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 de modèles dissipatifs multi-échelles

EDOs à variété centrale et modèles cinétiques

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

Rapporteurs avant soutenance :

Prénom NOM Fonction et établissement d'exercice
Prénom NOM Fonction et établissement d'exercice
Prénom NOM Fonction et établissement d'exercice

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 :	Prénom NOM	Fonction et établissement d'exercice
Co-dir. de thèse :	Prénom NOM	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

Contextualisation

Le développement de modèles mathématiques en sciences naturelles bénéficie toujours d'avancées mathématiques qui permettent de vérifier la 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 principale se penche sur les modèles double-échelle. Dans ces modèles, on distingue deux dynamiques : une d'échelle caractéristiques « rapide » ε et l'autre d'échelle 1. Dans ce manuscrit, on s'intéresse principalement 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 des systèmes dérivés de problèmes non-linéaires [JX95]. On les observe également en dynamiques des populations, e.g. dans [GHM94 ; AP96 ; Sán+00 ; 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 céleste, ou le problème de Fermi-Pasta-Ulam-Tsingou [For92] en théorie du chaos. Si E est un espace fonctionnel, 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 particulier, 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 en fait une brève description du comportement. À cet égard, on introduit deux exemples de systèmes « jouet » qui 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 la forme

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

On considère ce problème dans un Banach $(E, |\cdot|)$, 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 un problème jouet

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

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 nature différente, 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.

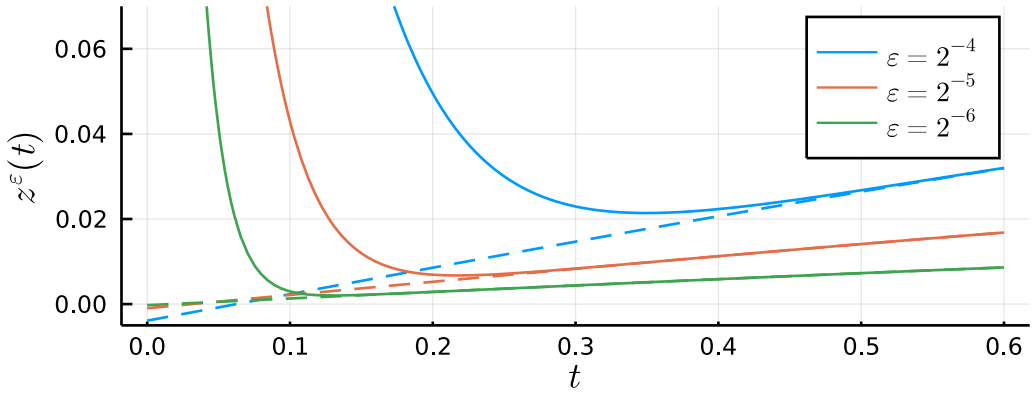

FIGURE 1 – Comportement de la solution (4) pour différentes valeurs de ε, avec chaque variété centrale associée en pointillés.

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])**.** *Si le champ f est de classe C^1, alors il existe un morphisme $x \mapsto z = \varepsilon h^\varepsilon(x)$, et un taux $\mu > 0$ tels que*

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

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

$$\big|x(t) - \varphi_t^\varepsilon\big(x_0^\varepsilon\big)\big| \lesssim \varepsilon e^{-\mu t/\varepsilon},$$

où φ_t^ε est le t-flot associé au champs de vecteur $x \mapsto a\big(x, \varepsilon h^\varepsilon(x)\big)$. On appelle l'ensemble des $\big(x, \varepsilon h^\varepsilon(x)\big)$ la variété centrale, qui attire la solution.

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

$$h^\varepsilon(x) = \frac{1}{1+\varepsilon^2}\big(\sin(x) - \varepsilon\cos(t)\big).$$

En général, 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, et $z \approx \varepsilon h^\varepsilon(x)$ et 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 $h^{[0]} = 0$ et itérer de manière explicite

$$\varepsilon \partial_x h^{[n]}(x) \cdot a(x, \varepsilon h^{[n]}(x)) = -h^{[n+1]}(x) + b(x, \varepsilon h^{[n]}(x)). \tag{5}$$

Le calcul de la donnée initiale de l'ombre x_0^ε n'est en revance 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

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 possédent une structure algébrique élégante basée sur des opérations sur les arbres.

que (5).

Remarque. *Le phénomène de donnée initiale de l'ombre n'est cependant pas visible sur 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. En effet, chaque itération vient demander un ordre de dérivation supplémentaire dans a et b, ce qui peut croître comme $n!$ par exemple avec $a = 1$ et $b(x, z) = 1/(1 + x)$. Même dans l'exemple jouet (3), cette méthode génère

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

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. $h^{[1]}(x) = \sin(x)$, $h^{[2]}(x) = \sin(x) + \varepsilon\cos(x)$ etc. Ainsi, le développement n'est convergent que pour $\varepsilon < 1$, alors que le résultat de variété centrale est valide pour tout $\varepsilon > 0$. Cette limitation apparaîtra couramment au cours de ce manuscrit sous le format plus contraignant

$$\varepsilon \leq \frac{\varepsilon_0}{n+1},$$

qu'on peut remarquer par exemple avec $b(x, z) = (1 + x)^{-1}$ dont la taille des dérivées évolue de manière factorielle avec l'ordre de dérivation.

Définitions et hypothèses

Problème uniformément bien posé

Propriété. *Le problème* (1) *est uniformément bien posé, i.e. il existe un temps d'existence $T > 0$ valide pour tout $\varepsilon < \varepsilon_0$.*

Ouvert borné $\mathcal{K}$

Régularité + « chaussette » autour de la solution

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 **REF**, la phase transitoire n'est pas calculée.

Plus récemment dans [CCS16], les auteurs capturent aussi la phase transitoire, mais la méthode est très difficile à s'approprier et n'est valide que dans la limite $\varepsilon \to 0$. 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.

Dans cette section, 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 reconnues dans 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

$$\begin{cases} \partial_t v_1 = v_2, & v_1(0) = 1, \quad (6a) \\ \partial_t v_2 = -\dfrac{1}{\varepsilon}\big(v_1 + v_2\big), & v_2(0) = 0. \quad (6b) \end{cases}$$

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

$$\begin{cases} \partial_t x = -x + z, & x(0) = 1, \quad (7a) \\ \partial_t z = -\dfrac{1}{\varepsilon} z - x + z, & z(0) = 1. \quad (7b) \end{cases}$$

Ce problème est linéaire et se diagonalise sans problème 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 $r_\varepsilon = \frac{1}{2\varepsilon}\big(1 - \sqrt{1 - 4\varepsilon}\big)$. On obtient directement une expression explicite pour $u = \begin{pmatrix} x \\ z \end{pmatrix}$, qui est

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

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 $-\sqrt{1-4\varepsilon}$.*

Introduction aux résolutions numériques

Dans les faits, on ne saura pas résoudre tous les systèmes de la forme (1) de manière exacte. Donc on va appliquer des méthodes d'*approximation numérique* pour calculer une solution approchée. À partir l'exemple (7), on va étudier la précision de ces méthodes. Voici la manière dont on procède :

Discrétisation de t de manière uniforme

Def : erreur d'une méthode numérique

Rq : on pourrait considérer une interpolation des données et définir une erreur dans le monde continu mais c'est compliqué et déjà si l'erreur ponctuelle est bien on est content

On suppose qu'on sait calculer $\exp(-tA)$ pour Strang ou expRK, ou qu'on sait facilement inverser $\mathrm{id} + \Delta t\, A$ pour IMEX-BDF.

Def : ce qu'on entend par un schéma numérique,

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

Footnote : en vérité, quitte à remplacer u_n par $U_n = (u_{n+s-1}, \dots, u_n)$, on peut se ramener à des méthodes à une seule étape

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 $\varepsilon \to 0$.

Attention : on observe le comportement de l'erreur et sa relation avec Δt mais aussi avec ε !

Disclaimer : je n'ai pas étudié les schémas en eux-mêmes, je les ai juste compilés et ai étudié leur comportement

Rq : Je n'ai pas étudié de méthodes complètement implicites parce qu'elles sont très coûteuses, pas très adapté pour l'extension aux EDP... Les méthodes purement explicites demandent $\Delta t < \varepsilon$, ce qui est beaucoup trop coûteux. On cherche justement à se débarasser de cette contrainte.

Splitting de Strang

Une approche courante est de séparer le problème (1) en deux parties, une raide et une 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, $\varphi^{(1)}_t$ et $\varphi^{(2)}_t$ 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 non en général, mais on peut *approcher* φ à partir des autres avec des compositions successives. C'est cette approche qu'on appelle *splitting*. Le plus couramment utilisé est le splitting de Strang, qui s'écrit

$$\varphi_t = \varphi^{(1)}_{t/2} \circ \varphi^{(2)}_t \circ \varphi^{(1)}_{t/2} + \mathcal{O}(t^3).$$

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. Ce schéma peut être obtenu par symétrie à partir du splitting de Lie $\Phi_t = \varphi^{(2)}_t \circ \varphi^{(1)}_t$, d'ordre 1. Le splitting est exact si et seulement si les champs A et f commutent, c'est-à-dire

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

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 (7), on a $[A, f] = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$.

Figure de convergence en fonction de Δt et ε, avec Lie en haut et Strang en bas. Séparer x et z ?

Dans cette figure, on observe que le comportement de la solution est le bon 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 de l'erreur. Cette augmentation, de tendance prédite en $1/\varepsilon$, entraîne une *réduction d'ordre*, c'est-à-dire qu'il n'y a pas de constante d'erreur C telle que

$$\sup_{\varepsilon} \mathrm{err} \leq C\Delta t^2.$$

Annexe B : Stormer-Verlett est un cas particulier de Strang mélangé à Euler explicite. En effet avec $\dot{q} = v$ et $\dot{v} = F(q)$,

$$
\begin{aligned}
v_{n+1/2} &= v_n + \frac{\Delta t}{2} F(q_n) \\
q_{n+1} &= q_n + \Delta t v_{n+1/2} \\
v_{n+1} &= v_{n+1/2} + \frac{\Delta t}{2} F(q_{n+1}).
\end{aligned}
$$

Ce qui revient à séparer le système en $\partial_t(q, v) = (0, F(q))$ et $\partial_t(q, v) = (v, 0)$.

Présentation rapide du splitting de Lie et de Strang, avec une résolution exacte des flots

Fig : convergence de Lie (Δt à gauche, ε à droite)

Fig : convergence de Strang (Δt à gauche, ε à droite)

Rq : réduction d'ordre notée dans un article de Sportisse en 2000 `https://www.sciencedirect.com/science/article/pii/S0021999100964957`

Méthode IMEX-BDF Justification de l'utilisation de cette méthode avec le côté "UA" et les IMEX-LM en cinétique

Formule de IMEX-BDF Euler avec justif

Fig : convergence de la méthode ordre 1

Fig : convergence de la méthode ordre 2

Méthodes exponentielles Runge-Kutta Formulation intégrale avec semi-groupe

Justification de l'utilisation de cette méthode avec la norme "relative"

Formule de expRK Euler

Rq : aussi considéré dans un article de Sportisse (`https://ir.cwi.nl/pub/4597`) pour compenser les failles d'erreur du splitting

Fig : convergence méthode ordre 1

Fig : convergence méthode ordre 2

Analyse des résultats

Toutes les méthodes se comportent de la même manière... Il y a réduction d'ordre

Def : convergence AP

Def : convergence UA

Rq : Souvent, la littérature parle de convergence UA lorsqu'il s'agit d'une convergence UA « faible », i.e. avec une donnée bien préparée $z(0) = \varepsilon h^\varepsilon(x(0)) + \mathcal{O}(\varepsilon^n)$.

Contribution personnelle

Suite aux résultats de précision uniforme (i.e. indépendante de ε) obtenus pour les problèmes hautement oscillants [Cha+15 ; Cha+20a ; CJL17], on étudie dans cette section quelques méthodes numériques à la pointe de l'état de l'art qui concernent

Résultat de convergence uniforme avec IMEX-BDF ou expRK

Explication de la méthode à détailler (sans prendre le crédit) :

— Lien avec moyennisation

— Formes normales, splitting quasi-exact

— Conservation exacte du défaut

Ouverture au cinétique

Annonce plan

Chapitre 1

La moyennisation en bref

Lors de ma troisième année de thèse, j'ai eu l'occasion d'un peu plus me pencher sur les méthodes de moyennisation, et notamment de rédiger un mini-article compilant certains résultats du sujet. Une partie est ici, l'autre sera présentée en Annexe quelle annexe ?

1.1 Présentation d'une méthode

1.1.1 L'équation homologique

Dérivation de l'équation homologique

Distinction entre averaging standard et stroboscopique

1.1.2 Définition d'une décomposition approchée

Relation de récurrence et résultat sur les bornes

Discussion autour de la méthode pour les résultats (boules imbriquées, estimations de Cauchy...)

1.2 Contexte d'un problème autonome

$$\partial_t y = \frac{1}{\varepsilon} G(y) + K(y) \tag{1.1}$$

On décompose

$$y^\varepsilon(t) = \Omega^\varepsilon_{t/\varepsilon} \circ \Psi^\varepsilon_t \circ \left(\Omega^\varepsilon_0\right)^{-1} \tag{1.2}$$

1.2.1 Un résultat géométrique

Ω est un flot, et commute avec Ψ.

1.2.2 Cas d'un opérateur linéaire

Averaging standard : crochets de Lie
Formes normales

1.3 Aspect numérique

Si on ne connaît pas le défaut, il est clair qu'on peut calculer numériquement

$$\partial_t v^{[n]} = F^{[n]}(v^{[n]}) \tag{1.3}$$

et alors

$$u^\varepsilon(t) = \Phi^{[n]}_{t/\varepsilon}\big(v^{[n]}(t)\big) + \mathcal{O}(\varepsilon^{n+!}) \tag{1.4}$$

Mais en fait si on montre ça, on peut montrer mieux (en supposant un peu de régularité). Section basée principalement sur l'article avec Gilles.

1.3.1 Définition d'un nouveau problème

Micro-macro ou pullback
Raideur "retardée"

1.3.2 Convergence uniforme

Résultat de convergence uniforme
Présentation de schémas intégraux et composition de schémas

Chapitre 2

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

Dans cet article, on ...

2.1 Introduction

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

$$\begin{cases} \dot{x}^\varepsilon = a(x^\varepsilon, z^\varepsilon), & x^\varepsilon(0) = x_0, \\ \dot{z}^\varepsilon = -\dfrac{1}{\varepsilon} A z^\varepsilon + b(x^\varepsilon, z^\varepsilon), & z^\varepsilon(0) = z_0, \end{cases} \tag{2.1}$$

with $\varepsilon \in (0, 1]$ a small parameter, A 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 $\frac{1}{\varepsilon}\Lambda$. 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

$$\dot{u}^\varepsilon = -\frac{1}{\varepsilon}\Lambda u^\varepsilon + f(u^\varepsilon), \quad u^\varepsilon(0) = u_0, \tag{2.2}$$

where $u = \begin{pmatrix} x \\ z \end{pmatrix}$, $\Lambda = \begin{pmatrix} 0 & 0 \\ 0 & A \end{pmatrix}$ 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 (2.2) recurrently appear in population dynamics (see [**castella.2015** GHM94 ; AP96 ; Sán+00]), where A 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]

$$\sup_{\varepsilon \in (0,1]} E_\varepsilon(\Delta t) \leq C \Delta t^s. \tag{2.3}$$

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 $\Delta t = \mathcal{O}(\varepsilon^{r/q})$.

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,

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.

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 (2.2) as the following composition of maps

$$u^\varepsilon(t) = \Omega^\varepsilon_{t/\varepsilon} \circ \Gamma^\varepsilon_t \circ \left(\Omega^\varepsilon_0\right)^{-1}(u_0) \tag{2.4}$$

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 $(\tau, u) \mapsto e^{-\tau\Lambda}u$ 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

$$u^\varepsilon(t) = \Omega^{[n]}_{t/\varepsilon}\left(v^{[n]}(t)\right) + w^{[n]}(t) \tag{2.5}$$

holds true, where $v^{[n]}(t) = \Gamma^{[n]}_t \circ \left(\Omega^{[n]}_0\right)^{-1}(u_0)$ 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 (2.3) 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 2.2.9.

The present work is related to the recent paper [CCS16], where asymptotic expansions of the solution of (2.1) are constructed for the special case where A 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 [Cha+20a] and later modified to become amenable for actual computations at any order [Cha+20b]. As a matter of fact, the technical arguments that sustain decomposition (2.4) are essentially adapted from [Cas+15] in a way that will be fully explained in Section 2.3.

The rest of the paper is organized as follows. In Section 2.2, 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 2.3, we give proofs of all the results from Section 2.2. In Section 2.4, 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 2.5, 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.

2.2 Uniform accuracy from a decomposition

We start by considering the solution u of

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

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 $\tau \in \mathbb{R}_+$,

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

In order for our approach to be rigorous, we start by introducing some definitions and assumptions in Subsection 2.2.1. We then present a way to approach these maps at any rank $n \in \mathbb{N}$ by $\Gamma^{[n]}$ and $\Omega^{[n]}$ in Subsection 2.2.2. This approximation is such that the error in (2.7) is of size $\mathcal{O}(\varepsilon^{n+1})$. In Subsection 2.2.3, 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 (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 2.3.

2.2.1 Definitions and assumptions

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

Assumption 2.2.1. *The matrix Λ is diagonal with nonnegative integer eigenvalues, and these values are nondecreasing when following the diagonal. In other words, $\Lambda = \mathrm{Diag}(\lambda_1, \dots, \lambda_d)$ with $(\lambda_i)_{1\le i\le d} \in \mathbb{N}^d$ and $\lambda_1 \le \cdots \le \lambda_d$.*

Thanks to this assumption, we write $u = \begin{pmatrix} x \\ z \end{pmatrix}$, with (x,z) such that $\Lambda u = \begin{pmatrix} 0 \\ Az \end{pmatrix}$ for some A positive definite. The dimension of z may be zero without making our results invalid.

Assumption 2.2.2. *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 (2.2) 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 (2.1). It also states that all solutions $(x^\varepsilon, z^\varepsilon)$ of (2.1) converge towards it exponentially quickly, i.e. there exists $\mu > 0$ independent of ε such that

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

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 2.2.3. *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 (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

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

which help quantify the distance to the solution u^ε. By Assumption 2.2.3, the solution of (2.2) is in $\mathcal{K}_0$ at all time.

Definition 2.2.4. *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* $\sigma \in [0, 6|||\Lambda|||]$,

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

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,

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

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

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

2.2.2 Constructing the micro-macro problem

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

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

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

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

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

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

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

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

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

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

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

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

Therefore, we write equation (2.13) as follows

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

and apply the Fourier operator (2.16) to get

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

Taking now $k = 0$ and using definition (2.15) we get the expression (2.14). This framework of exponential series comes naturally thanks to Assumption 2.2.1.

The homological equation (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

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

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

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

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

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

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

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

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

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

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 2.2.5. *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 2.2.4. 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 (2.20) and (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 (2.10) *and* (2.11)*, the following bounds are satisfied for all* $0 \leq \nu \leq p+1$*,*

$$(i) \quad \left\|\Omega^{[n]} - e^{-\tau\Lambda}\right\|_R \leq 4\varepsilon M, \qquad (ii) \quad \left\|\partial_\theta^\nu \left[\Omega^{[n]} - e^{-\tau\Lambda}\right]\right\|_R \leq 8\big(1 + |||\Lambda|||\big)^\nu \varepsilon M\, \nu!$$

$$(iii) \quad \|F^{[n]}\|_R \leq 2M \qquad (iv) \quad \|\eta_\tau^{[n]}(u)\|_{R,p} \leq 2M\big(1 + |||\Lambda|||\big)^p \left(2\mathcal{Q}_p \frac{\varepsilon}{\varepsilon_n}\right)^n$$

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 2.3.1, and this results remains valid with the choice $\Omega_0^{[n]} = \mathrm{id}$.

2.2.3 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

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

into Problem (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 (2.23) we obtain the following system

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

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 (2.22),

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

We get the micro-macro problem

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

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

Theorem 2.2.6. *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 2.2.4. For all $\varepsilon \leq \varepsilon_n$, Problem (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 $E^{[n]} = \partial_t w^{[n]} + \frac{1}{\varepsilon}\Lambda w^{[n]}$.

Remark 2.2.7. *The attentive reader may notice that, while we made the computation of $F^{[n]}$ easy with (2.21), the initial condition of the macro part, $v^{[n]}(0) = \big(\Omega^{[n]}_0\big)^{-1}(u_0)$, 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]

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

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

2. The above formula is a consequence of the behaviour of the error, $\Omega^{[n+1]} = \Omega^{[n]} + \mathcal{O}(\varepsilon^{n+1})$ (see [Cha+20a]), 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.

of 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 (2.25) *at each rank* n *and truncating the* $\mathcal{O}(\varepsilon^{n+2})$ *term. The initial condition of the micro part becomes*

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

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

Using a standard explicit scheme to solve Problem (2.24) cannot work due to the term $\frac{1}{\varepsilon}\Lambda w^{[n]}$. 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 (2.2) would yield

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

with $\varphi(-h\Lambda) = \frac{1}{h}\int_0^h e^{-s\Lambda}\mathrm{d}s$. Because Λ 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

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

This norm is interesting because after a short time $t \geq \varepsilon \log(1/\varepsilon)$, the z-component of the solution u^ε of (2.2) is of size ε, as evidenced by the center manifold theorem in (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 2.2.8. *Under the assumptions of Theorem 2.2.6 and denoting $T_n \leq T$ a final time such that Problem (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 (2.24) using an exponential Runge-Kutta scheme of order $q := \min(n, p) + 1$ yields a uniform error of order q, i.e.*

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

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 2.2.9. *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 Λ. 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.

2.3 Proofs of theorems from Section 2.2

2.3.1 Proof of Theorem 2.2.5 : 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 (2.20) and (2.21). From a straightforward induction using Assumptions 2.2.1 and 2.2.2, 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

$$\Phi_\theta^{[n]}(u) = \Xi^{[n]}(e^{i\theta};u) = \Omega_{-i\theta}^{[n]}(u) = \sum_{k\in\mathbb{N}} e^{ik\theta}\widehat{\Omega^{[n]}}_k(u). \tag{2.29}$$

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

$$\big(\partial_\theta - i\Lambda\big)\Phi_\theta^{[n+1]} = -i\varepsilon\left(f\circ\Phi_\theta^{[n]} - \partial_u\Phi_\theta^{[n]}\cdot F^{[n]}\right). \tag{2.30}$$

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

$$\delta_\theta^{[n]} = \frac{1}{\varepsilon}\big(\partial_\theta - i\Lambda\big)\Phi_\theta^{[n]} + if\circ\Phi_\theta^{[n]}. - i\partial_u\Phi_\theta^{[n]}\cdot F^{[n]} \tag{2.31}$$

Note that these relations both use the identity

$$\sum_{k\in\mathbb{N}} \xi^k \widehat{f\circ\Omega^{[n]}}_k = f\left(\sum_{k\in\mathbb{N}} \xi^k \widehat{\Omega^{[n]}}_k\right) \tag{2.32}$$

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 $\widetilde{\Phi}_\theta^{[n]} = e^{-i\theta\Lambda}\Phi_\theta^{[n]}$, it satisfies

$$\partial_\theta\widetilde{\Phi}_\theta^{[n+1]} = \varepsilon\left(g_\theta\circ\widetilde{\Phi}_\theta^{[n]} - \partial_u\widetilde{\Phi}_\theta^{[n]}\cdot G^{[n]}\right) \tag{2.33}$$

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

Property 2.3.1. *Assumptions 2.2.2 and 2.2.3 ensure the following properties, with R, M and p given in Definition 2.2.4 :*

(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 $\sigma \in [0,3]$,

$$\forall\ 0 \leq \nu \leq p+2, \quad \frac{\sigma^\nu}{\nu!}\left\|\partial_\theta^\nu g\right\|_{\mathbb{T},2R} \leq M, \tag{2.34}$$

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

$$\widetilde{\Phi}_\theta^{[n+1]} = \mathrm{id} + \varepsilon\big(T_\theta^{[n]} - \Pi(T^{[n]})\big) \quad \text{and} \quad \Phi_\theta^{[n+1]} = e^{i\theta\Lambda}\Phi_\theta^{[n+1]} \tag{2.35}$$

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 (2.33) gives the vector field

$$G^{[n]} = \Pi\big(g \circ \widetilde{\Phi}^{[n]}\big). \tag{2.36}$$

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

$$\|\varphi\|_{\mathbb{T},\rho} := \sup_{(\theta,u)\in\mathbb{T}\times\mathcal{K}_\rho} |\varphi_\theta(u)| \quad \text{and} \quad \|\varphi\|_{\mathbb{T},\rho,\nu} := \max_{0\leq\alpha\leq\nu} \|\varphi_\theta(u)\|_{\mathbb{T},\rho} \tag{2.37}$$

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

Theorem 2.3.2 (from [Cha+20a] and [**castella.2018.stroboscopic**])**.** *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 (2.35) and (2.36). 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 $\widetilde{\Phi}_\theta^{[n]} = e^{-i\theta\Lambda}\Phi_\theta^{[n]}$ and $\widetilde{\delta}^{[n]} = e^{-i\theta\Lambda}\delta_\theta^{[n]}$ correspond to the filtered equation (2.33), and $\mathcal{Q}_p$ is a p-dependent constant.

In order to prove Theorem 2.2.5, 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 $\Phi_\theta^{[0]} = e^{i\theta\Lambda}$ since Λ 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 $u \in \mathcal{K}_R$.

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

By maximum modulus principle,

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

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

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

and $\left\|\partial_\theta^\nu[\Phi^{[n]} - e^{i\theta\Lambda}]\right\|_{\mathbb{T},R} \leq \left(1+|||\Lambda|||\right)^\nu \|\partial_\theta^\nu\widetilde{\Phi}^{[n]}\|_{\mathbb{T},R,\nu}$. 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 $e^{-i\theta\Lambda}$ 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 (2.32) and (2.31). The maximum modulus principle can finally be applied to the couple ($\eta^{[n]}$, $\delta^{[n]}$) in order to obtain the last estimate of Theorem 2.2.5.

□

2.3.2 Proof of Theorem 2.2.6 : well-posedness of the micro-macro problem

This proof is in several parts : first we show that problem (2.24a) is well-posed, and use this result to show that the bound on $w^{[n]}$ is satisfied, thereby also proving that (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 2.2.5, 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 2.2.5 :

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

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 2.2.5, we deduce $\Omega_\tau^{[n]}\big(v^{[n]}(t)\big) \in \mathcal{K}_{5R/4}$.

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

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

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

$$\begin{aligned}\left|w^{[n]}(t)\right| &\leq \left|\int_0^t e^{\frac{s-t}{\varepsilon}\Lambda} L^{[n]}\big(s/\varepsilon, s, w^{[n]}(s)\big) w^{[n]}(s) \mathrm{d}s + \int_0^t e^{\frac{s-t}{\varepsilon}\Lambda} S^{[n]}(s/\varepsilon, s) \mathrm{d}s\right| \\ &\leq \int_0^t \frac{2M}{R}\left|w^{[n]}(s)\right| \mathrm{d}s + \left|\int_0^t e^{\frac{s-t}{\varepsilon}\Lambda} S^{[n]}(s/\varepsilon, s) \mathrm{d}s\right| \end{aligned} \tag{2.38}$$

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

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

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

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

where $\|\cdot\|_{\mathbb{T},R}$ is given by (2.37). Using Theorem 2.3.2, there exists a constant $M_n > 0$

such that for all $t \in [0, T_v]$,

$$\left| \int_0^t e^{\frac{s-t}{\varepsilon}\Lambda} S^{[n]}(s/\varepsilon, s) \mathrm{d}s \right| \leq M_n \left(\frac{\varepsilon}{\varepsilon_n} \right)^{n+1}. \tag{2.39}$$

Using Gronwall's lemma in (2.38) 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 (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 2.2.5.

□

2.3.3 Proof of Theorem 2.2.8 : 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

$$\left|w^{[n]}(t_i) - w_i\right|_\varepsilon \leq C\Delta t^q \Big(\|\partial_t^{q-1} E^{[n]}\|_\infty + \|\partial_t^q E^{[n]}\|_{L^1} \Big). \tag{2.40}$$

The bounds on $E^{[n]} = \partial_t w^{[n]} + \frac{1}{\varepsilon}\Lambda w^{[n]}$ and its derivatives w.r.t. ε can be found in Theorem 2.2.6, rendering the computation of bounds on the error of the micro part straightforward. From Theorem 2.2.5.*(i)*, $\Omega_\tau^{[n]}(u) = e^{-\tau\Lambda}u + \mathcal{O}(\varepsilon)$, therefore the error on $\Omega_{t/\varepsilon}^{[n]}(v^{[n]})$ is of the form

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

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 $\mathbb{R}^d$,

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

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

□

2.4 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 2.4.1), or as a scalar value function (relaxed conservation law, Subsection 2.4.2). 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 2.5.1.

2.4.1 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 $(t, x) \in [0, T] \times \mathbb{R}/2\pi\mathbb{Z}$

$$\begin{cases} \partial_t \rho + \partial_x j = 0, \\ \partial_t j + \dfrac{1}{\varepsilon} \partial_x \rho = -\dfrac{1}{\varepsilon} j, \end{cases} \tag{2.41}$$

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 2.4.1.

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

$$z_k(t) = \exp\left(-\lambda \frac{t}{\varepsilon}\right) z_k(0) - ik^3 \int_0^t e^{(s-t)\lambda/\varepsilon} \rho_k(s) \mathrm{d}s. \tag{2.42}$$

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

$$\begin{cases} \partial_t \rho_k = -\dfrac{k^2}{1 + \alpha\varepsilon k^2} \rho_k - ikz_k, \\ \partial_t z_k = -\dfrac{1}{\varepsilon} z_k + \dfrac{k^2}{1 + \alpha\varepsilon k^2} z_k - \dfrac{ik^3}{1 + \alpha\varepsilon k^2} \left(\alpha + \dfrac{1}{1 + \alpha\varepsilon k^2}\right) \rho_k. \end{cases} \tag{2.43}$$

As in (2.42), 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 $\Lambda = \mathrm{Diag}(0, 1)$ such that $\partial_t u_k = -\frac{1}{\varepsilon}\Lambda u_k + f(u_k)$ with

$$f(u) = \begin{pmatrix} -\dfrac{k^2}{1 + \alpha\varepsilon k^2} u_1 - iku_2 \\ \dfrac{k^2}{1 + \alpha\varepsilon k^2} u_2 - \dfrac{ik^3}{1 + \alpha\varepsilon k^2} \left(\alpha + \dfrac{1}{1 + \alpha\varepsilon k^2}\right) u_1 \end{pmatrix}. \tag{2.44}$$

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

$$F^{[0]}(\rho, z) = \hat{k}^2 \begin{pmatrix} -\rho \\ z \end{pmatrix} \qquad \text{with} \qquad \hat{k} = \frac{k}{\sqrt{1 + \alpha\varepsilon k^2}}. \tag{2.45}$$

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 (2.41), $u_k(t) = \Omega_{t/\varepsilon}^{[0]}\left(v^{[0]}(t)\right) + w^{[0]}(t)$, 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 $e^{-\frac{t}{\varepsilon}\Lambda} v^{[0]}(t)$ as can be seen in the upcoming expression of $\eta^{[0]}$ and using $\Omega_\tau^{[0]}(u) = e^{-\tau\Lambda} u$. This means that the interesting quantity is

$$e^{-\frac{t}{\varepsilon}\Lambda} v^{[0]}(t) = \begin{pmatrix} e^{-\hat{k}^2 t} & 0 \\ 0 & e^{-(1-\varepsilon\hat{k}^2)\frac{t}{\varepsilon}} \end{pmatrix} v^{[0]}(0). \tag{2.46}$$

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 $e^{-\frac{t}{\varepsilon}\Lambda} v^{[0]}(t)$ 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

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

The rescaled macro variable $e^{-\frac{t}{\varepsilon}\Lambda}v^{[0]}(t)$ is given by relation (2.46) 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 (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 $e^{-\frac{t}{\varepsilon}\Lambda}v^{[1]}(t)$. 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 (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

$$\Omega_0^{[1]}(\rho,z)=\begin{pmatrix} \rho+\dfrac{\varepsilon}{1+\alpha\varepsilon k^2}ikz \\ z-\dfrac{\varepsilon}{1+\alpha\varepsilon k^2}\hat{k}^2\left(\alpha+\dfrac{1}{1+\alpha\varepsilon k^2}\right)ik\rho \end{pmatrix},\tag{2.47}$$

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 $\Omega_\tau^{[1]}(u) = \Omega_0^{[1]}(e^{-\tau\Lambda}u)$ and $\eta_\tau^{[1]}(u) = \eta_0^{[1]}(e^{-\tau\Lambda}u)$ are satisfied, therefore the interesting variable is $e^{-\frac{t}{\varepsilon}\Lambda}v^{[1]}(t)$. 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

$$\partial_t w^{[1]} = -\frac{1}{\varepsilon}\Lambda w^{[1]} + f(w^{[1]}) - \eta_0^{[1]}\left(e^{-\frac{t}{\varepsilon}\Lambda}v^{[1]}(t)\right), \quad w^{[1]}(0) = u_k(0) - \Omega_0^{[1]}\left(v^{[1]}(0)\right) \tag{2.48}$$

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

$$\eta_\tau^{[1]}(\rho, z) = ik \cdot \varepsilon \hat{k}^2 \left(\alpha + \hat{I}\left(2 + \varepsilon \hat{k}^2(\alpha + \hat{I})\right)\right) \begin{pmatrix} e^{-\tau} z \\ \hat{k}^2(\alpha + \hat{I})\rho \end{pmatrix} \tag{2.49}$$

$$\text{and} \quad v^{[1]}(0) = \begin{pmatrix} \rho_k(0) - \varepsilon \hat{I} i k z_k(0) \\ z_k(0) + \varepsilon \hat{k}^2(\alpha + \hat{I}) i k \rho_k(0) \end{pmatrix}. \tag{2.50}$$

We approached the initial condition using Remark 2.2.7, 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 2.4.1. *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 (2.43) can be decomposed into*

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

where $\Omega_0^{[1]}$ is given by (2.47) and $w^{[1]}(t) = \mathcal{O}(\varepsilon^2)$. The macro component $v^{[1]}$ is given by

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

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* (2.50). *The micro component* $w^{[1]}$ *is the solution to*

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

with f *and* $\eta_0^{[1]}$ *given respectively by* (2.44) *and* (2.49). *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 (2.48) 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.

2.4.2 Relaxed conservation law

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

$$\begin{cases} \partial_t u + \partial_x \tilde{u} = 0, \\ \partial_t \tilde{u} + \partial_x u = \dfrac{1}{\varepsilon}(g(u) - \tilde{u}), \end{cases} \tag{2.51}$$

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

Remark 2.4.2. *Note that the assumption that* Λ *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* (2.4) *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 :

$$|g'(u)| < 1 \tag{2.52}$$

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

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

Using an upwind scheme after diagonalising problem (2.51) yields

$$\begin{cases} \partial_t U + D\tilde{U} - \dfrac{\Delta x}{2} LU = 0, \\ \partial_t \tilde{U} + DU - \dfrac{\Delta x}{2} L\tilde{U} = \dfrac{1}{\varepsilon}(g(U) - \tilde{U}). \end{cases} \tag{2.53}$$

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

$$\begin{cases} \partial_t U_1 = -D\big(U_2 + g(U_1)\big), \\ \partial_t U_2 = -\dfrac{1}{\varepsilon} U_2 + g'(U_1) DU_2 - T(U_1) \end{cases} \tag{2.54}$$

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

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

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

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

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

Remark 2.4.3. *Remember that for the telegraph equation, the macro variable $v^{[1]}(t)$ needed to be rescaled by $e^{-t\Lambda/\varepsilon}$. 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 $\phi^\varepsilon(u_1, u_2) = g''(u_1)\left(2g'(u_1) - \varepsilon(1 - g'(u_1)^2)\partial_x u_1\right) u_2$. 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 (2.52) *but the result would be different.*

Because D^2 is sparse, it is not too costly to compute $\left(I_N - \varepsilon D^2\right)^{-1}$, 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

$$\mu_k = 1 + \frac{\varepsilon}{\Delta x^2} \sin^2\left(2\pi \frac{k}{N}\right) \tag{2.55}$$

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.

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

$$\begin{aligned} \eta_0^{[1]}(U_1, U_2)_{U_1} = {} & D\big(g(U_1 + \varepsilon\widetilde{D}W) - g(U_1)\big) + (D - \widetilde{D})U_2 \\ & + \varepsilon\widetilde{D}\Big(g'(U_1)DW - \varepsilon T'(U_1)\widetilde{D}W - \varepsilon^2 g''(U_1)\big(T(U_1), \widetilde{D}W\big)\Big), \end{aligned} \tag{2.56a}$$

$$\begin{aligned} \eta_0^{[1]}(U_1, U_2)_{U_2} = {} & -\big(g'(U_1 + \varepsilon\widetilde{D}U_2) - g'(U_1)\big)DU_2 \\ & + T(U_1 + \varepsilon\widetilde{D}U_2) - T(U_1) - \varepsilon T'(U_1)\widetilde{D}U_2 \\ & + \varepsilon g'(U_1 + \varepsilon\widetilde{D}U_2)DT(U_1) - \varepsilon^2 g''(U_1)\big(\widetilde{D}U_2, T(U_1)\big) \\ & + \varepsilon T'(U_1)\left(Dg(U_1) - \varepsilon T(U_1)\right). \end{aligned} \tag{2.56b}$$

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

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

2.5 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 (2.1), thereby exhibiting the problem of order reduction.

In Subsection 2.5.1 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 2.5.2, we showcase the results of uniform convergence for the partial differential equations of Section 2.4. 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 2.5.3, we share our thoughts on some remaining avenues of research following this paper.

2.5.1 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 :

$$\begin{cases} \dot{x} = (1-z)\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} x \\ \dot{z} = -\dfrac{1}{\varepsilon} z + x_1^2 x_2^2 \end{cases} \tag{2.57}$$

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

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

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

$$\big(\partial_\tau + \Lambda\big)\Omega_\tau^{[n+1]} = \varepsilon\left(f \circ \Omega_\tau^{[n]} - \partial_u \Omega_\tau^{[n]} F^{[n]}\right) \tag{2.58}$$

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

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

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 $\delta^{[1]} = \frac{1}{\varepsilon}(\partial_\tau +$

$\Lambda)\Omega^{[1]} - T^{[1]}$. From a direct calculation this writes,

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

where for clarity we defined

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

$$Z_1 = -2x_1 x_2 ({x_1}^2 - {x_2}^2)\left(1 - \varepsilon (x_1 x_2)^2 + \varepsilon e^{-3\tau} z^{\ 3}\right) \quad \text{and} \quad Z_2 = -e^{-2\tau} (2u_1 u_2 u_3)^2.$$

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 (2.58). This yields [4]

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

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

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

Remark 2.5.1. *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].

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.

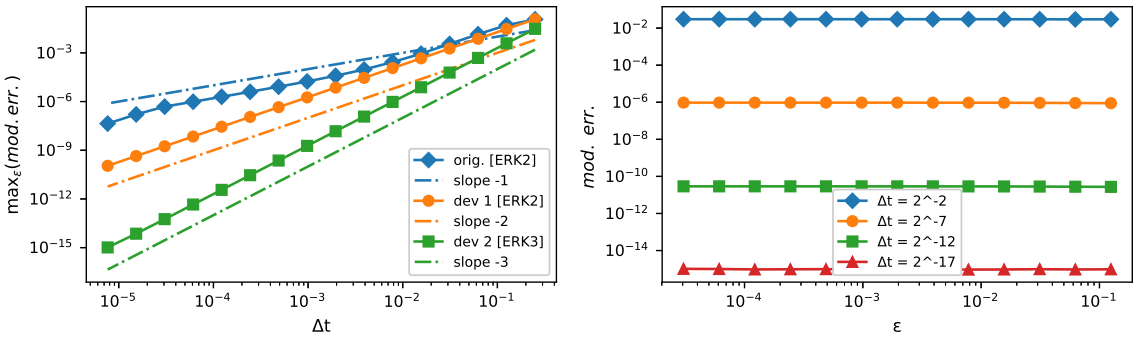

FIGURE 2.1 – Oscillating case : On the left, maximum error on ε (for $\varepsilon = 2^{-k}$ with k spanning $\{3, \dots, 15\}$) as a function of Δt when using exponential RK schemes (abbr. ERK) of different orders. On the right, the error as a function of ε when solving the micro-macro problem of order 2 using ERK3.

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

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

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 $u_i = \Omega^{[k]}_{t_i/\varepsilon}(v_i) + w_i$. Initial conditions $v^{[k]}(0)$ and $w^{[k]}(0)$ are computed using Remark 2.2.7.

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

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

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

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

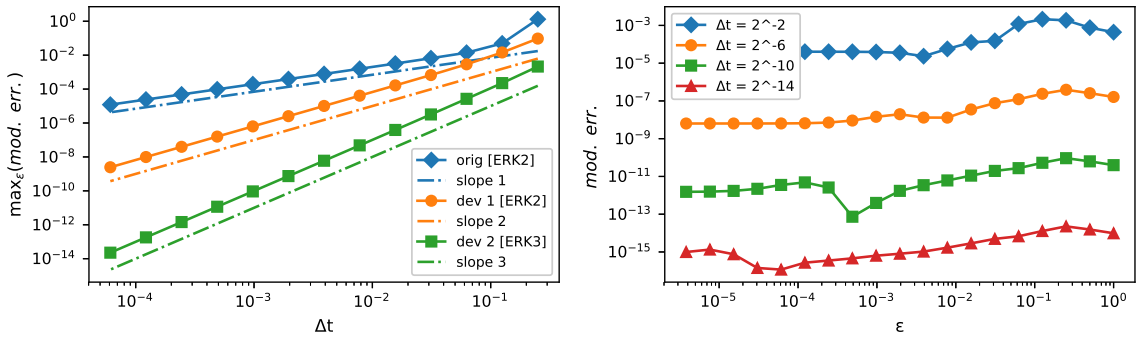

FIGURE 2.2 – PDE-inspired problem : On the left, maximum error on ε (for $\varepsilon = 2^{-k}$ with k spanning $\{1, \dots, 18\}$) as a function of Δt when using exponential RK schemes (abbr. ERK) of different orders. On the right, the error as a function of ε when solving the micro-macro problem of order 2 using ERK3.

for some smooth map $g : \mathbb{R}^d \mapsto \mathbb{R}^d$. This can be transformed in a system of the form (2.1) by setting $x = u$ and $z = g(u) - \tilde{u}$, yielding the problem

$$\begin{cases} \dot{x} = g(x) - z, & x(0) = u(0), \\ \dot{z} = -\dfrac{1}{\varepsilon} z + g'(x)\big(g(x) - z\big), & z(0) = g(u(0)) - \tilde{u}(0). \end{cases} \tag{2.59}$$

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 $\frac{1}{\varepsilon}(\partial_t + \Lambda)\Omega_\tau^{[2]}(x, z)$ 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 $e^{-\tau\Lambda}$-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 2.1 and 2.2 showcase the phenomenon of order reduction when solving the original problem (2.57) : 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

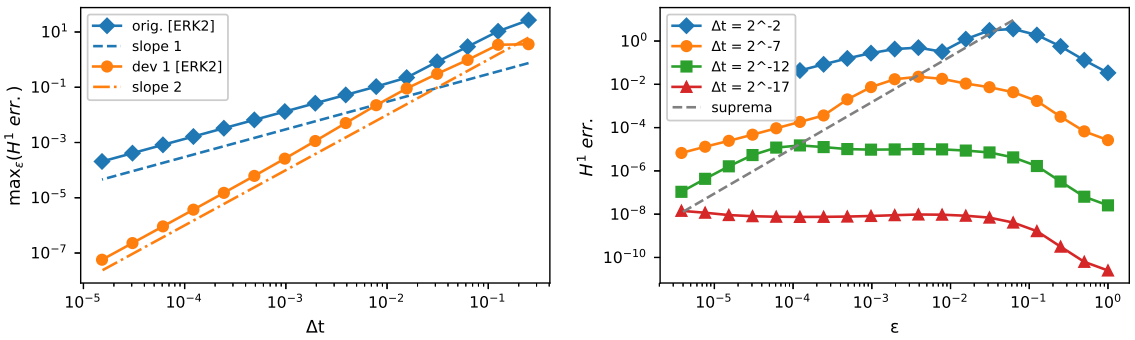

FIGURE 2.3 – Telegraph equation : Absolute H^1 error on the solution of (2.41) computed by an ERK3 scheme. Supremum on ε as a function of Δt (left) and evolution of this error as a function of ε for the 1st-order decomposition (right).

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 2.2.8 are confirmed. Indeed, using a scheme of order 2 (resp. 3) on the micro-macro problem of order 1 (resp. 2) generates a uniform error of the expected order of convergence, with no order reduction.

2.5.2 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 (2.43). The micro-macro decomposition of order 1 is summarized in Property 2.4.1, and its construction is detailed in Subsection 2.4.1. 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 2.3 when using a scheme of order 2. When solving

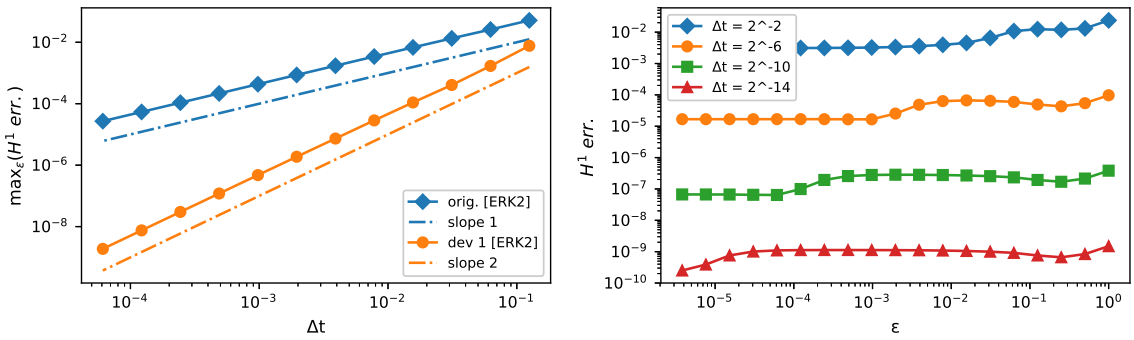

FIGURE 2.4 – Relaxed Burgers-type problem : Maximum modified H^1 error (for ε spanning 1 to 2^{-18} using an ERK3 scheme as a function of Δt (left), and H^1 error as a function of ε for the micro-macro problem of order 1 (right).

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 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 (2.1) 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 2.4.2.

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 2.4, 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.

2.5.3 Thoughts

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 (2.56) 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 (2.1), then it is close to the center manifold up to $\mathcal{O}(\varepsilon)$, and Problem (2.2) can be solved with uniform accuracy of order 2 but only when considering the absolute error $|\cdot|$, not the modified error $|\cdot|_\varepsilon$ from (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

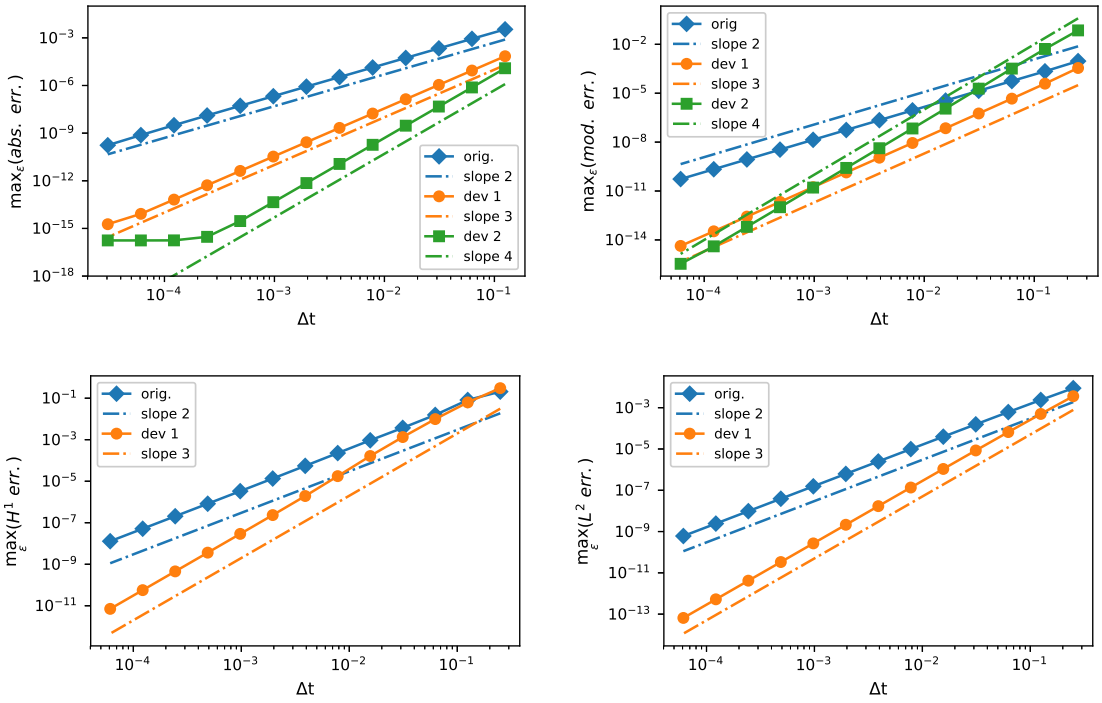

FIGURE 2.5 – In reading order, errors when solving the oscillating toy problem, the PDE-inspired problem, the telegraph equation and the relaxed conservation law. All systems start near equilibrium and are solved with exponential Runge-Kutta schemes of the observed order of convergence.

micro-macro decomposition of order 2 for the oscillating toy problem. These results can be seen in Figure 2.5 and will be studied in future works.

Chapitre 3

Discussion d'extension des résultats

> Closure. I keep hearing that word. It's the theater of the absurd. Everybody knows that on television they'll see the end of the story in the last 15 minutes of the thing. It's like a drug. To me, that's the beauty of 'Twin Peaks.' We throw in some curve balls. **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

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

3.1 Coût de calcul, erreurs d'arrondis, derivative-free, pullback

Coût de calcul avec les non-linéarités
Difficulté du calcul de la relaxation pour certains schémas
Gain d'ordre avec $z(0) = 0$ (figure avec err sup sur ε)
Donner une clé pour le gain d'ordre : $v_z(0) = \mathcal{O}(\varepsilon)$

3.2 Autour de l'équation de télégraphe

Annexe A

Un développement double-échelle

Résultats du stage et du début de thèse

Annexe B

Présentation de schémas numériques

Méthodes composées

Tableaux de Butcher

Splting

Stormer-Verlett est un cas particulier de Strang mélangé à Euler explicite. En effet avec $\dot{q} = v$ et $\dot{v} = F(q)$,

$$
\begin{aligned}
v_{n+1/2} &= v_n + \frac{\Delta t}{2} F(q_n) \\
q_{n+1} &= q_n + \Delta t v_{n+1/2} \\
v_{n+1} &= v_{n+1/2} + \frac{\Delta t}{2} F(q_{n+1}).
\end{aligned}
$$

Ce qui revient à séparer le système en $\partial_t(q, v) = (0, F(q))$ et $\partial_t(q, v) = (v, 0)$.

Annexe C

Autour de notions géométriques

C.1 Algèbre de Lie

C.2 Géométrie de la moyennisation stroboscopique

BIBLIOGRAPHIE

[ACM99] Georgios Akrivis, Michel Crouzeix et Charalambos Makridakis, « Implicit-explicit multistep methods for quasilinear parabolic equations », in : *Numerische Mathematik* 82.*4* (1999), Publisher : Springer, p. 521-541.

[ADP20] Giacomo Albi, Giacomo Dimarco et Lorenzo Pareschi, « Implicit-explicit multistep methods for hyperbolic systems with multiscale relaxation », in : *SIAM Journal on Scientific Computing* 42.*4* (2020), Publisher : SIAM, A2402-A2435.

[AP96] Pierre Auger et Jean-Christophe Poggiale, « Emergence of population growth models : fast migration and slow growth », in : *Journal of Theoretical Biology* 182.*2* (1996), Publisher : Elsevier, p. 99-108.

[ARW95] Uri M Ascher, Steven J Ruuth et Brian TR Wetton, « Implicit-explicit methods for time-dependent partial differential equations », in : *SIAM Journal on Numerical Analysis* 32.*3* (1995), Publisher : SIAM, p. 797-823.

[Bam03] Dario Bambusi, « Birkhoff normal form for some nonlinear PDEs », in : *Communications in mathematical physics* 234.*2* (2003), Publisher : Springer, p. 253-285.

[BD12] Weizhu Bao et Xuanchun Dong, « Analysis and comparison of numerical methods for the Klein–Gordon equation in the nonrelativistic limit regime », in : *Numerische Mathematik* 120.*2* (2012), p. 189-229.

[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 », in : *Physical review* 94.*3* (1954), p. 511.

[BPR17] Sebastiano Boscarino, Lorenzo Pareschi et Giovanni Russo, « A unified IMEX Runge–Kutta approach for hyperbolic systems with multiscale relaxation », in : *SIAM Journal on Numerical Analysis* 55.*4* (2017), Publisher : SIAM, p. 2085-2109.

[Car82] Jack Carr, *Applications of centre manifold theory*, t. 35, Applied Mathematical Sciences, Springer-Verlag New York, 1982.

[Cas+15] F. Castella et al., « Stroboscopic Averaging for the Nonlinear Schrödinger Equation », in : *Foundations of Computational Mathematics* 15.*2* (avr. 2015), p. 519-559, ISSN : 1615-3383, DOI : 10.1007/s10208-014-9235-7, URL : https://doi.org/10.1007/s10208-014-9235-7 (visité le 17/06/2021).

[CCS16] François Castella, Philippe Chartier et Julie Sauzeau, « A formal series approach to the center manifold theorem », in : *Foundations of Computational Mathematics* (2016), Publisher : Springer, p. 1-38.

[CCS18] Francois Castella, Philippe Chartier et Julie Sauzeau, « Analysis of a time-dependent problem of mixed migration and population dynamics », in : *arXiv preprint, arXiv :1512.01880* (2018).

[Cha+15] Philippe Chartier et al., « Uniformly accurate numerical schemes for highly oscillatory Klein–Gordon and nonlinear Schrödinger equations », in : *Numerische Mathematik* 129.*2* (2015), Publisher : Springer, p. 211-250.

[Cha+20a] Philippe Chartier et al., « A New Class of Uniformly Accurate Numerical Schemes for Highly Oscillatory Evolution Equations », in : *Foundations of Computational Mathematics* 20.*1* (fév. 2020), p. 1-33, ISSN : 1615-3375, 1615-3383, DOI : 10.1007/s10208-019-09413-3, URL : http://link.springer.com/10.1007/s10208-019-09413-3 (visité le 13/06/2020).

[Cha+20b] Philippe Chartier et al., « Derivative-free high-order uniformly accurate schemes for highly-oscillatory systems », in : *submitted preprint* (2020).

[CHV10] Philippe Chartier, Ernst Hairer et Gilles Vilmart, « Algebraic structures of B-series », in : *Foundations of Computational Mathematics* 10.*4* (2010), p. 407-427.

[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 », in : *Mathematical Models and*

Methods in Applied Sciences 27.*11* (2017), Publisher : World Scientific, p. 2031-2070.

[DM04] Stéphane Descombes et Marc Massot, « Operator splitting for nonlinear reaction-diffusion systems with an entropic structure : singular perturbation and order reduction », in : *Numerische Mathematik* 97.*4* (2004), p. 667-698.

[DP11] Giacomo Dimarco et Lorenzo Pareschi, « Exponential Runge–Kutta methods for stiff kinetic equations », in : *SIAM Journal on Numerical Analysis* 49.*5* (2011), Publisher : SIAM, p. 2057-2077.

[For92] Joseph Ford, « The Fermi-Pasta-Ulam problem : paradox turns discovery », in : *Physics Reports* 213.*5* (1992), p. 271-310.

[GHM94] Günther Greiner, JAP Heesterbeek et Johan AJ Metz, « A singular perturbation theorem for evolution equations and time-scale arguments for structured population models », in : *Canadian applied mathematics quarterly* 3.*4* (1994), Publisher : Applied mathematics institute of the University of Alberta, p. 435-459.

[GM03] Thierry Goudon et Antoine Mellet, « Homogenization and diffusion asymptotics of the linear Boltzmann equation », in : *ESAIM : Control, Optimisation and Calculus of Variations* 9 (2003), p. 371-398.

[GV11] Benoit Grébert et Carlos Villegas-Blas, « On the energy exchange between resonant modes in nonlinear Schrödinger equations », in : *Annales de l'Institut Henri Poincaré C, Analyse non linéaire* 28.*1* (jan. 2011), p. 127-134, issn : 0294-1449, doi : 10.1016/j.anihpc.2010.11.004, url : https://www.sciencedirect.com/science/article/pii/S0294144910000818 (visité le 17/06/2021).

[HH64] Michel Hénon et Carl Heiles, « The applicability of the third integral of motion : some numerical experiments », in : *The astronomical journal* 69 (1964), p. 73.

[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, isbn : 978-3-540-30663-4, doi :

10.1007/3-540-30666-8, URL : https://www.springer.com/gp/book/9783540306634 (visité le 17/06/2021).

[HO04] Marlis Hochbruck et Alexander Ostermann, « Exponential Runge–Kutta methods for parabolic problems », in : *Applied Numerical Mathematics* 53.*2* (2004), Publisher : Elsevier, p. 323-339.

[HO05] Marlis Hochbruck et Alexander Ostermann, « Explicit exponential Runge–Kutta methods for semilinear parabolic problems », in : *SIAM Journal on Numerical Analysis* 43.*3* (2005), Publisher : SIAM, p. 1069-1090.

[HR07] Willem Hundsdorfer et Steven J Ruuth, « IMEX extensions of linear multistep methods with general monotonicity and boundedness properties », in : *Journal of Computational Physics* 225.*2* (2007), Publisher : Elsevier, p. 2016-2042.

[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 », in : *Mathematics of Computation* 90.*328* (2021), p. 641-670.

[HW96] Ernst Hairer et Gerhard Wanner, *Solving ordinary differential equations II. Stiff and Differential-Algebraic Problems*, Springer Berlin Heidelberg, 1996.

[Jin99] Shi Jin, « Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations », in : *SIAM Journal on Scientific Computing* 21.*2* (1999), Publisher : SIAM, p. 441-454.

[JPT00] Shi Jin, Lorenzo Pareschi et Giuseppe Toscani, « Uniformly accurate diffusive relaxation schemes for multiscale transport equations », in : *SIAM Journal on Numerical Analysis* 38.*3* (2000), Publisher : SIAM, p. 913-936.

[JPT98] Shi Jin, Lorenzo Pareschi et Giuseppe Toscani, « Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations », in : *SIAM Journal on Numerical Analysis* 35.*6* (1998), Publisher : SIAM, p. 2405-2439.

[JX95] Shi Jin et Zhouping Xin, « The relaxation schemes for systems of conservation laws in arbitrary space dimensions », in : *Communications on pure and applied mathematics* 48.*3* (1995), Publisher : Wiley Online Library, p. 235-276.

[LM08] Mohammed Lemou et Luc Mieussens, « A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit », in : *SIAM Journal on Scientific Computing* 31.*1* (2008), Publisher : SIAM, p. 334-368.

[LM88] P. Lochak et C. Meunier, *Multiphase Averaging for Classical Systems : With Applications to Adiabatic Theorems*, Applied Mathematical Sciences, New York : Springer-Verlag, 1988, isbn : 978-0-387-96778-3, doi : 10.1007/978-1-4612-1044-3, url : https://www.springer.com/gp/book/9780387967783 (visité le 17/06/2021).

[MS11] Omar Morandi et Ferdinand Schürrer, « Wigner model for quantum transport in graphene », in : *Journal of Physics A : Mathematical and Theoretical* 44.*26* (2011), p. 265301.

[Mur06] James Murdock, *Normal forms and unfoldings for local dynamical systems*, Springer Science & Business Media, 2006.

[MZ09] Stefano Maset et Marino Zennaro, « Unconditional stability of explicit exponential Runge-Kutta methods for semi-linear ordinary differential equations », in : *Mathematics of computation* 78.*266* (2009), p. 957-967.

[Per69] Lawrence M. Perko, « Higher Order Averaging and Related Methods for Perturbed Periodic and Quasi-Periodic Systems », in : *SIAM Journal on Applied Mathematics* 17.*4* (juill. 1969), p. 698-724, issn : 0036-1399, 1095-712X, doi : 10.1137/0117065, url : http://epubs.siam.org/doi/10.1137/0117065 (visité le 17/06/2021).

[Rob14] Anthony John Roberts, *Model emergent dynamics in complex systems*, t. 20, Section : IV, SIAM, 2014.

[Sak90] Kunimochi Sakamoto, « Invariant manifolds in singular perturbation problems for ordinary differential equations », in : *Proceedings of the Royal Society of Edinburgh Section A : Mathematics* 116.*1* (1990), Publisher : Royal Society of Edinburgh Scotland Foundation, p. 45-78.

[Sán+00] Eva SÁNCHEZ et al., « A singular perturbation in an age-structured population model », in : *SIAM Journal on Applied Mathematics* 60.*2* (2000), Publisher : SIAM, p. 408-436.

[Spo00] Bruno SPORTISSE, « An analysis of operator splitting techniques in the stiff case », in : *Journal of computational physics* 161.*1* (2000), p. 140-168.

[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, ISBN : 978-0-387-48916-2, DOI : 10.1007/978-0-387-48918-6, URL : https://www.springer.com/gp/book/9780387489162 (visité le 17/06/2021).

[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 », in : *Russian Mathematical Surveys* 18.*3* (1963), Publisher : IOP Publishing, p. 13.

UNIVERSITÉ DE RENNES 1

Titre : titre (en français).....

Mot clés : de 3 à 6 mots clefs

Résumé : Eius populus ab incunabulis primis ad usque pueritiae tempus extremum, quod annis circumcluditur fere trecentis, circummurana pertulit bella, deinde aetatem ingressus adultam post multiplices bellorum aerumnas Alpes transcendit et fretum, in iuvenem erectus et virum ex omni plaga quam orbis ambit inmensus, reportavit laureas et triumphos, iamque vergens in senium et nomine solo aliquotiens vincens ad tranquilliora vitae discessit. Hoc inmaturo interitu ipse quoque sui pertaesus excessit e vita aetatis nono anno atque vicensimo cum quadriennio imperasset. natus apud Tuscos in Massa Veternensi, patre Constantio Constantini fratre imperatoris, matreque Galla. Thalassius vero ea tempestate praefectus praetorio praesens ipse quoque adrogantis ingenii, considerans incitationem eius ad multorum augeri discrimina, non maturitate vel consiliis mitigabat, ut aliquotiens celsae potestates iras principum molliverunt, sed adversando iurgandoque cum parum congrueret, eum ad rabiem potius evibrabat, Augustum actus eius exaggerando creberrime docens, idque, incertum qua mente, ne lateret adfectans. quibus mox Caesar acrius efferatus, velut contumaciae quoddam vexillum altius erigens, sine respectu salutis alienae vel suae ad vertenda opposita instar rapidi fluminis irrevocabili impetu ferebatur. Hae duae provinciae bello quondam piratico catervis mixtae praedonum.

Title: titre (en anglais).....

Keywords: de 3 à 6 mots clefs

Abstract: Eius populus ab incunabulis primis ad usque pueritiae tempus extremum, quod annis circumcluditur fere trecentis, circummurana pertulit bella, deinde aetatem ingressus adultam post multiplices bellorum aerumnas Alpes transcendit et fretum, in iuvenem erectus et virum ex omni plaga quam orbis ambit inmensus, reportavit laureas et triumphos, iamque vergens in senium et nomine solo aliquotiens vincens ad tranquilliora vitae discessit. Hoc inmaturo interitu ipse quoque sui pertaesus excessit e vita aetatis nono anno atque vicensimo cum quadriennio imperasset. natus apud Tuscos in Massa Veternensi, patre Constantio Constantini fratre imperatoris, matreque Galla. Thalassius vero ea tempestate praefectus praetorio praesens ipse quoque adrogantis ingenii, considerans incitationem eius ad multorum augeri discrimina, non maturitate vel consiliis mitigabat, ut aliquotiens celsae potestates iras principum molliverunt, sed adversando iurgandoque cum parum congrueret, eum ad rabiem potius evibrabat, Augustum actus eius exaggerando creberrime docens, idque, incertum qua mente, ne lateret adfectans. quibus mox Caesar acrius efferatus, velut contumaciae quoddam vexillum altius erigens, sine respectu salutis alienae vel suae ad vertenda opposita instar rapidi fluminis irrevocabili impetu ferebatur. Hae duae provinciae bello quondam piratico catervis mixtae praedonum.