)$ régulière. Il est à noter que dans [HO05], les auteurs mentionnent une problématique de réduction d'ordre. En particulier, ils décrivent les schémas de Cox & Matthews dans [CM02] (d'ordres 3 et 4) comme pouvant entraîner une réduction à l'ordre 2 du fait d'un non-respect des conditions d'ordre des schémas. Cette notion est à distinguer de la réduction d'ordre qu'on observe en Introduction et en Chapitre II, qui provient d'une dégénérescence de l'erreur à cause des dérivées de la solution qui ne sont pas bornées.

Entre un temps t_n et un temps $t_{n+1} = t_n + \Delta t$, on considère qu'on dispose d'une approximation $u_n \approx u(t_n)$ et on cherche à trouver $u_{n+1} \approx u(t_{n+1})$. La méthode utilisé est un schéma de Runge-Kutta explicite exponentiel à s étages, qui peut s'écrire

$$\begin{aligned} u_{n+1} &= e^{-\Delta t\, A/\varepsilon} u_n + \Delta t \sum_{i=1}^{s} b_i(-\tfrac{\Delta t}{\varepsilon} A) f(U_{ni}), \\ U_{ni} &= e^{-c_i \Delta t\, A/\varepsilon} u_n + \Delta t \sum_{j=1}^{i-1} a_{ij}(-\tfrac{\Delta t}{\varepsilon} A) f(U_{nj}), \qquad i = 1, \dots, s. \end{aligned}$$

La différence par rapport à un schéma standard est que les coefficients du schéma dépendent de l'opérateur A qui est « filtré ».

Étant donné que l'argument au sein des coefficients est toujours le même, on peut écrire ce schéma sous forme de tableau de Butcher, de la même manière que les méthodes de Runge-Kutta standards :

$$\begin{array}{c|cccc} 0 & 0 & & & \\ c_2 & a_{21} & 0 & & \\ \vdots & \vdots & \ddots & \ddots & \\ c_s & a_{s1} & \cdots & a_{s,s-1} & 0 \\ \hline & b_1 & \cdots & \cdots & b_s \end{array}$$

Par exemple la méthode d'Euler explicite exponentielle

$$u_{n+1} = e^{-\Delta t\, A/\varepsilon} u_n + \left(\int_0^{\Delta t} e^{(\Delta t - \tau)A/\varepsilon} \mathrm{d}\tau \right) f(u_n)$$

devient sous forme de tableau, en définissant $\varphi_1(-hA) = h^{-1} \int_0^h e^{(h-\tau)A} \mathrm{d}\tau$,

$$\begin{array}{c|c} 0 & 0 \\ \hline & \varphi_1 \end{array}$$

avec le raccourci $\varphi_1 = \varphi_1(-\frac{\Delta t}{\varepsilon} A)$. Pour les tableaux suivants, on définit

$$\varphi_j(-hA) = h^{-j} \int_0^h \frac{\tau^{j-1}}{(j-1)!} e^{(h-\tau)A} \mathrm{d}\tau,$$

ce qui correspond au reste intégral renormalisé dans le développement de l'exponentiel. De sorte à avoir des coefficients dont l'argument est toujours $-\frac{\Delta t}{\varepsilon} A$ dans les tableaux, on définit aussi

$$\varphi_{i,j}(-hA) = \varphi_j(-c_i hA).$$

On peut maintenant énoncer les tableaux de Butcher des schémas utilisés dans le Chapitre II. Le schéma d'ordre 2 choisi est

$$\begin{array}{c|cc} 0 & & \\ \frac{1}{2} & \frac{1}{2}\varphi_{1,2} & \\ \hline & \varphi_1 - 2\varphi_2 & 2\varphi_2 \end{array}$$

Pour le schéma d'ordre 3, on choisit celui de Cox et Matthews [CM02],

0			
$\frac{1}{2}$	$\frac{1}{2}\varphi_{1,2}$		
1	$-\varphi_{1,3}$	$2\varphi_{1,3}$	
	$4\varphi_3 - 3\varphi_2 + \varphi_1$	$-8\varphi_3 + 4\varphi_2$	$4\varphi_3 - \varphi_2$

de même que celui d'ordre 4,

0				
$\frac{1}{2}$	$\frac{1}{2}\varphi_{1,2}$			
$\frac{1}{2}$	0	$\frac{1}{2}\varphi_{1,3}$		
1	$\frac{1}{3}(\varphi_{0,3} - 1)$	0	$\varphi_{1,3}$	
	$\varphi_1 - 3\varphi_2 + 4\varphi_3$	$2\varphi_2 - 4\varphi_3$	$2\varphi_2 - 4\varphi_3$	$4\varphi_3 - \varphi_2$

Dans [HO05], d'autres schémas sont proposés, mais après implémentation le comportement de l'erreur lors de la résolution de (B.1) est le même.

BIBLIOGRAPHIE

[ACM99] Georgios AKRIVIS, Michel CROUZEIX et Charalambos MAKRIDAKIS, « Implicit-explicit multistep methods for quasilinear parabolic equations », *Numerische Mathematik* 82.*4* (1999), Publisher : Springer, p. 521-541 (cf. p. 22, 52).

[ADP20] Giacomo ALBI, Giacomo DIMARCO et Lorenzo PARESCHI, « Implicit-explicit multistep methods for hyperbolic systems with multiscale relaxation », *SIAM Journal on Scientific Computing* 42.*4* (2020), Publisher : SIAM, A2402-A2435 (cf. p. 23, 27, 67).

[All92] Grégoire ALLAIRE, « Homogenization and two-scale convergence », *SIAM Journal on Mathematical Analysis* 23.*6* (1992), p. 1482-1518 (cf. p. 95).

[ARW95] Uri M ASCHER, Steven J RUUTH et Brian TR WETTON, « Implicit-explicit methods for time-dependent partial differential equations », *SIAM Journal on Numerical Analysis* 32.*3* (1995), Publisher : SIAM, p. 797-823 (cf. p. 22, 52).

[AP96] Pierre AUGER et Jean-Christophe POGGIALE, « Emergence of population growth models : fast migration and slow growth », *Journal of Theoretical Biology* 182.*2* (1996), Publisher : Elsevier, p. 99-108 (cf. p. 7, 50).

[BG96] George A BAKER JR et Peter GRAVES-MORRIS, *Pade Approximants : Encyclopedia of Mathematics and It's Applications, Vol. 59 George A. Baker, Jr., Peter Graves-Morris*, t. 59, Cambridge University Press, 1996 (cf. p. 92).

[Bam03] Dario BAMBUSI, « Birkhoff normal form for some nonlinear PDEs », *Communications in mathematical physics* 234.*2* (2003), Publisher : Springer, p. 253-285 (cf. p. 13, 31).

[Bam06] Dario BAMBUSI, « Birkhoff normal form for some quasilinear Hamiltonian PDEs », *XIVth International Congress on Mathematical Physics*, World Scientific, 2006, p. 273-280 (cf. p. 31).

[Bam08] Dario Bambusi, « A Birkhoff normal form theorem for some semilinear PDEs », *Hamiltonian dynamical systems and applications*, Springer, 2008, p. 213-247 (cf. p. 31).

[BB05] Dario Bambusi et Massimiliano Berti, « A Birkhoff–Lewis-Type Theorem for Some Hamiltonian PDEs », *SIAM Journal on Mathematical Analysis* 37.*1* (2005), Publisher : SIAM, p. 83-102 (cf. p. 31).

[BCZ14] Weizhu Bao, Yongyong Cai et Xiaofei Zhao, « A Uniformly Accurate Multiscale Time Integrator Pseudospectral Method for the Klein–Gordon Equation in the Nonrelativistic Limit Regime », *SIAM Journal on Numerical Analysis* 52.*5* (jan. 2014), p. 2488-2511, DOI : 10.1137/130950665 (cf. p. 29).

[BD12] Weizhu Bao et Xuanchun Dong, « Analysis and comparison of numerical methods for the Klein–Gordon equation in the nonrelativistic limit regime », *Numerische Mathematik* 120.*2* (2012), p. 189-229 (cf. p. 7).

[BZ19] Weizhu Bao et Xiaofei Zhao, « Comparison of numerical methods for the nonlinear Klein-Gordon equation in the nonrelativistic limit regime », *Journal of Computational Physics* 398 (déc. 2019), p. 108886, DOI : 10.1016/j.jcp.2019.108886 (cf. p. 29).

[BV20] Guillaume Bertoli et Gilles Vilmart, « Strang splitting method for semilinear parabolic problems with inhomogeneous boundary conditions : a correction based on the flow of the nonlinearity », *SIAM Journal on Scientific Computing* 42.*3* (2020), A1913-A1934 (cf. p. 20).

[BGK54] Prabhu Lal Bhatnagar, Eugene P Gross et Max Krook, « A model for collision processes in gases. I. Small amplitude processes in charged and neutral one-component systems », *Physical review* 94.*3* (1954), p. 511 (cf. p. 7).

[BPR17] Sebastiano Boscarino, Lorenzo Pareschi et Giovanni Russo, « A unified IMEX Runge–Kutta approach for hyperbolic systems with multiscale relaxation », *SIAM Journal on Numerical Analysis* 55.*4* (2017), Publisher : SIAM, p. 2085-2109 (cf. p. 23, 24, 27, 67).

[BR09] Sebastiano Boscarino et Giovanni Russo, « On a class of uniformly accurate IMEX Runge–Kutta schemes and applications to hyperbolic systems with relaxation », *SIAM Journal on Scientific Computing* 31.*3* (2009), p. 1926-1945 (cf. p. 24).

[Bou96] J. Bourgain, « Construction of approximative and almost periodic solutions of perturbed linear schrödinger and wave equations », *Geometric & Functional Analysis GAFA* 6.*2* (mar. 1996), p. 201-230, doi : 10.1007/BF02247885 (cf. p. 31).

[BNSTC19] Benjamin Boutin, Thi Hoai Thuong Nguyen, Abraham Sylla, Sébastien Tran-Tien et Jean-François Coulombel, « High order numerical schemes for transport equations on bounded domains », *arXiv preprint arXiv :1912.03097* (2019) (cf. p. 97).

[Bri26] Léon Brillouin, « Remarques sur la mécanique ondulatoire », *J. phys. radium* 7.*12* (1926), p. 353-368 (cf. p. 30).

[CR17] Begoña Cano et Nuria Reguera, « Avoiding order reduction when integrating nonlinear Schrödinger equation with Strang method », *Journal of Computational and Applied Mathematics* 316 (2017), p. 86-99 (cf. p. 20).

[Car82] Jack Carr, *Applications of centre manifold theory*, t. 35, Applied Mathematical Sciences, Springer-Verlag New York, 1982 (cf. p. 9, 51).

[CCM19] Fernando Casas, Philippe Chartier et Ander Murua, « Continuous changes of variables and the Magnus expansion », *Journal of Physics Communications* 3.*9* (sept. 2019), p. 095014, doi : 10.1088/2399-6528/ab42c1 (cf. p. 30).

[CCMM15] F. Castella, Ph. Chartier, F. Méhats et A. Murua, « Stroboscopic Averaging for the Nonlinear Schrödinger Equation », *Foundations of Computational Mathematics* 15.*2* (avr. 2015), p. 519-559, doi : 10.1007/s10208-014-9235-7 (cf. p. 26, 29, 30, 42, 52, 63).

[CCS18] Francois Castella, Philippe Chartier et Julie Sauzeau, « Analysis of a time-dependent problem of mixed migration and population dynamics », *arXiv preprint, arXiv :1512.01880* (2018) (cf. p. 7, 50).

[CCS16] François Castella, Philippe Chartier et Julie Sauzeau, « A formal series approach to the center manifold theorem », *Foundations of Computational Mathematics* (2016), Publisher : Springer, p. 1-38 (cf. p. 10, 13, 52, 76, 78).

[CMS10] P. Chartier, A. Murua et J. M. Sanz-Serna, « Higher-Order Averaging, Formal Series and Numerical Integration I : B-series », *Foundations of Computational Mathematics* 10.*6* (déc. 2010), p. 695-727, doi : 10.1007/s10208-010-9074-0 (cf. p. 26, 30, 32).

[CCLM15] Philippe Chartier, Nicolas Crouseilles, Mohammed Lemou et Florian Méhats, « Uniformly accurate numerical schemes for highly oscillatory Klein–Gordon and nonlinear Schrödinger equations », *Numerische Mathematik* 129.*2* (2015), Publisher : Springer, p. 211-250 (cf. p. 13, 26, 29, 95, 98).

[CCLM20] Philippe Chartier, Nicolas Crouseilles, Mohammed Lemou et Florian Méhats, « Averaging of highly-oscillatory transport equations », *Kinetic & Related Models* 13.*6* (2020), p. 1107, doi : 10.3934/krm.2020039 (cf. p. 32).

[CCLMZ20] Philippe Chartier, Nicolas Crouseilles, Mohammed Lemou, Florian Méhats et Xiaofei Zhao, « Uniformly Accurate Methods for Three Dimensional Vlasov Equations under Strong Magnetic Field with Varying Direction », *SIAM Journal on Scientific Computing* 42.*2* (jan. 2020), B520-B547, doi : 10.1137/19M127402X (cf. p. 29).

[CHV10] Philippe Chartier, Ernst Hairer et Gilles Vilmart, « Algebraic structures of B-series », *Foundations of Computational Mathematics* 10.*4* (2010), p. 407-427 (cf. p. 10).

[CLMV20] Philippe Chartier, Mohammed Lemou, Florian Méhats et Gilles Vilmart, « A New Class of Uniformly Accurate Numerical Schemes for Highly Oscillatory Evolution Equations », *Foundations of Computational Mathematics* 20.*1* (fév. 2020), p. 1-33, doi : 10.1007/s10208-019-09413-3 (cf. p. 26, 29, 30, 52, 59, 63, 88, 91).

[CLMZ20] Philippe Chartier, Mohammed Lemou, Florian Méhats et Xiaofei Zhao, « Derivative-free high-order uniformly accurate schemes for highly-oscillatory systems », *submitted preprint* (2020) (cf. p. 32, 52, 88, 95, 98, 100).

[CLT21] Philippe Chartier, Mohammed Lemou et Léopold Trémant, « A uniformly accurate numerical method for a class of dissipative systems », à paraître dans *Mathematics of Computation* (2021) (cf. p. 27).

[CMTZ17] Philippe Chartier, Florian Méhats, Mechthild Thalhammer et Yong Zhang, « Convergence of multi-revolution composition time-splitting methods for highly oscillatory differential equations of Schrödinger type », *ESAIM : Mathematical Modelling and Numerical Analysis* 51.*5* (sept. 2017), p. 1859-1882, doi : 10.1051/m2an/2017010 (cf. p. 31).

[CMS12a] Philippe Chartier, Ander Murua et Jesus Maria Sanz-Serna, « A formal series approach to averaging : exponentially small error estimates », *Discrete and Continuous Dynamical Systems-Series A* 32.*9* (2012) (cf. p. 26, 30).

[CMS12b] Philippe Chartier, Ander Murua et Jesus Maria Sanz-Serna, « Higher-order averaging, formal series and numerical integration II : the quasi-periodic case », *Foundations of Computational Mathematics* 12.*4* (2012), Publisher : Springer, p. 471-508 (cf. p. 31).

[CMS15] Philippe Chartier, Ander Murua et Jesus Maria Sanz-Serna, « Higher-order averaging, formal series and numerical integration III : error bounds », *Foundations of Computational Mathematics* 15.*2* (2015), Publisher : Springer, p. 591-612 (cf. p. 42).

[CKSTT10] James Colliander, Markus Keel, Gigiola Staffilani, Hideo Takaoka et Terence Tao, « Transfer of energy to high frequencies in the cubic defocusing nonlinear Schrödinger equation », *Inventiones mathematicae* 181.*1* (2010), Publisher : Springer, p. 39-113 (cf. p. 31).

[CKO12] James Colliander, Soonsik Kwon et Tadahiro Oh, « A remark on normal forms and the "upside-down" I-method for periodic NLS : growth of higher Sobolev norms », *Journal d'Analyse Mathématique* 118.*1* (2012), Publisher : Springer, p. 55-82 (cf. p. 31).

[CM02] Steven M Cox et Paul C Matthews, « Exponential time differencing for stiff systems », *Journal of Computational Physics* 176.*2* (2002), p. 430-455 (cf. p. 101, 102).

[CEM20] Nicolas Crouseilles, Lukas Einkemmer et Josselin Massot, « Exponential methods for solving hyperbolic problems with application to collisionless kinetic equations », *Journal of Computational Physics* 420 (2020), p. 109688 (cf. p. 90).

[CJL17] Nicolas Crouseilles, Shi Jin et Mohammed Lemou, « Nonlinear geometric optics method-based multi-scale numerical schemes for a class of highly oscillatory transport equations », *Mathematical Models and Methods in Applied Sciences* 27.*11* (2017), Publisher : World Scientific, p. 2031-2070 (cf. p. 7, 26, 73).

[Cro80] Michel Crouzeix, « Une méthode multipas implicite-explicite pour l'approximation des équations d'évolution paraboliques », *Numerische Mathematik* 35.*3* (1980), p. 257-276 (cf. p. 22).

[Deg04] Pierre Degond, « Macroscopic limits of the Boltzmann equation : a review », *Modeling and computational methods for kinetic equations* (2004), p. 3-57 (cf. p. 13).

[DM04] Stéphane Descombes et Marc Massot, « Operator splitting for nonlinear reaction-diffusion systems with an entropic structure : singular perturbation and order reduction », *Numerische Mathematik* 97.*4* (2004), p. 667-698 (cf. p. 18).

[DP11] Giacomo Dimarco et Lorenzo Pareschi, « Exponential Runge–Kutta methods for stiff kinetic equations », *SIAM Journal on Numerical Analysis* 49.*5* (2011), Publisher : SIAM, p. 2057-2077 (cf. p. 27, 81).

[DP17] Giacomo Dimarco et Lorenzo Pareschi, « Implicit-explicit linear multistep methods for stiff kinetic equations », *SIAM Journal on Numerical Analysis* 55.*2* (2017), p. 664-690 (cf. p. 22, 27).

[Dra84] André Draux, « The Padé approximants in a non-commutative algebra and their applications », *Padé Approximation and its Applications Bad Honnef 1983*, Springer, 1984, p. 117-131 (cf. p. 93).

[Dra88] André Draux, « Convergence of Padé approximants in a non-commutative algebra », *Approximation and Optimization*, Springer, 1988, p. 118-130 (cf. p. 93).

[DF01] Tobin A Driscoll et Bengt Fornberg, « A Padé-based algorithm for overcoming the Gibbs phenomenon », *Numerical Algorithms* 26.*1* (2001), p. 77-92 (cf. p. 92).

[EO15] Lukas Einkemmer et Alexander Ostermann, « Overcoming order reduction in diffusion-reaction splitting. Part 1 : Dirichlet boundary conditions », *SIAM Journal on Scientific Computing* 37.*3* (2015), A1577-A1592 (cf. p. 20).

[FOS15] Erwan Faou, Alexander Ostermann et Katharina Schratz, « Analysis of exponential splitting methods for inhomogeneous parabolic equations », *IMA Journal of Numerical Analysis* 35.*1* (2015), p. 161-178 (cf. p. 20).

[For92] Joseph Ford, « The Fermi-Pasta-Ulam problem : paradox turns discovery », *Physics Reports* 213.*5* (1992), p. 271-310 (cf. p. 7).

[FS00] E. Frénod et E. Sonnendrücker, « Long time behavior of the two-dimensional vlasov equation with a strong external magnetic field », *Mathematical Models and Methods in Applied Sciences* 10.*4* (juin 2000), p. 539-553, doi : 10.1142/S021820250000029X (cf. p. 29).

[FSS09] Emmanuel Frénod, Francesco Salvarani et Eric Sonnendrücker, « Long time simulation of a beam in a periodic focusing channel via a two-scale pic-method », *Mathematical Models and Methods in Applied Sciences* 19.*2* (fév. 2009), p. 175-197, doi : 10.1142/S0218202509003395 (cf. p. 29).

[GSS98] Bosco Garcia-Archilla, Jesús Maria Sanz-Serna et Robert D Skeel, « Long-time-step methods for oscillatory differential equations », *SIAM Journal on Scientific Computing* 20.*3* (1998), Publisher : SIAM, p. 930-963 (cf. p. 29).

[GM03] Thierry Goudon et Antoine Mellet, « Homogenization and diffusion asymptotics of the linear Boltzmann equation », *ESAIM : Control, Optimisation and Calculus of Variations* 9 (2003), p. 371-398 (cf. p. 13).

[GT12] Benoit Grébert et Laurent Thomann, « Resonant dynamics for the quintic nonlinear Schrödinger equation », *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, t. 29, Issue : 3, Elsevier, 2012, p. 455-477 (cf. p. 31).

[GV11] Benoit Grébert et Carlos Villegas-Blas, « On the energy exchange between resonant modes in nonlinear Schrödinger equations », *Annales de l'Institut Henri Poincaré C, Analyse non linéaire* 28.*1* (jan. 2011), p. 127-134, doi : 10.1016/j.anihpc.2010.11.004 (cf. p. 7, 31).

[GHM94] Günther Greiner, JAP Heesterbeek et Johan AJ Metz, « A singular perturbation theorem for evolution equations and time-scale arguments for structured population models », *Canadian applied mathematics quarterly* 3.*4* (1994), Publisher : Applied mathematics institute of the University of Alberta, p. 435-459 (cf. p. 7, 50).

[HLW06] Ernst Hairer, Christian Lubich et Gerhard Wanner, *Geometric Numerical Integration : Structure-Preserving Algorithms for Ordinary Differential Equations*, 2e éd., Springer Series in Computational Mathematics, Berlin Heidelberg : Springer-Verlag, 2006, doi : 10.1007/3-540-30666-8 (cf. p. 7, 10, 38).

[HW96] Ernst Hairer et Gerhard Wanner, *Solving ordinary differential equations II. Stiff and Differential-Algebraic Problems*, Springer Berlin Heidelberg, 1996 (cf. p. 9, 17, 50).

[HH64] Michel Hénon et Carl Heiles, « The applicability of the third integral of motion : some numerical experiments », *The astronomical journal* 69 (1964), p. 73 (cf. p. 7).

[HLO20] Marlis Hochbruck, Jan Leibold et Alexander Ostermann, « On the convergence of Lawson methods for semilinear stiff problems », *Numerische Mathematik* 145 (2020), p. 553-580 (cf. p. 20, 90).

[HO04] Marlis Hochbruck et Alexander Ostermann, « Exponential Runge–Kutta methods for parabolic problems », *Applied Numerical Mathematics* 53.*2* (2004), Publisher : Elsevier, p. 323-339 (cf. p. 20, 60, 66).

[HO05] Marlis Hochbruck et Alexander Ostermann, « Explicit exponential Runge–Kutta methods for semilinear parabolic problems », *SIAM Journal on Numerical Analysis* 43.*3* (2005), Publisher : SIAM, p. 1069-1090 (cf. p. 20, 52, 60, 76, 82, 101, 103).

[HS21] Jingwei Hu et Ruiwen Shu, « On the uniform accuracy of implicit-explicit backward differentiation formulas (IMEX-BDF) for stiff hyperbolic relaxation systems and kinetic equations », *Mathematics of Computation* 90.*328* (2021), p. 641-670 (cf. p. 24, 52, 61, 73, 82).

[HR07] Willem Hundsdorfer et Steven J Ruuth, « IMEX extensions of linear multistep methods with general monotonicity and boundedness properties », *Journal of Computational Physics* 225.*2* (2007), Publisher : Elsevier, p. 2016-2042 (cf. p. 22, 50).

[Jin99] Shi Jin, « Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations », *SIAM Journal on Scientific Computing* 21.*2* (1999), Publisher : SIAM, p. 441-454 (cf. p. 27, 50).

[JPT98] Shi Jin, Lorenzo Pareschi et Giuseppe Toscani, « Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations », *SIAM Journal on Numerical Analysis* 35.*6* (1998), Publisher : SIAM, p. 2405-2439 (cf. p. 67, 68).

[JPT00] Shi Jin, Lorenzo Pareschi et Giuseppe Toscani, « Uniformly accurate diffusive relaxation schemes for multiscale transport equations », *SIAM Journal on Numerical Analysis* 38.*3* (2000), Publisher : SIAM, p. 913-936 (cf. p. 24, 67).

[JX95] Shi Jin et Zhouping Xin, « The relaxation schemes for systems of conservation laws in arbitrary space dimensions », *Communications on pure and applied mathematics* 48.*3* (1995), Publisher : Wiley Online Library, p. 235-276 (cf. p. 7, 53, 73).

[JM80] William B Jones et Arne Magnus, « Computation of poles of two-point Padé approximants and their limits », *Journal of Computational and Applied Mathematics* 6.*2* (1980), p. 105-119 (cf. p. 92).

[JNT83] William B Jones, Olav Njåstad et Wolfgang J Thron, « Two-point Padé expansions for a family of analytic functions », *Journal of Computational and Applied Mathematics* 9.*2* (1983), p. 105-123 (cf. p. 92).

[Kra26] Hendrik Anthony Kramers, « Wellenmechanik und halbzahlige Quantisierung », *Zeitschrift für Physik* 39.*10* (1926), Publisher : Springer, p. 828-840 (cf. p. 30).

[Law67] J Douglas Lawson, « Generalized Runge-Kutta processes for stable systems with large Lipschitz constants », *SIAM Journal on Numerical Analysis* 4.*3* (1967), p. 372-380 (cf. p. 20).

[LM08] Mohammed Lemou et Luc Mieussens, « A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit », *SIAM Journal on Scientific Computing* 31.*1* (2008), Publisher : SIAM, p. 334-368 (cf. p. 7, 23, 27, 53, 68).

[LM88] P. Lochak et C. Meunier, *Multiphase Averaging for Classical Systems : With Applications to Adiabatic Theorems*, Applied Mathematical Sciences, New York : Springer-Verlag, 1988, doi : 10.1007/978-1-4612-1044-3 (cf. p. 13, 26, 30).

[MZ09] Stefano Maset et Marino Zennaro, « Unconditional stability of explicit exponential Runge-Kutta methods for semi-linear ordinary differential equations », *Mathematics of computation* 78.*266* (2009), p. 957-967 (cf. p. 60).

[MS11] Omar Morandi et Ferdinand Schürrer, « Wigner model for quantum transport in graphene », *Journal of Physics A : Mathematical and Theoretical* 44.*26* (2011), p. 265301 (cf. p. 7).

[Mur06] James Murdock, *Normal forms and unfoldings for local dynamical systems*, Springer Science & Business Media, 2006 (cf. p. 13, 56, 78, 89).

[Nei84] A. I. Neishtadt, « The separation of motions in systems with rapidly rotating phase », *Journal of Applied Mathematics and Mechanics* 48.*2* (jan. 1984), p. 133-139, doi : 10.1016/0021-8928(84)90078-9 (cf. p. 30, 42).

[Per69] Lawrence M. Perko, « Higher Order Averaging and Related Methods for Perturbed Periodic and Quasi-Periodic Systems », *SIAM Journal on Applied Mathematics* 17.*4* (juil. 1969), p. 698-724, doi : 10.1137/0117065 (cf. p. 13, 26, 30).

[Rob14] Anthony John Roberts, *Model emergent dynamics in complex systems*, t. 20, Section : IV, SIAM, 2014 (cf. p. 78).

[Sak90] Kunimochi SAKAMOTO, « Invariant manifolds in singular perturbation problems for ordinary differential equations », *Proceedings of the Royal Society of Edinburgh Section A : Mathematics* 116.*1* (1990), Publisher : Royal Society of Edinburgh Scotland Foundation, p. 45-78 (cf. p. 51).

[SAAP00] Eva SÁNCHEZ, Ovide ARINO, Pierre AUGER et Rafael Bravo de la PARRA, « A singular perturbation in an age-structured population model », *SIAM Journal on Applied Mathematics* 60.*2* (2000), Publisher : SIAM, p. 408-436 (cf. p. 7, 50).

[SVM07] Jan A. SANDERS, Ferdinand VERHULST et James MURDOCK, *Averaging Methods in Nonlinear Dynamical Systems*, 2e éd., Applied Mathematical Sciences, New York : Springer-Verlag, 2007, DOI : 10.1007/978-0-387-48918-6 (cf. p. 13, 26, 27, 30, 37).

[SBD86] Andres SANTOS, J Javier BREY et James W DUFTY, « Divergence of the Chapman-Enskog expansion », *Physical review letters* 56.*15* (1986), p. 1571 (cf. p. 13).

[Spo00] Bruno SPORTISSE, « An analysis of operator splitting techniques in the stiff case », *Journal of computational physics* 161.*1* (2000), p. 140-168 (cf. p. 18, 20).

[TLH12] Arnel L TAMPOS, Jose Ernie C LOPE et Jan S HESTHAVEN, « Accurate reconstruction of discontinuous functions using the singular pade-chebyshev method », *IAENG International Journal of Applied Mathematics* 42.*ARTICLE* (2012), p. 242-249 (cf. p. 92).

[Vas63] Adelaida Borisovna VASIL'EVA, « Asymptotic behaviour of solutions to certain problems involving non-linear differential equations containing a small parameter multiplying the highest derivatives », *Russian Mathematical Surveys* 18.*3* (1963), Publisher : IOP Publishing, p. 13 (cf. p. 51).

[VS98] Jan G VERWER et Bruno SPORTISSE, « A note on operator splitting in a stiff linear case », *Modelling, Analysis and Simulation [MAS] R 9830* (1998) (cf. p. 21).

[Wen26] Gregor WENTZEL, « Eine verallgemeinerung der quantenbedingungen für die zwecke der wellenmechanik », *Zeitschrift für Physik* 38.*6* (1926), Publisher : Springer, p. 518-529 (cf. p. 30).

UNIVERSITÉ DE RENNES 1

Titre : Méthodes d'analyse asymptotique et d'approximation numérique – Problèmes d'évolution multi-échelles de type oscillatoire ou dissipatif

Mot clés : décomposition micro-macro, précision uniforme, relaxation rapide, moyennisation

Résumé : Les problèmes à relaxation rapide apparaissent dans de nombreux systèmes physiques ou biologiques, notamment dans le cadre de modèles cinétiques avec collisions. Leur comportement mélange une dynamique de relaxation de temps caractéristique ε et une partie lente d'interactions (généralement non-linéaire) ou de transport. Naturellement, on cherche à résoudre ce type de problème numériquement. Malgré le développement depuis les années 1980 de méthodes de résolution adaptées peu coûteuses (i.e. stables et essentiellement explicites), un problème demeure : la précision des méthodes est dégradée lorsque le pas de discrétisation est d'ordre ε. Dans ce manuscrit, on présente une méthode pour dépasser cette limite. L'approche mise en œuvre consiste à effectuer des développements asymptotiques par rapport au paramètre ε de sorte à pouvoir séparer le modèle asymptotique et son erreur ; on parle alors d'un problème micro-macro. Ce nouveau problème peut être résolu numériquement et on reconstruit la solution du problème d'origine avec une précision indépendante du paramètre ε. Nos développements asymptotiques font appel à des résultats récents de moyennisation, si bien qu'un chapitre de ce manuscrit est dédié à l'exposition de preuves originales de certains résultats de moyennisation connus. On discute en outre d'extensions possibles de nos résultats.

Title: Some methods for asymptotic analysis and numerical approximation – Multi-scale evolution problems, of oscillatory or relaxation behavior

Keywords: micro-macro decomposition, uniform accuracy, stiff relaxation, averaging

Abstract: Stiff relaxation problems appear in numerous physical and biological systems, most notably in kinetic models with collisions. The solutions of such problems present two dynamics which are intertwined: a relaxation of characteristic time ε, and a slow part of interactions or transport. Since the 1980s, efficient and stable methods have been developed to solve these problems numerically, however one issue remains: the accuracy of the method degrades when the time-step is of size ε. In this work, we present a method to overcome this limit. Our approach consists in performing asymptotic developments with relation to ε to construct an *non-stiff* asymptotic model and its error. We then consider this asymptotic behavior and the error separately – this is a micro-macro decomposition. This new problem may be solved and the solution of the original problem recovered, all with an accuracy independent of ε. As our asymptotic developments use results from averaging, a chapter of this work is dedicated to averaging results. Specifically, we present original proofs of known results, using recent frameworks which make algebraic reasonings straightforward. A brief discussion surrounding possible extensions of our results is conducted at the end of the manuscript.