ORSAY
n^0 d'ordre :

UNIVERSITÉ DE PARIS-SUD

CENTRE D'ORSAY

THÈSE

présentée
pour obtenir

Le GRADE de DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS XI ORSAY

Spécialité : Mathématiques

PAR

Guy MOEBS

Sujet : **APPLICATION DE METHODES SPECTRALES MULTI-NIVEAUX A DIFFERENTS PROBLEMES DE LA PHYSIQUE MATHEMATIQUE**

Soutenue le : 10 Juillet 1998 devant la Commission d'examen

Mme	C. BERNARDI	Présidente
M.	F. JAUBERTEAU	
M.	G. LABROSSE	
M.	J. LIANDRAT	Rapporteur
M	J. SHEN	Rapporteur
M.	R. TEMAM	
M.	J. ZYSS	Invité

A Sophie,
à mes parents

Remercier constitue, à mon avis, une tâche agréable mais néanmoins difficile. Agréable, parce qu'une fois accomplie une étape importante de la vie on a très envie de dire merci à toutes celles et à tous ceux qui ont collaboré d'une façon directe ou indirecte à son aboutissement. Difficile, parce qu'ils sont nombreux et qu'en imposant une liste de noms, on risque toujours d'oublier quelqu'un.

Je tiens à exprimer ma gratitude à Monsieur Roger Temam, mon directeur de thèse. Ses conseils et sa confiance m'ont permis de mener à bien ce travail au Laboratoire d'Analyse Numérique d'Orsay.
Je souhaiterais remercier Messieurs Jacques Liandrat et Jie Shen pour l'intérêt qu'ils accordent à mon travail en me faisant l'honneur de bien vouloir être les rapporteurs de cette thèse; de même que Mme Christine Bernardi, la présidente du jury, et Monsieur Joseph Jyss qui ont accepté de faire partie du jury.
Je remercie également Gérard Labrosse et François Jauberteau pour avoir accepté de juger ce travail ainsi que Monsieur David Gottlieb pour m'avoir tous trois consacré un peu de leur temps et de leur savoir.

C'est avec un réel plaisir que je remercie Thierry Dubois qui a guidé mes premier pas de jeune numéricien et cela dès le DEA. Je le remercie aussi pour sa patience, l'aide et le soutien qu'il m'a apportés ainsi que pour les innombrables questions auxquelles il a répondu.

Je tiens également à exprimer ma reconnaissance aux membres numériciens (Jacques Laminie, Frédéric Pascal, Hervé Le Meur, Jean-Pierre Croisille, Claude Jouron et Laurent Di Menza) et théoriciens (notamment Anne de Bouard, Olivier Goubet et Arnaud Debussche) du Laboratoire d'Analyse Numérique d'Orsay pour avoir eu la gentillesse de me consacrer du temps, sans oublier Danièle Le Meur pour sa disponibilité et son aide ainsi que notre ingénieur système Adrien Ramparison pour sa compétence.

Je n'oublie pas mes compagnons doctorants pour leur sympathie et leur aide au cours de ces presque quatre années : Caterina, Catherine, Farid, François, Bruno, Khalid, Michael pour ne citer qu'eux ainsi que mes ainés qui m'ont précédé : notamment Olivier, Pascal, Jean-Paul et Ezzedine.

Merci également au personnel de la bibliothèque mathématique d'Orsay pour sa compétence et sa gentillesse.

Enfin, je n'oublie pas de remercier ma famille, mes amis et Sophie pour leurs encouragements et leur soutien et surtout mes parents sans lesquels je ne serais pas là aujourd'hui. Qu'ils sachent à quel point je les en remercie.

Guy Moebs :

Applications of multilevel spectral methods to some problems of the mathematical physics.

Abstract :

This work deals with some aspects of multilevel spectral methods applied to several problems of the mathematical physics.

In a first part, we consider the weakly damped nonlinear Schrödinger equation as a propagation model for solitons through optical fibers. We use the pseudospectral Fourier method associated to the Split-Step Agrawal temporal scheme. A study of the Fourier spectrum allows us to underline the main part of the linear operator in the high frequencies equation. This method is performed for several splitting levels in frequencies; moreover it turns out that the method saves computing time when computing high frequencies.

Next, we study the treatment of nonperiodic boundary conditions for a two level spectral method, that is the Tau-Legendre method applied to a 2D electromagnetic problem : the resonant cavity. We consider several conservative discretizations in time, either explicit or implicit. The non-commutativity of the derivative operators with the orthogonal projectors onto the low and the high frequencies spaces, provides a coupling system for the projected equations, which is enforced by the global nature of the boundary conditions. An underrelaxed iterative algorithm of Gauss-Seidel type allows us to solve this system. Then we carry out a comparative study of the different schemes.

At least, we present results obtained with the stochastic Burgers equation as a turbulence model. For periodic boundary conditions we study the time evolution of the averaged quantities from the projections on the large and the small scales. For non-slip boundary conditions we compare results obtained by a spectral spatial discretization with those obtained by a finite differences method.

Key words :

spectral methods – multilevel methods – scales splitting – Schrödinger equation – Maxwell equation – Burgers' equation – orthogonal projections

n^0 d'impression : 2012
$3^{ème}$ trimestre 1998

Introduction

L'objet de ce travail est l'étude de certains aspects des méthodes spectrales multi-niveaux appliquées à différents problèmes de la physique mathématique.

Les méthodes spectrales sont des techniques d'approximation des solutions d'équations aux dérivées partielles. Les solutions discrètes sont construites comme le développement en série tronqué de polynômes, appelés fonctions de base. La précision de ces méthodes n'est limitée que par la régularité de la fonction à approcher : dans le cas de fonctions infiniment dérivables on parle alors de précision spectrale.
Développées depuis une trentaine d'années, leurs premières applications ont été des problèmes munis de conditions aux limites périodiques, compétitives avec d'autres méthodes de discrétisation grâce aux Transformées de Fourier Rapides, conçues en 1965 par Cooley et Turkey ([16]). Un premier ouvrage de référence est le livre de Gottlieb et Orszag paru en 1977, ([26]), rassemblant un grand nombre des premiers résultats.
Par la suite, des problèmes plus complexes avec des conditions aux limites de type Dirichlet ou Neumann ont été étudiés. Les solutions discrètes sont alors décomposées en base polynomiale de Chebyshev et de Legendre, cette dernière étant plus intéressante puisque les polynômes de Legendre sont orthogonaux pour la mesure de Lebesgue.
D'autres livres traitant des développements nouveaux des méthodes spectrales sont parus; citons notamment ([13, 25, 6]).

La première partie, préliminaire, de cette thèse est consacrée aux notions de base employées dans le cadre d'une discrétisation spatiale de type spectral. Nous introduisons les bases polynomiales les plus adaptées ainsi qu'un certain nombre de propriétés de ces polynômes (parité, dérivation, ...). Nous présentons aussi deux méthodes spectrales classiques : les méthodes Fourier-Galerkin et Tau-Chebyshev, dans le cadre de l'équation de Burgers 1D. Des tests numériques de convergence et de stabilité des discrétisations spatiale et temporelle clôturent cette première partie.

Les trois chapitres suivants décrivent plus directement notre travail.
L'équation de Schroedinger non linéaire faiblement amortie permet de modéliser la transmission de signaux binaires via une fibre optique. Nous considérons dans un premier temps la méthode usuelle pour ce type de propagation : la méthode Fourier-Collocation. Associée à un schéma en temps conservatif (schéma Split-Step Agrawal), elle fournit de bons résultats. Ensuite une étude de son spectre de Fourier nous permet de décomposer le signal (un train de solitons) en basses et hautes fréquences. De là, nous en déduisons le rôle prépondérant de l'opérateur linéaire dans l'équation régissant les hautes fréquences. La validation de cette nouvelle méthode appelée méthode multi-niveaux Split-Step Agrawal

effectuée pour différents niveaux de séparation entre basses et hautes fréquences met en évidence la possibilité de calculer ces dernières de manière plus économique.

Le troisième chapitre nous sert de cadre à l'étude du traitement de conditions aux limites non périodiques pour une méthode spectrale à deux niveaux : la méthode Tau-Legendre. Pour cela, Nous considérons le problème d'électromagnétisme appelé cavité résonnante. Il s'agit d'un domaine borné en dimension deux d'espace où les champs électrique et magnétique sont indépendants de la troisième variable d'espace. Le domaine est supposé entouré d'un conducteur parfait ce qui se traduit par des conditions aux limites de type Dirichlet homogène suivant la direction et l'inconnue considérées. La méthode Tau-Legendre usuelle est modifiée de telle sorte que l'on obtienne un problème semi-discrétisé en espace bien posé. Une discrétisation temporelle conservative pour préserver l'invariant nous délivre des contraintes de stabilité. Nous construisons alors les projections des différents champs sur leurs basses et hautes fréquences. La non commutativité des opérateurs de dérivation spatiale spectrale et de projection orthogonale constitue la difficulté principale. Une bonne imposition des conditions aux limites, qui nécessitent tous les modes de l'inconnue, entraine un couplage des équations obtenues par les projections. L'usage d'un algorithme itératif de type Gauss-Seidel par blocs permet de résoudre ce problème. Nous l'optimisons à l'aide d'une sous-relaxation (SOR). Cette décomposition de la solution en basses et hautes fréquences est réalisée indépendemment de toute contrainte; seul l'algorithme de point fixe restreint quelque peu le pas de temps. Une étude de la parallélisation des différents algorithmes achève cette partie.

Dans un quatrième et dernier chapitre, nous présentons des résultats plus anciens mais qui gardent un intérêt certain. Nous considérons l'équation de Burgers stochastique comme modèle de turbulence simplifié des équations de Navier-Stokes. Cette étude est réalisée dans le cas de conditions aux limites périodiques et de non glissement. Nous étudions alors d'une part l'évolution en temps des quantités moyennées issues des projections de l'équation sur les grandes et les petites structures et d'autre part nous comparons les résultats obtenus par discrétisation spatiale spectrale avec ceux obtenus à l'aide d'une discrétisation par différences finies.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Méthodes spectrales et applications.

Introduction.

Les méthodes spectrales peuvent être vues comme un développement extrême de la classe des schémas de discrétisation pour les équation différentielles appelée généralement méthode des résidus pondérés (MRP). Les éléments-clé de la MRP sont les fonctions d'approximation et les fonctions tests.
Les premières sont utilisées comme les fonctions de base pour le développement en série tronquée de la solution. Les dernières sont employées pour permettre à la série tronquée de satisfaire l'équation différentielle aussi fidèlement que possible.
Cela est atteint en minimisant le résidu, i.e. l'erreur dans l'équation différentielle engendrée en prenant le développement tronqué à la place de la solution exacte pour une norme convenable. Une formulation équivalente veut que le résidu satisfasse une condition d'orthogonalité appropriée avec chaque fonction test.

Le choix des fonctions de base est une des caractéristiques qui distingue les méthodes spectrales des méthodes d'éléments finis ou de différences finies. Ces fonctions, pour les méthodes spectrales, sont infiniment différentiables sur tout le domaine.
Dans le cas des éléments finis, le domaine est divisé en de petits éléments et une fonction de base est associée à chacun de ces éléments. De même les fonctions de base pour la méthode des différences finies sont locales.

Le choix des fonctions tests permet de distinguer entre-eux les trois schémas spectraux les plus employés, à savoir : la méthode de Galerkin, la méthode de Collocation et la méthode Tau.
Dans l'approche de Galerkin, les fonctions tests sont les mêmes que les fonctions de base. Par conséquent, ce sont des fonctions infiniment différentiables qui satisfont individuellement les conditions de frontière. L'équation différentielle est vérifiée en demandant que l'intégrale du résidu multiplié par chaque fonction test soit nulle.
Dans la méthode de Collocation, les fonctions tests sont des translations de la fonction de Dirac centrées aux points de "collocation". Cette approche impose la vérification de l'équation différentielle en chacun de ces points.
Les méthodes spectrales Tau sont similaires aux méthodes de Galerkin dans le sens où l'équation différentielle doit être vérifiée. Cependant, aucune des fonctions tests ne doit

satisfaire les conditions de frontière. On ajoute donc un ensemble supplémentaire d'équations pour ces conditions.
Avant de décrire plus précisément les méthodes de Galerkin et Tau, nous allons introduire les polynômes qui seront utilisés comme fonctions tests et fonctions de base et énoncer des propriétés utiles pour la mise en oeuvre de ces méthodes.

1.1 Le système de Fourier.

1.1.1 Développement continu de Fourier.

L'ensemble des fonctions

$$\Phi_k(x) = e^{ikx}, k \in \mathbb{Z}$$

constitue un système orthogonal dans l'intervalle $(0, 2\pi)$:

$$\int_0^{2\pi} \Phi_k(x)\overline{\Phi_l(x)}\,\mathrm{d}x = 2\pi\delta_{k,l} = \begin{cases} 2\pi & , \text{ si } k = l \\ 0 & , \text{ si } k \neq l \end{cases} \quad k, l \in \mathbb{Z}$$

Pour une fonction u à valeurs complexes définie sur l'intervalle $(0, 2\pi)$, nous définissons les coefficients de Fourier de u par

$$\hat{u}_k = \frac{1}{2\pi}\int_0^{2\pi} u(x)e^{-ikx}\,\mathrm{d}x, \quad k \in \mathbb{Z} \tag{1.1}$$

La relation (1.1) qui associe à u la suite de nombres complexes $\{\hat{u}_k\}_{k\in\mathbb{Z}}$ est appelée Transformée de Fourier de u.
De plus, si u est une fonction à valeurs réelles, alors les $\hat{u}_k$ sont des nombres complexes vérifiant

$$\hat{u}_{-k} = \overline{\hat{u}_k}$$

La série de Fourier de la fonction u est définie par

$$Su = \sum_{k\in\mathbb{Z}} \hat{u}_k\Phi_k$$

Pour N un entier strictement positif, posons

$$\mathbb{I}_N = \left[\left[1 - \frac{N}{2}, \frac{N}{2}\right]\right]$$

Soit $P_N u$ la troncature à l'ordre N de la série de Fourier de u.

$$P_N u = \sum_{k\in\mathbb{N}} \hat{u}_k\Phi_k \tag{1.2}$$

Nous avons choisi d'utiliser la formule (mathématiquement non conventionnelle) (1.2) parce qu'elle correspond directement à la façon dont les méthodes spectrales sont, en fait, mises en œuvre.
Nous rappelons maintenant des résultats sur la convergence des séries de Fourier.

Théorème *1.1*

a) Si u est continue, périodique et à variation bornée sur [0,2π] alors Su converge uniformément vers u, i.e.

$$\max_{x\in[0,2\pi]} |u(x) - P_N u(x)| \longrightarrow 0 \ \textit{quand} \ N \longrightarrow +\infty$$

b) Si u est à variation bornée sur [0,2π] alors $P_N u(x)$ converge localement vers $\frac{1}{2}\left(u(x^+) + u(x^-)\right)$ $\forall x \in [0, 2\pi]$ (ici $u(0^-) = u(2\pi^-)$)

c) Si u est continue et périodique, sa série de Fourier ne converge pas nécessairement en chaque point $x \in [0, 2\pi]$.

Nous considérons maintenant le problème de la vitesse de convergence de $P_N u$ vers u.
Les fonctions de $L^2(0, 2\pi)$ — i.e. les fonctions définies sur $(0, 2\pi)$ de carré intégrable au sens de Lebesgue — peuvent être caractérisées par leurs coefficients de Fourier $\{\hat{u}_k\}_{k\in\mathbb{Z}}$ grâce à l'identité de Parseval :

$$|\,u\,|_{L^2(0,2\pi)} = \left\{ 2\pi \sum_{k\in\mathbb{Z}} |\hat{u}_k|^2 \right\}^{\frac{1}{2}} \tag{1.3}$$

où $|\,.\,|_{L^2(0,2\pi)}$ désigne la norme de l'espace $L^2(0, 2\pi)$:

$$|\,u\,|_{L^2(0,2\pi)} = \left\{ \int_0^{2\pi} |u(x)|^2 \, \mathrm{d}x \right\}^{\frac{1}{2}}$$

Par (1.3) nous avons

$$|\,u - P_N u\,|_{L^2(0,2\pi)} = \left\{ 2\pi \sum_{k\in\mathbb{Z}\setminus\mathbb{I}_N} |\hat{u}_k|^2 \right\}^{\frac{1}{2}}$$

D'autre part, pour u suffisamment régulière

$$\max_{x\in[0,2\pi]} |u(x) - P_N u(x)| \leqslant \sum_{k\in\mathbb{Z}\setminus\mathbb{I}_N} |\hat{u}_k|^2$$

puisque $|\Phi_k(x)| \leqslant 1, \forall k \in \mathbb{Z}, \forall x \in [0, 2\pi]$.
Cela montre que la convergence de $P_N u$ vers u dépend de la rapidité de décroissance des coefficients de Fourier de u lorsque $|k|$ augmente.
Nous avons le résultat suivant :

Théorème *1.2*

Si u est m-fois continûment différentiable dans $[0, 2\pi]$, ($m \geqslant 1$) et si $u^{(j)}$ est périodique pour tout $j \leqslant$ m-2 alors

$$|\hat{u}_k| = \mathcal{O}(k^{-m}),\ k \in \mathbb{Z}^*$$

où $u^{(j)}$ désigne la $j^{ème}$ dérivée de u.

Corollaire *1.1*

Le $k^{ème}$ coefficient de Fourier d'une fonction infiniment différentiable et périodique décroît plus vite que toute puissance négative de k.

On parle alors de précision spectrale ou de convergence exponentielle.

1.1.2 Développement discret de Fourier.

Il arrive parfois que les méthodes numériques basées sur les séries de Fourier ne peuvent pas être mises en œuvre directement. L'utilisation de la Transformée de Fourier Discrète et les relations sur les séries discrètes de Fourier permettent de résoudre ces difficultés. On obtient des valeurs approchées des coefficients de Fourier de u à partir des valeurs de u en un nombre discret de points de l'espace physique appelés noeuds.
Pour un entier $N > 0$, on considère l'ensemble des points

$$x_j = \frac{2j\pi}{N},\ j \in \mathbb{J}_N = [\![0,\ N-1]\!] \tag{1.4}$$

Pour $u \in L^2(0, 2\pi)$, le polynôme d'interpolation $I_N u$ de u associé aux noeuds $\{x_j\}_{j\in\mathbb{J}_N}$ est défini par

$$I_N u(x) = \sum_{k\in\mathbb{I}_N} \tilde{u}_k e^{ikx} \tag{1.5}$$

Comme

$$I_N(x_j) = u(x_j),\ \forall j \in \mathbb{J}_N$$

alors on en déduit

$$\tilde{u}_k = \frac{1}{N}\sum_{j=0}^{N-1} u(x_j) e^{-ikx_j}, \qquad \forall k \in \mathbb{I}_N \tag{1.6}$$

grâce à la relation d'orthogonalité :

$$\frac{1}{N}\sum_{j=0}^{N-1} e^{ipx_j} = \delta_{p,N\mathbb{Z}} = \begin{cases} 1 & ,\ \text{si}\ \ p \in N\mathbb{Z} \\ 0 & ,\ \text{sinon} \end{cases} \tag{1.7}$$

Les équations (1.5) et (1.6) nous permettent de passer de l'espace physique $\{u(x_j)\}_{j\in\mathbb{J}}$ à l'espace transformé $\{\tilde{u}_k\}_{k\in\mathbb{Z}}$ et inversement. Les coefficients de Fourier discrets de u, $\tilde{u}_k$, peuvent être exprimés en fonction des coefficients de Fourier de u, $\hat{u}_k$:

$$\tilde{u}_k = \hat{u}_k + \sum_{m\in\mathbb{Z}^*} \hat{u}_{k+mN}, \qquad k \in \mathbb{I}_N \tag{1.8}$$

grâce à (1.6) et (1.7).
La formule (1.8) montre que le $k^{ème}$ mode de $I_N u$ dépend non seulement du $k^{ème}$ mode de u mais aussi de tous les modes qui aliasent Φ_k sur le maillage de l'espace physique. Ceci est dû au fait que

$$\Phi_{k+mN}(x_j) = \Phi_k(x_j), \ m \in \mathbb{Z}^*$$

Une formulation équivalente de (1.8) est

$$I_N u = P_N u \ + \ R_N u$$

avec

$$R_N u = \sum_{k \in \mathbb{I}_N} \left\{ \sum_{m \in k+N\mathbb{Z}^*} \hat{u}_m \right\} \Phi_k \tag{1.9}$$

L'erreur $R_N u$ entre le polynôme d'interpolation $I_N u$ et la série de Fourier tronquée $P_N u$ est appelée "erreur d'aliasing".
De plus, nous avons

$$| u - I_N u |^2_{L^2(0,2\pi)} = | u - P_N u |^2_{L^2(0,2\pi)} \ + \ | R_N u |^2_{L^2(0,2\pi)}$$

par orthogonalité de $u - P_N u$ avec $R_N u$.
On en déduit que l'erreur d'interpolation est toujours plus grande que l'erreur de troncature.

1.1.3 Différentiation.

Soit $H^1_{per}(0, 2\pi)$ l'espace des fonctions périodiques dans $L^2(0, 2\pi)$ ainsi que leur dérivée première u'.
Si l'on se place dans l'espace transformé, soit

$$Su = \sum_{k \in \mathbb{Z}} \hat{u}_k \Phi_k$$

la série de Fourier d'une fonction u, alors

$$Su' = \sum_{k \in \mathbb{Z}} ik\hat{u}_k \Phi_k$$

est la série de Fourier de la dérivée.
Par conséquent

$$(P_N u)' = P_N u'$$

i.e. la différentiation et la troncature commutent.
Pour approcher par méthode spectrale u' dans l'espace transformé, il suffit de dériver l'approximation de u. Les propriétés de convergence de l'approximation u' seront les mêmes que celles de u.

Si l'on se place dans l'espace physique, la différentiation est basée sur les valeurs de la fonction u aux noeuds $\{x_j\}$ du maillage (1.4). Elles sont utilisées pour l'évaluation des coefficients de Fourier discrets de u par (1.6). Ceux-ci sont ensuite multipliés par ik

et les coefficients obtenus sont ensuite transformés dans l'espace physique grâce à (1.5). Ainsi les valeurs $(\mathcal{D}_N u)_l$ de l'approximation de la dérivée aux noeuds sont données par

$$(\mathcal{D}_N u)_l = \sum_{k \in \mathbb{I}_N} d_k e^{ikx_j}, \qquad l \in \mathbb{J}_N$$

où

$$d_k = \frac{ik}{N} \sum_{j \in \mathbb{J}_N} u(x_j) e^{-ikx_j}$$

Cette procédure revient à calculer les valeurs nodales de la dérivée de la série discrète de Fourier, i.e.

$$\mathcal{D}_N u = (I_N u)'$$

Interpolation et différentiation ne commutent pas : $(I_N u)' \neq I_N(u')$

1.1.4 Produit de convolution.

Nous considérons le traitement du terme quadratique

$$w = uv \tag{1.10}$$

En effet un cas particulier de terme quadratique apparait dans les équations de Burgers et de Navier-Stokes, il s'agit du terme non linéaire $u\dfrac{\partial u}{\partial x}$ (i.e. $v = \dfrac{\partial u}{\partial x}$).
Dans le cas de développement en série infinie, nous avons la somme de convolution

$$\hat{w}_k = \sum_{\substack{m+n=k \\ m,n \in \mathbb{Z}}} \hat{u}_m \hat{v}_n \tag{1.11}$$

où

$$\begin{aligned} u(x) &= \sum_{m \in \mathbb{Z}} \hat{u}_m e^{imx} \\ v(x) &= \sum_{n \in \mathbb{Z}} \hat{u}_n e^{inx} \\ \hat{w}_k &= \frac{1}{2\pi} \int_0^{2\pi} w(x) e^{-ikx} \, \mathrm{d}x \end{aligned}$$

Lorsque u, v, w sont approchées respectivement par $P_N u$, $P_N v$, $P_N w$, dans (1.10), (1.11) devient

$$\hat{w}_k = \sum_{\substack{m+n=k \\ m,n \in \mathbb{I}_N}} \hat{u}_m \hat{v}_n, \ k \in \mathbb{I}_N \tag{1.12}$$

Le calcul direct de (1.12) nécessite $\mathcal{O}(N^2)$ opérations en dimension 1. Or pour l'évaluation d'un terme non linéaire, un algorithme de différences finies ne demande que $\mathcal{O}(N)$ opérations.
Cependant, l'utilisation de Transformées de Fourier Rapides (FFT) et le passage dans l'espace physique permettent de ramener ce coût à $\mathcal{O}(N \log_2(N))$ opérations, ce qui rend les méthodes spectrales utilisables pour les approximations d'équations aux dérivées partielles non linéaires avec un grand nombre de modes.

Pour évaluer (1.12) on utilise la Transformée de Fourier Discrète Inverse afin de transformer les suites $(\hat{u}_m)_{m\in\mathbb{I}_N}$ et $(\hat{v}_n)_{n\in\mathbb{I}_N}$ dans l'espace physique. On évalue alors dans l'espace physique une multiplication similaire à (1.10) puis on utilise la Transformée de Fourier Discrète pour déterminer $\hat{w}_k$.

Nous rappelons les notations

$$\mathbb{I}_N = \left[\!\!\left[1-\frac{N}{2},\,\frac{N}{2}\right]\!\!\right],\quad \mathbb{J}_N = [\![0,\ N-1]\!],\quad x_j = \frac{2j\pi}{N},\ j\in\mathbb{J}_N \quad \text{pour } N>0.$$

Posons

$$\left\{\begin{array}{lcl} U_j &=& \displaystyle\sum_{k\in\mathbb{I}_N} \hat{u}_k e^{ikx_j} \\ V_j &=& \displaystyle\sum_{k\in\mathbb{I}_N} \hat{v}_k e^{ikx_j} \end{array}\right. \quad ,\quad j\in\mathbb{J}_N$$

et

$$W_j = U_j\,V_j \quad j\in\mathbb{J}_N$$

On construit

$$\hat{W}_k = \frac{1}{N}\sum_{j\in\mathbb{J}_N} W_j e^{-ikx_j},\ k\in\mathbb{I}_N$$

La relation d'orthogonalité discrète (1.7) implique

$$\hat{W}_k = \sum_{m+n=k} \hat{u}_m\hat{v}_n + \sum_{m+n=k\pm N} \hat{u}_m\hat{v}_n$$

i.e.

$$\hat{W}_k = \hat{w}_k + \sum_{m+n=k\pm N} \hat{u}_m\hat{v}_n \tag{1.13}$$

Le second terme à droite de l'égalité (1.13) constitue l'erreur d'aliasing. La somme de convolution est évaluée au moyen de 3 FFT et N multiplications, soit un nombre d'opérations en $\mathcal{O}(N\log_2(N))$ en dimension 1.

Il existe différentes méthodes pour supprimer l'erreur d'aliasing. L'une des moins chères est la méthode du padding ou règle des *3/2*. Cette technique de déaliasing utilise une transformation discrète avec M points au lieu de N. En faisant le calcul précédent avec M points au lieu de N et en posant

$$\hat{u}^*_k = \left\{\begin{array}{ll} \hat{u}_k & ,\ \text{si}\ \ k\in\mathbb{I}_N \\ 0 & ,\ \text{si}\ \ k\in\mathbb{I}_M\setminus\mathbb{I}_N \end{array}\right.$$

on montre que l'erreur d'aliasing disparait si $M\geqslant\frac{3}{2}N$; i.e.

$$\hat{W}^*_k = \hat{w}_k \quad \forall k\in\mathbb{I}_N$$

Le nombre d'opérations pour cette transformation est $\mathcal{O}\left(M\log_2 M\right)$ soit environ 50% plus grand que dans la méthode sans déaliaser.

1.2 Polynômes orthogonaux.

1.2.1 Système de polynômes orthogonaux.

Nous considérons ici le problème du développement d'une fonction en terme d'un système de polynômes orthogonaux.
Nous notons $\mathcal{P}_N$ l'espace de tous les polynômes de degré inférieur ou égal à N. Soit $\{p_k\}_{k\in\mathbb{N}}$ un système de polynômes algébriques (avec degré de p_k égal à k), qui sont orthogonaux entre-eux sur l'intervalle $(-1,+1)$ pour la fonction poids w, i.e.

$$\int_{-1}^{1} p_k(x)p_l(x)w(x)\,\mathrm{d}x = 0 \qquad \text{pour} \;\; k \neq l$$

Le théorème de Weierstrass implique qu'un tel système est complet dans l'espace $L^2_w(-1,1)$, où $L^2_w(-1,1)$ représente l'espace des fonctions v telles que la norme

$$|\, v\,|_{L^2_w(-1,1)} = \left\{ \int_{-1}^{1} |v(x)|^2 w(x)\,\mathrm{d}x \right\}^{\frac{1}{2}} \tag{1.14}$$

soit finie. Le produit scalaire associé est

$$(u,v)_w = \int_{-1}^{1} u(x)v(x)w(x)\,\mathrm{d}x \tag{1.15}$$

Le développement en série d'une fonction $u \in L^2_w(-1,1)$ dans le système $\{p_k\}_{k\in\mathbb{N}}$ est

$$Su = \sum_{k\in\mathbb{N}} \hat{u}_k p_k$$

où les coefficients $\hat{u}_k$ du développement sont définis par

$$\hat{u}_k = \frac{1}{|\, p_k\,|^2_{L^2_w(-1,1)}} \int_{-1}^{1} u(x)p_k(x)w(x)\,\mathrm{d}x \tag{1.16}$$

Pour un entier $N > 0$, la série tronquée à l'ordre N de u est le polynôme

$$P_N u = \sum_{k=0}^{N} \hat{u}_k p_k \tag{1.17}$$

Par (1.14), $P_N u$ est la projection orthogonale de u dans $\mathcal{P}_N$ pour le produit scalaire (1.15); i.e.

$$(P_N u, v)_w = (u,v)_w \,,\; \forall v \in \mathcal{P}_N$$

Le système $\{p_k\}_{k\in\mathbb{N}}$ étant complet, nous avons

$$|\, u - P_N u\,|_{L^2_w(-1,1)} \longrightarrow 0 \;\; quand \;\; N \longrightarrow +\infty$$

Nous avons donc l'assurance de la convergence de $P_N u$ vers u (puisque l'erreur de troncature $u - P_N u$ tend vers 0) mais nous n'avons aucune indication sur la rapidité de la convergence, i.e. aucun renseignement sur la décroissance des coefficients $\hat{u}_k$ même lorsque u est de classe $\mathbb{C}^\infty$.

1.2.2 Problèmes de Sturm-Liouville.

L'importance des problèmes de Sturm-Liouville pour les méthodes spectrales réside dans le fait que l'approximation spectrale de la solution d'un problème différentiel est généralement regardée comme un développement fini en fonctions propres d'un problème de Sturm-Liouville adéquat.
Nous rappelons qu'un problème de Sturm-Liouville est un problème de valeurs propres de la forme

$$\begin{cases} -(pu')' + qu = \lambda wu \text{ dans l'intervalle } (-1,1) \\ \text{u munie de conditions aux limites convenables} \end{cases} \tag{1.18}$$

Les coefficients p, q, w sont trois fonctions à valeurs réelles données telles que : p est continûment différentiable, strictement positive dans $(-1,1)$ et continue en $x = \pm 1$.
q est continue, non négative et bornée dans $(-1,1)$. La fonction de poids w est continue, non négative et intégrable sur $(-1,1)$.
Seuls les développements en termes de fonctions propres de problèmes de Sturm- Liouville singuliers (i.e. lorsque p s'annule au bord) permettent d'obtenir la précision spectrale dans le cas de fonctions de classe $\mathbb{C}^\infty$ ([26], [13]).

1.2.3 Polynômes de Chebyshev.

1.2.3.1 Formules de base.

Les polynômes de Chebyshev de première espèce $\{T_k\}_{k\in\mathbb{N}}$ sont les fonctions propres du problème de Sturm-Liouville singulier :

$$\left(\sqrt{1-x^2}\,T_k'(x)\right)' + \frac{k^2}{\sqrt{1-x^2}}T_k(x) = 0$$

qui correspond à (1.18) avec

$$p(x) = (1-x^2)^{\frac{1}{2}}\ ,\ q(x) = 0 \text{ et } w(x) = (1-x^2)^{-\frac{1}{2}}.$$

Pour tout k, $T_k(x)$ a la parité de k.
Si T_k est tel que $T_k(1) = 1$ alors

$$T_k(x) = \cos\left(k \arccos(x)\right),\ |x| \leqslant 1,\ k \geqslant 0 \tag{1.19}$$

Ils vérifient la relation de récurrence

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x),\ |x| \leqslant 1,\ k \geqslant 0$$

$$T_0(x) = 1,\ T_1(x) = x,\ T_2(x) = 2x^2 - 1$$

Nous allons énoncer quelques propriétés des polynômes de Chebyshev dont nous aurons besoin ultérieurement

$$\begin{aligned} T_k(-x) &= (-1)^k T_k(x), \qquad |x| \leqslant 1, \qquad k \geqslant 0 \\ T_k(\pm 1) &= (\pm 1)^k \\ \left(\frac{dT_k}{dx}\right)(\pm 1) &= (\pm 1)^{k+1} k^2 \\ 2T_k(x) &= \frac{1}{k+1}\frac{dT_{k+1}(x)}{dx} - \frac{1}{k-1}\frac{dT_{k-1}(x)}{dx}, \qquad k \geqslant 2 \end{aligned} \tag{1.20}$$

Pour $m \geqslant 2$, on a

$$\frac{dT_m(x)}{dx} = 2(m+1) \sum_{\substack{k=0 \\ m-k \ pair}}^{m} \frac{1}{c_k} T_k(x)$$

Les constantes $(c_k)_k$ étant définies par

$$\begin{cases} c_0 &= 2 \\ c_k &= 1, \ k \geqslant 1 \end{cases}$$

1.2.3.2 Développement de Chebyshev continu.

Les polynômes de Chebyshev $\{T_k\}_k$ sont orthogonaux sur l'intervalle $[-1,1]$ pour le produit scalaire à poids

$$(f,g)_w = \int_{-1}^{1} f(x)g(x)w(x)\,\mathrm{d}x \qquad \text{avec } w(x) = (1-x^2)^{-\frac{1}{2}}$$

pour $f, g \in L^2_w(-1,1)$ et ils vérifient la relation d'orthogonalité

$$(T_m, T_n)_w = \frac{\pi}{2} c_m \delta_{m,n}$$

Le développement en série de Chebyshev d'une fonction $f \in L^2_w(-1,1)$ est

$$g(x) = \sum_{k \in \mathbb{N}} \hat{g}_k T_k(x), \quad \text{pour } |x| \leqslant 1$$

avec

$$\hat{g}_k = \frac{2}{\pi c_k} \int_{-1}^{1} f(x) T_k(x) w(x)\,\mathrm{d}x$$

Puisque $T_n(\cos\theta) = \cos(n\theta)$ alors $G(\theta) = g(\cos\theta)$ est la série de Fourier cosinus de $F(\theta) = f(\cos\theta)$ alors nous obtenons pour le spectre de Chebychev des résultats analogues à ceux du spectre de Fourier.

Théorème *1.3*

a) *Si f est continue par morceaux et à variation totale bornée pour $-1 \leqslant x \leqslant 1$ alors $g(x) = \frac{1}{2}(\ f(x^+) + f(x^-)\ )$ pour $-1 < x < 1$ et $g(1) = f(1^-)$, $g(-1) = f(-1^+)$*

b) *Si la dérivée $p^{ème}$, $f^{(p)}(x)$, est continue pour tout x, $|x| \leqslant 1$, pour $p = 0, 1, ...,$ $(n-1)$ et $f^{(n)}(x)$ est intégrable, alors $|\hat{a}_k| \leqslant \frac{1}{k^n}$ quand $n \longrightarrow +\infty$*

c) *Si f est de classe $\mathbb{C}^\infty(-1,1)$ alors $|\hat{a}_k|$ décroît plus vite que toute puissance négative de k.*

On note $P_N f$ la série de Chebyshev tronquée à l'ordre N de f

$$P_N f = \sum_{k=0}^{N} \hat{g}_k T_k$$

1.2.3.3 Développement de Chebyshev discret.

Comme le montre la formule (1.16), le calcul des coefficients $\{\hat{u}_k\}_k$ nécessite la connaissance des valeurs de u en chacun des points de l'espace physique. En fait, on approche ces coefficients $\{\hat{u}_k\}_k$ par une formule de quadrature. Dans le cas des polynômes de Chebyshev nous utilisons la formule de quadrature de Gauss-Lobatto.

Formule d'intégration de Gauss-Lobatto :
Pour un entier $N > 0$ donné, on considère l'ensemble des points

$$x_j = \cos\left(\frac{j\pi}{N}\right) \qquad j \in [\![0,\, N]\!] \tag{1.21}$$

appelés noeuds de Gauss-Lobatto. On définit de même les poids correspondants $\{w_j\}_j$, $j \in [\![0,\, N]\!]$ par :

$$w_j = \frac{\pi}{\overline{c_j} N}$$

avec

$$\begin{cases} \overline{c_0} &= \overline{c_N} = 2 \\ \overline{c_j} &= 1, \quad j \in [\![1,\, N-1]\!] \end{cases}$$

Nous avons le résultat suivant :

$$\sum_{j=0}^{N} p(x_j) w_j = \int_{-1}^{1} p(x)\, w(x)\, \mathrm{d}x\,, \ \forall p \in \mathcal{P}_{2N-1} \tag{1.22}$$

Soit u une fonction régulière définie sur $[-1, 1]$. Le polynôme d'interpolation $I_N u$ de u associé aux noeuds $\{x_j\}_j$ est un élément de $\mathcal{P}_N$ qui satisfait

$$I_N u(x_j) = u(x_j) \quad j \in [\![0,\, N]\!] \tag{1.23}$$

$I_N u$ est unique puisque les points x_j sont distincts. Comme il est de degré N, il peut s'écrire sous la forme

$$I_N u = \sum_{k=0}^{N} \tilde{u}_k\, T_k$$

et de manière triviale

$$u(x_j) = \sum_{k=0}^{N} \tilde{u}_k\, T_k(x_j) \tag{1.24}$$

Les $\{\tilde{u_j}\}_j$ sont appelés coefficients polynomiaux discrets de u. La relation d'orthogonalité discrète des polynômes de Chebyshev

$$\sum_{j=0}^{N} \frac{1}{\overline{c_j}} T_n(x_j) T_m(x_j) = \frac{N}{2}\, c_n\, \delta_{m,n}\,, \quad m,\, n \in [\![0,\, N]\!]$$

nous donne la relation inverse de (1.24)

$$\tilde{u}_k = \frac{2}{c_k N} \sum_{j=0}^{N} \frac{1}{\overline{c_j}}\; u(x_j)\, T_k(x_j) \tag{1.25}$$

Les équations (1.24) et (1.25) permettent de passer de l'espace physique $\{u(x_j)\}_j$ à l'espace transformé $\{\tilde{u}_k\}_k$ et inversement. Une telle transformation pour les polynômes orthogonaux est l'analogue des relations (1.5) et (1.6). Nous les appèlerons Transformations de Chebyshev Discrètes. Considérons le produit scalaire discret défini pour u, v continues sur $[-1,1]$:

$$(u,v)_N = \sum_{j=0}^{N} u(x_j)\, v(x_j)\, w_j$$

La formule d'intégration (1.22) implique

$$(u,v)_N = (u,v)_w \ \ , \text{ si } uv \in \mathcal{P}_{2N-1} \tag{1.26}$$

Pour toute fonction v, (1.23) nous donne

$$(I_N u, v)_N = (u,v)_N \tag{1.27}$$

L'orthogonalité des polynômes de Chebyshev $(T_k)_k$ et la relation (1.26) impliquent

$$(T_m, T_k)_N = \frac{\pi}{2}\, c_k\, \delta_{k,m} \ \ , \ \ k, m \in [\![0,\, N]\!] \tag{1.28}$$

De (1.27) et (1.28) nous obtenons

$$(u, T_k)_N = (I_N u, T_k)_N = \sum_{m=0}^{N} \tilde{u}_m\, (T_m, T_k)_N = \frac{\pi}{2} c_k \tilde{u}_k$$

que l'on reécrit

$$\tilde{u}_k = \frac{2}{\pi c_k} (u, T_k)_N \ \ , \quad k \in [\![0,\, N]\!] \tag{1.29}$$

De là

$$\begin{aligned} \tilde{u}_k &= \frac{2}{\pi c_k} \sum_{m=0}^{\infty} \hat{u}_m\, (T_m, T_k)_N \\ &= \hat{u}_k + \frac{2}{\pi c_k} \sum_{m>N} \hat{u}_m\, (T_m, T_k)_N \end{aligned}$$

qui est une conséquence de (1.28) et (1.29). En utilisant (1.19) et (1.7) avec $2N$ au lieu de N, nous obtenons pour $k \in [\![0,\, N]\!]$

$$(T_k, T_m)_N = \begin{cases} (T_k, T_k)_N & , \text{ si } m = 2pN \pm k,\ p \in \mathbb{N} \\ 0 & , \text{ sinon} \end{cases}$$

De là,

$$\tilde{u}_k = \hat{u}_k + \sum_{\substack{m=2pN\pm k \\ m>N}} \hat{u}_m \tag{1.30}$$

De même que dans le cas Fourier, le $k^{ème}$ mode de Chebyshev du polynôme d'interpolation dépend de tous les modes de Chebyshev qui aliasent T_k aux noeuds $\{x_j\}_j$.
On peut écrire de manière équivalente

$$I_N u = P_N u + R_N u \tag{1.31}$$

avec

$$R_N u = \sum_{k=0}^{N} \left\{ \frac{2}{\pi c_k} \sum_{m>N} \hat{u}_m \left(T_m, T_k\right)_N \right\} T_k \tag{1.32}$$

$$P_N u = \sum_{k=0}^{N} \hat{u}_k T_k$$

$R_N u$ peut être assimilée à l'erreur d'aliasing dûe à l'interpolation par analogie avec (1.9). De plus, l'erreur d'aliasing est orthogonale à l'erreur de troncature, $u - P_N u$, de telle sorte que

$$| u - I_N u |^2_{L^2_w(-1,1)} = | u - P_N u |^2_{L^2_w(-1,1)} + | R_N u |^2_{L^2_w(-1,1)}$$

1.2.3.4 Différentiation.

Si l'on se place dans l'espace transformé, notons $u_N = P_N u$ la troncature à l'ordre N de son développement en série de Chebyshev :

$$u_N(x) = \sum_{k=0}^{N} \hat{u}_k T_k(x)$$

alors

$$u_N^{(1)}(x) = \sum_{k=0}^{N} \hat{u}_k T_k^{(1)}(x) = \sum_{k=0}^{N} \hat{u}_k^{(1)} T_k(x)$$

Nous allons exprimer les coefficients de la dérivée en fonction de ceux de u_N

$$\begin{aligned} \hat{u}_m^{(1)} &= \left(u_N^{(1)}, T_m\right)_w = \left(\sum_{p=0}^{N} \hat{u}_p^{(1)} T_p, T_m\right)_w \\ &= \left(\sum_{n=0}^{N} \hat{u}_n T_n^{(1)}, T_m\right)_w = \sum_{n=0}^{N} \hat{u}_n \left(T_n^{(1)}, T_m\right)_w \end{aligned}$$

Ainsi

$$\begin{aligned} U_N^{(1)} &= \mathcal{D}^{(1)} U_N \quad &, \quad \mathcal{D}^{(1)}_{m,n} &= \left(T_m, T_n^{(1)}\right)_w \\ U_N^{(1)} &= \left(\hat{u}_0^{(1)}, —, \hat{u}_N^{(1)}\right)^T \quad &\text{et} \quad U_N &= \left(\hat{u}_0, —, \hat{u}_N\right)^T \end{aligned}$$

$\mathcal{D}^{(1)}$ est la matrice de dérivation première dont les coefficients non nuls sont donnés par

$$\mathcal{D}^{(1)}_{m,n} = \frac{2n}{c_m}, \text{ avec } m - n + 1 \leqslant 0 \text{ et pair}$$

De même pour la matrice de dérivation seconde dont les coefficients non nuls sont donnés par

$$\mathcal{D}^{(2)}_{m,n} = \frac{n}{c_m}\left(n^2 - m^2\right), \text{ avec } m - n + 2 \leqslant 0 \text{ et pair}$$

Il existe des relations de récurrence pour déterminer les $\hat{u}_k^{(1)}$ et $\hat{u}_k^{(2)}$ que l'on obtient à partir de (1.20) :

$$\hat{u}_k^{(1)} = \frac{2(k+1)}{c_k} \hat{u}_{k+1} + \frac{e_{k+3}}{c_k} \hat{u}_{k+2}^{(1)}, \quad k = (N-1), —, 0 \tag{1.33}$$

$$\text{avec} \quad \left\{ \begin{array}{lcl} e_k & = & 1 \quad , \quad \text{si } k \leqslant N \\ e_k & = & 0 \quad , \quad \text{si } k > N \end{array} \right. \quad \text{et} \quad \left\{ \begin{array}{lcl} c_0 & = & 2 \\ c_k & = & 1 \quad , \quad \text{si } k \geqslant 1 \end{array} \right. \tag{1.34}$$

et pour la dérivée seconde

$$\hat{u}_k^{(2)} = \frac{2(k+1)}{c_k} \hat{u}_{k+1}^{(1)} + \frac{e_{k+4}}{c_k} \hat{u}_{k+2}^{(2)},\ k = (N-2), \text{—}, 0 \tag{1.35}$$

Les dérivées première et seconde de $I_N u(x)$ sont calculées aux points de collocation (1.21) à l'aide de matrices pleines dont une forme explicite peut être trouvée dans ([39, 53]). Contrairement au cas Fourier, la troncature et l'interpolation ne permutent pas avec la différentiation :

$$(P_N u)' \neq P_N (u')$$
$$(I_N u)' \neq I_N (u')$$

1.2.3.5 Produit de convolution.

Les non-linéarités quadratiques génèrent aussi des produits de convolution comme dans le cas Fourier

$$w(x) = u(x)v(x)$$

avec

$$\hat{w}_k = \widehat{(uv)}_k = \frac{1}{2} \left\{ \sum_{m+n=k} \hat{u}_m \hat{v}_n + \sum_{|m-n|=k} \hat{u}_m \hat{v}_n \right\}$$

On utilise les Transformées de Chebyshev Discrètes et on obtient

$$\tilde{W}_k = \hat{w}_k + \frac{1}{2} \left\{ \sum_{m,n=0}^{N} \hat{u}_m \hat{v}_n \delta_{m+n,2N\pm k} + \sum_{m,n=0}^{N} \hat{u}_m \hat{v}_n \delta_{|m-n|,2N\pm k} \right\} \tag{1.36}$$

Le second terme à droite dans l'égalité (1.36) constitue l'erreur d'aliasing. Nous utilisons là aussi la méthode du padding pour déaliaser, i.e. des Transformations de Chebyshev Discrètes avec $M = \dfrac{3}{2}N$ points qui nécessitent $\mathcal{O}(M \log_2 M)$ opérations.

1.2.4 Polynômes de Legendre.

1.2.4.1 Formules de base.

Les polynômes de Legendre $\{L_k\}_{k\in\mathbb{N}}$ sont les fonctions propres du problème de Sturm-Liouville singulier :

$$\left(\sqrt{1-x^2} L_k'(x)\right)' + k(k+1) L_k(x) = 0$$

qui correspond à (1.18) avec

$$p(x) = 1 - x^2 \quad , \quad q(x) = 0 \quad \text{et} \quad w(x) = 1.$$

Pour tout k, $L_k(x)$ a la parité de k et on le normalise de telle sorte que $L_k(1) = 1$.
Ils vérifient la relation de récurrence

$$(k+1)L_{k+1}(x) = (2k+1)xL_k(x) - kL_{k-1}(x)$$
$$L_0(x) = 1 \quad , \quad L_1(x) = x \ , \ L_2(x) = \frac{3}{2}(x^2 - 1)$$

Nous allons énoncer quelques propriétés des polynômes de Legendre dont nous aurons besoin ultérieurement

$$\begin{aligned} |L_k(x)| &\leqslant 1 \qquad |x| \leqslant 1, \qquad k \geqslant 0 \\ L_k(\pm 1) &= (\pm 1)^k \\ (2k+1)L_k(x) &= L'_{k+1}(x) - L'_{k-1}(x) \qquad k \geqslant 0 \\ \int_{-1}^{1} L_k^2(x)\,\mathrm{d}x &= \frac{2}{2k+1} \end{aligned} \tag{1.37}$$

1.2.4.2 Développement de Legendre continu.

Les polynômes de Legendre $\{L_k\}_k$ sont orthogonaux sur l'intervalle $[-1,1]$ pour le produit scalaire usuel :

$$(f,g) = \int_{-1}^{1} f(x)g(x)\,\mathrm{d}x$$

pour $f, g \in L^2(-1,1)$ et ils vérifient la relation d'orthogonalité

$$(L_m, L_n) = \frac{2}{2n+1}\delta_{m,n}$$

Le développement en série de Legendre d'une fonction $f \in L^2(-1,1)$ est

$$u(x) = \sum_{k\in\mathbb{N}} \hat{u}_k L_k(x)$$

avec

$$\hat{u}_k = \frac{2}{2k+1}\int_{-1}^{1} f(x)L_k(x)\,\mathrm{d}x$$

Contrairement aux polynômes de Chebyshev, nous ne pouvons pas facilement nous ramener à une série de Fourier pour obtenir des résultats de décroissance de $|u_k|$.

1.2.4.3 Développement de Legendre discret.

Bien que des formules explicites donnant les noeuds de quadrature ne soient pas connues, il est possible d'approcher numériquement ces derniers. Nous avons le résultat suivant :

Formule d'intégration de Gauss-Lobatto :
Pour un entier $N > 0$ donné, on considère l'ensemble des points

$$-1 = x_0, x_j (j = 1, ..., N-1) \text{ les racines de } L'_N(x), x_N = 1$$

et l'ensemble des poids correspondants

$$w_j = \frac{2}{N(N+1)}\frac{1}{[L_N(x_j)]^2}\,, \quad j \in \mathbb{J}_N$$

Alors nous avons le résultat suivant :

$$\sum_{j=0}^{N} p(x_j)w_j = \int_{-1}^{1} p(x)w(x)\,\mathrm{d}x\,,\ \forall p \in \mathcal{P}_{2N-1}$$

Bien qu'il n'existe pas de véritable Transformée de Legendre Rapide (FLT) on peut cependant évaluer les coefficient polynomiaux discrets en deux étapes :

- on effectue une Transformée de Chebyshev Rapide
- on effectue un changement de base polynomiale entre les polynômes de Chebyshev et les polynômes de Legendre ([42]).

Le coût en terme de nombre d'opérations est plus important que dans le cas Fourier ou Chebyshev. L'intérêt est donc moindre pour des problèmes non linéaires nécessitant de fréquents allers-retours entre l'espace physique et l'espace spectral pour l'évaluation des termes non linéaires.

1.2.4.4 Différentiation.

Soit u une fonction admettant le développement

$$u(x) = \sum_{k\in\mathbb{N}} \hat{u}_k L_k(x)$$

Alors sa dérivée, $u^{(1)}$, peut être (formellement) représentée par

$$u^{(1)} = \sum_{k\in\mathbb{N}} \hat{u}_k^{(1)} L_k(x)$$

avec

$$\hat{u}_k^{(1)} = (2k+1) \sum_{\substack{p=k+1 \\ p+k \text{ impair}}}^{\infty} \hat{u}_p \tag{1.38}$$

Soit N un entier et $u_N = P_N u$ la troncature à l'ordre N de son développement en série de Legendre. On note

$$\begin{array}{rclrcl} U_N &=& (\hat{u}_0, \text{—}, \hat{u}_N)^T & U_N^{(1)} &=& \mathcal{D}^{(1)}\, U_N \\ U_N^{(1)} &=& \left(\hat{u}_0^{(1)}, \text{—}, \hat{u}_N^{(1)}\right)^T & \mathcal{D}^{(1)}_{m,n} &=& \left(L_m, L_n^{(1)}\right) \end{array}$$

$\mathcal{D}^{(1)}$ est la matrice de dérivation première dont les coefficients non nuls sont donnés par

$$\mathcal{D}^{(1)}_{m,n} = (2m+1)\,, \text{ avec } m-n+1 \leqslant 0 \text{ et pair}$$

De même pour la matrice de dérivation seconde dont les coefficients non nuls sont donnés par

$$\mathcal{D}^{(2)}_{m,n} = \frac{2m+1}{2}\left[n(n+1) - m(m+1)\right]\,, \text{ avec } m-n+2 \leqslant 0 \text{ et pair}$$

Il existe des relations de récurrence pour déterminer les $\hat{u}_k^{(1)}$ et les $\hat{u}_k^{(2)}$ que l'on obtient à partir de (1.37) :

$$\hat{u}_{k-1}^{(1)} = (2k-1)\left[\hat{u}_k + \frac{e_{k+2}}{2k+3}\hat{u}_{k+1}^{(1)}\right], \quad k = N, \text{—}, 1$$

et pour la dérivée seconde

$$\hat{u}_{k-1}^{(2)} = (2k-1)\left[\hat{u}_k^{(1)} + \frac{e_{k+3}}{2k+3}\hat{u}_{k+1}^{(2)}\right], \; k = (N-1), \text{—}, 1$$

Une conséquence de (1.38) est la non-commutativité de la troncature et la différentiation.

$$(P_N u)' \neq P_N (u')$$

Nous présentons maintenant les deux méthodes pseudo-spectrales que nous appliquerons à l'équation de Burgers stochastique. Nous décrivons leur mise en œuvre sur l'équation de Burgers déterministe. Dans le cas de conditions aux limites périodiques, nous nous intéressons au découpage de la solution en deux niveaux (en basses et en hautes fréquences) ainsi qu'une comparaison numérique de l'erreur d'aliasing et des instabilités qu'elle génère.
Pour des conditions aux limites de non glissement (Dirichlet homogène), la discrétisation en temps conduit à un problème de Helmholtz que nous résolvons à l'aide d'un solveur direct. Pour ce cas nous étudions numériquement la stabilité des schémas en temps à l'aide du spectre d'énergie de la solution.

1.3 Méthode Fourier-Galerkin.

Nous allons maintenant présenter la méthode pseudo-spectrale Fourier-Galerkin appliquée à l'équation de Burgers déterministe.

1.3.1 Discrétisation en espace.

On considère l'équation de Burgers en une dimension d'espace

$$\frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} + u \frac{\partial u}{\partial x} = f \quad , \text{dans} \quad (0, 2\pi) \times \mathbb{R}_+ \tag{1.39}$$

munie de conditions aux limites périodiques

$$u(0,\, t) = u(2\pi,\, t),\ t \in \mathbb{R}_+ \tag{1.40}$$

et de la condition initiale

$$u(x, 0) = u_0(x) \quad , \ x \in (0, 2\pi) \tag{1.41}$$

Soit $\{\Phi_k\}_{k\in\mathbb{Z}}$, $\Phi_k(x) = e^{ikx}$, la famille des polynômes trigonométriques orthogonaux pour le produit scalaire

$$(u, v) = \int_0^{2\pi} u(x)\, \overline{v(x)}\, \mathrm{d}x$$

qui vérifient individuellement les conditions aux limites (1.40). La solution approchée de u, $u_N(x,t)$, est un polynôme de degré $N/2$ égal à la troncature de la série de Fourier de u à l'ordre N pour $N > 0$ la fréquence de coupure :

$$u_N(x, t) = P_N\left(u(x,t)\right) = \sum_{k\in\mathbb{I}_N} \hat{u}_k(t) \Phi_k(x)$$

avec

$$\mathbb{I}_N = \left[\!\left[1 - \frac{N}{2}, \frac{N}{2} \right]\!\right]$$

dont les coefficients du développement constituent les inconnues du problème. On demande l'orthogonalité du résidu de (1.39) à toutes les fonctions de $\mathcal{S}_N$, l'espace vectoriel engendré par $\{\Phi_k\}_{k\in\mathbb{I}_N}$. Pour cela, on projette l'équation (1.39) dans $\mathcal{S}_N$:

$$\int_0^{2\pi}\left\{\frac{\partial u_N}{\partial t}-\nu\frac{\partial^2 u_N}{\partial x^2}+P_N\left(u_N\frac{\partial u_N}{\partial x}\right)\right\}e^{-ikx}\,\mathrm{d}x=\int_0^{2\pi}P_N f e^{-ikx}\,\mathrm{d}x$$

pour $k\in\mathbb{I}_N$. Comme la famille $\{\Phi_k\}_{k\in\mathbb{I}_N}$ forme un système orthogonal pour le produit scalaire $(.,.)$, nous obtenons

$$\frac{d\hat{u}_k(t)}{dt}+\nu k^2\hat{u}_k(t)+\widehat{NL}_k(t)=\hat{f}_k(t),\quad k\in\mathbb{I}_N \tag{1.42}$$

avec

$$\widehat{NL}_k(t)=\left(\widehat{u_N\frac{\partial u_N}{\partial x}}\right)_k(t)=\frac{1}{2\pi}\int_0^{2\pi}P_N\left(u_N\frac{\partial u_N}{\partial x}\right)e^{-ikx}\,\mathrm{d}x \tag{1.43}$$

pour $k\in\mathbb{I}_N$.
La condition initiale s'écrit

$$\hat{u}_k(0)=\frac{1}{2\pi}\int_0^{2\pi}u_0(x)e^{-ikx}\,\mathrm{d}x \tag{1.44}$$

Les équations (1.42) et (1.44) forment un système complet d'équations différentielles ordinaires.

Remarque *1*

L'équation (1.43) est un cas particuler de terme non linéaire quadratique dont l'évaluation a été vue au paragraphe §1.1.4

1.3.2 Discrétisation en temps.

Après avoir décrit la discrétisation spatiale qui nous a permis d'obtenir le système d'équations différentielles ordinaires, (1.42) - (1.44), nous allons considérer maintenant l'intégration en temps de ce système. L'équation (1.42)

$$\frac{d\hat{u}_k(t)}{dt}+\nu k^2\hat{u}_k(t)+\widehat{NL}_k(t)=\hat{f}_k(t),\ k\in\mathbb{I}_N$$

est multipliée par $e^{\nu k^2 t}$, ce qui donne

$$\frac{d\left(e^{\nu k^2 t}\hat{u}_k(t)\right)}{dt}=\left[\hat{f}_k(t)-\widehat{NL}_k(t)\right]e^{\nu k^2 t},\ k\in\mathbb{I}_N \tag{1.45}$$

Nous écrivons le système (1.45) sous la forme matricielle

$$\frac{d\left(e^{At}u(t)\right)}{dt}=\left[f(t)-NL(u(t))\right]e^{At} \tag{1.46}$$

où

e^{At} est	la matrice diagonale de coefficient $\left(e^{At}\right)_{k,k} = e^{\nu k^2 t},\ k \in \mathbb{I}_N$
$u(t)$	le vecteur $(\hat{u}_k(t))_{k\in\mathbb{I}_N}$,
$NL(u(t))$	le vecteur $\left(\widehat{NL}_k(t)\right)_{k\in\mathbb{I}_N}$,
$f(t)$	le vecteur $\left(\hat{f}_k(t)\right)_{k\in\mathbb{I}_N}$.

On intègre alors l'équation (1.46) au moyen d'un schéma de Runge-Kutta d'ordre 3 explicite. Cette étape d'intégration s'exprime :
on note Δt le pas de la discrétisation en temps et $t_n = n\Delta t$.

$$\begin{aligned}
u_0 &= u^n = u(t_n)\\[1em]
G_1 &= e^{-\frac{\Delta t}{3}A}\Delta t\left\{NL(u_0) + f(t_n)\right\}\\
u_1 &= e^{-\frac{\Delta t}{3}A}u_0 + G_1\\[1em]
G_2 &= e^{-\frac{5\Delta t}{12}A}\left\{\Delta t\left[NL(u_1) + f(t_{n,2})\right] - \frac{5}{9}G_1\right\};t_{n,2} = t_n + \frac{\Delta t}{3}\\
u_2 &= e^{-\frac{5\Delta t}{12}A}u_1 + \frac{15}{16}G_2\\[1em]
G_3 &= e^{-\frac{\Delta t}{4}A}\left\{\Delta t\left[NL(u_2) + f(t_{n,3})\right] - \frac{153}{128}G_2\right\};t_{n,3} = t_n + \frac{3\Delta t}{4}\\
u_3 &= e^{-\frac{\Delta t}{4}A}u_0 + G_1\\[1em]
u^{n+1} &= u(t_{n+1}) = u_3
\end{aligned}$$

Cette technique d'intégration qui consiste à multiplier les équations (1.42) par $e^{\nu k^2 t}$ à gauche et à droite des égalités est appelée *méthode de quadrature*: elle permet une intégration en temps exacte de la partie linéaire (terme de diffusion). Ainsi le traitement du terme linéaire est inconditionnellement stable.
Les restrictions de précision temporelle et de stabilité proviennent seulement de l'intégration en temps du terme non linéaire (terme de convection).
L'ordre élevé du schéma d'intégration en temps (ordre 3) se justifie par le fait que l'approximation spatiale par une méthode spectrale conduit à des schémas très précis. Ce sera donc la précision du schéma en temps qui sera limitative et qui fixera la précision globale de résolution du système (1.39) - (1.41).

1.3.3 Stabilité.

La stabilité numérique d'un schéma signifie qu'une petite perturbation dans le calcul de la solution ne s'amplifie pas au cours du temps. Une perturbation désigne ici l'erreur dûe à la discrétisation du problème continu ou bien aussi les erreurs d'arrondi à chaque étape du calcul.
Une étude de stabilité pour le schéma de Runge-Kutta explicite d'ordre 3 a été menée par

F. Jauberteau ([31]). On obtient une contrainte de stabilité de type Courant-Friedrichs-Levy (C.F.L.) de la forme :

$$N\Delta t \, |u_N|_{L^\infty(0,2\pi)} \; < \; \text{constante.}$$

1.3.4 Convergence spatiale.

Nous définissons l'expression

$$E(k) = \sum_{\substack{|l|=k \\ l\in\mathbb{Z}}} |\hat{u}_l|^2 = |\hat{u}_{-k}|^2 + |\hat{u}_k|^2$$

qui quantifie l'énergie présente dans le spectre.
Nous distinguons différentes zones dans le tracé de $E(k)$:
- *zone d'injection ou de forçage*: elle correspond aux modes qui seront alimentés directement en énergie par la force extérieure (simulation d'un écoulement avec une force déterministe);
- *zone inertielle ou de transfert*: elle correspond aux transferts d'énergie des grandes échelles (i.e. les grandes structures de l'écoulement) vers les petites. Ce transfert d'énergie est dû, d'une part à une force extérieure suffisamment grande pour apporter suffisamment d'énergie sur les grandes structures et d'autre part à une viscosité suffisamment faible pour permettre le développement d'instabilités et la création de structures plus petites alimentées par les structures plus grandes. c'est le terme non linéaire.
- *zone visqueuse ou de dissipation*: elle correspond aux modes associés aux structures trop petites pour se développer. Cela se traduit numériquement par la présence dans les équations discrétisées du terme multiplicateur $e^{-\nu k^2 \Delta t}$ qui est d'autant plus petit que $|k|$ est grand. Ainsi la viscosité ν amortit plus, au cours du temps, les coefficients de u associés aux modes élevés (petites structures) que ceux associés aux petits modes (grandes structures). Cela explique la décroissance des coefficients du développement en série de Fourier de u et traduit une certaine régularité de la solution u dès que $\nu \neq 0$, cet amortissement étant d'autant plus faible que ν est petite.

1.3.5 Résolution sur deux niveaux.

Soit N la fréquence de coupure.
Pour un entier N_1, $0 < N_1 < N$, nous définissons les opérateurs de projection P_{N_1} et Q_{N_1} et les quantités y_{N_1} et z_{N_1} par

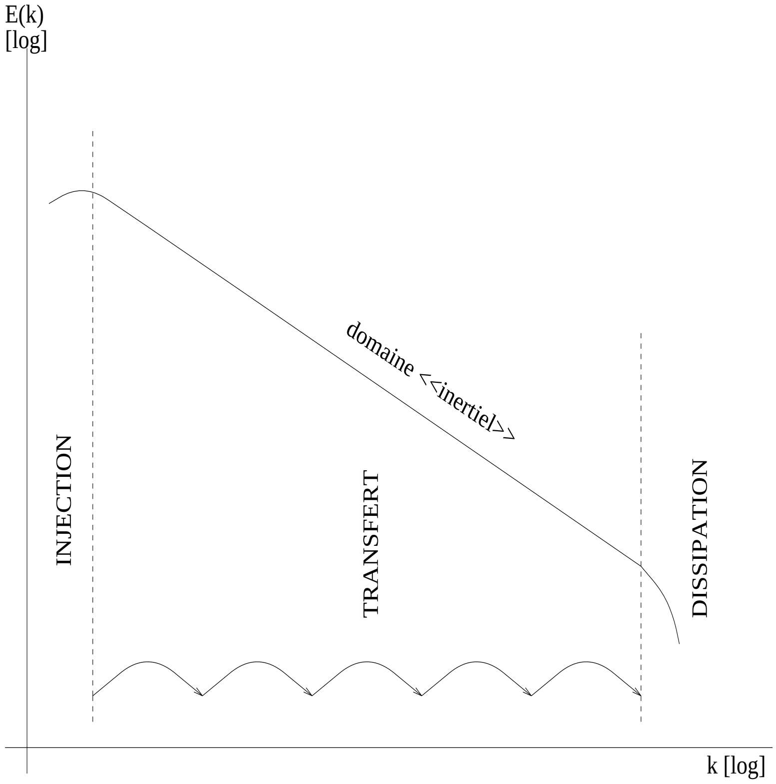
$$\begin{aligned} y_{N_1} &= P_{N_1}(u_N) = \sum_{k\in\mathbb{I}_N} \hat{u}_k \Phi_k \\ z_{N_1} &= Q_{N_1}(u_N) = (Id_N - P_{N_1})(u_N) \end{aligned}$$

ainsi que l'opérateur

$$B(u,v) = u\frac{\partial v}{\partial x}; \;\; u, v, \in \mathcal{S}_N.$$

De plus nous posons

$$B_{int}(y,z) = B(y,z) + B(z,y) + B(z,z)$$



Nous avons alors

$$\begin{aligned} B(u_N, u_N) &= B(y_{N_1}, z_{N_1}) \\ &= B(y_{N_1}, y_{N_1}) + B_{int}(y_{N_1}, z_{N_1}) \end{aligned}$$

Pour alléger les notations, remplaçons y_{N_1}, z_{N_1}, P_{N_1} et Q_{N_1} respectivement par y, z, P et Q. Les quantités y et z sont gouvernées par le système d'équations

$$\begin{aligned} \frac{\partial y}{\partial t} - \nu \frac{\partial^2 y}{\partial x^2} + PB(y,y) + PB_{int}(y,z) &= Pf \\ \frac{\partial z}{\partial t} - \nu \frac{\partial^2 z}{\partial x^2} + QB(y,y) + QB_{int}(y,z) &= Qf \end{aligned}$$

y désigne les grandes échelles et z les petites qui animent l'écoulement.
cette séparation en deux niveaux (ou plus si on utilise différentes valeurs pour N_1) permet de déterminer : le nombre de modes nécessaire pour discrétiser la solution (cela nous sera très utile pour les simulations dont nous ne connaissons pas les solutions exactes), et les parties du spectre de la solution qui contiennent l'énergie de l'écoulement et les termes dominants dans les équations régissant les petites et les grandes échelles.

1.3.6 Validation numérique.

On considère un exemple dont on connait la solution exacte afin, d'une part de tester la mise en œuvre de la méthode, d'autre part de vérifier la précision du schéma. On construira alors la force extérieure, i.e. le second membre de l'équation de Burgers, ad hoc :

$$f = \frac{\partial u_{ex}}{\partial t} - \nu \frac{\partial^2 u_{ex}}{\partial x^2} + u_{ex}\frac{\partial u_{ex}}{\partial x} \text{ , où } u_{ex} \text{ désigne la solution exacte.}$$

On choisit comme solution exacte $u_{ex}(x,t) = g(t)h(x)$ avec

$$\begin{aligned} g(t) &= \frac{1}{40}\left\{\frac{1}{10}\exp\left[\sin(2t) - 3\sin(2\pi t)\right] - \cos(2\sqrt{2}t)\right\} + 0,075 \ , \\ h(x) &= \cos(\alpha_1 x)\,\sin(\alpha_2 x)\,\exp\left[\cos(\alpha_1 x)\,\cos(\alpha_2 x)\right] \end{aligned}$$

et les scalaires valent $\alpha_1 = 3$, $\alpha_2 = 2$.
Les paramètres de l'exécution sont les suivants :
- la fréquence de coupure N est fixée à 96 modes,
- le pas de temps Δt est pris égal à 10^{-2},
- la viscosité ν vaut 10^{-4}.

Nous avons représenté la composante temporelle de la vitesse, $g(t)$. Celle-ci est à l'origine des variations des différentes quantités tracées.
En effet :

$$|u_N|_{L^\infty} = |g(t)\,h(x)|_{L^\infty} = |g(t)|\,\max_{j\in\mathbb{J}_N}|h(x_j)| = cste\,|g(t)|$$

Il en est de même pour le nombre de Courant et $|\,u_N\,|_{L^2(0,2\pi)}$ (figures 1.1 et 1.2). L'erreur obtenue est un peu supérieure à l'erreur théorique $\mathcal{O}(\Delta t^2) \sim \mathcal{O}(10^{-6})$. Elle reprend le comportement oscillant de la fonction $g(t)$.

1.3.7 Comparaison numérique aliasing-déaliasing.

Pour mettre en évidence l'erreur d'aliasing et son caractère destabilisant, nous utilisons une solution exacte : celle étudiée au paragraphe précédent. Nous travaillons avec une petite fréquence de coupure ($N = 48,\ 64,\ 96$) pour que l'erreur spatiale domine l'erreur du schéma en temps. Nous reprenons les autres paramètres, à savoir :
$\Delta t = 10^{-2}$,
$\nu \ = 10^{-4}$,
$\alpha_1 = 3$ et $\alpha_2 = 2$.
La figure 4 représente l'évolution de l'erreur globale pour les différentes valeurs de N, en laissant ou en ôtant les termes d'aliasing.
Pour N= 48 et 64 il y a explosion du calcul aliasé avant $t = 50$ et explosion vers $t = 100$ pour $N = 96$. On voit donc clairement l'effet néfaste de l'aliasing sur la stabilité et la précision des calculs; d'où la nécessité de la supprimer pour les simulations sur de grands temps, en particulier pour le cas de forçage aléatoire.

FIG. 1.1 – *Evolution en temps de la fonction $g(t)$ de la solution (gauche) et du nombre de Courant (droite).*

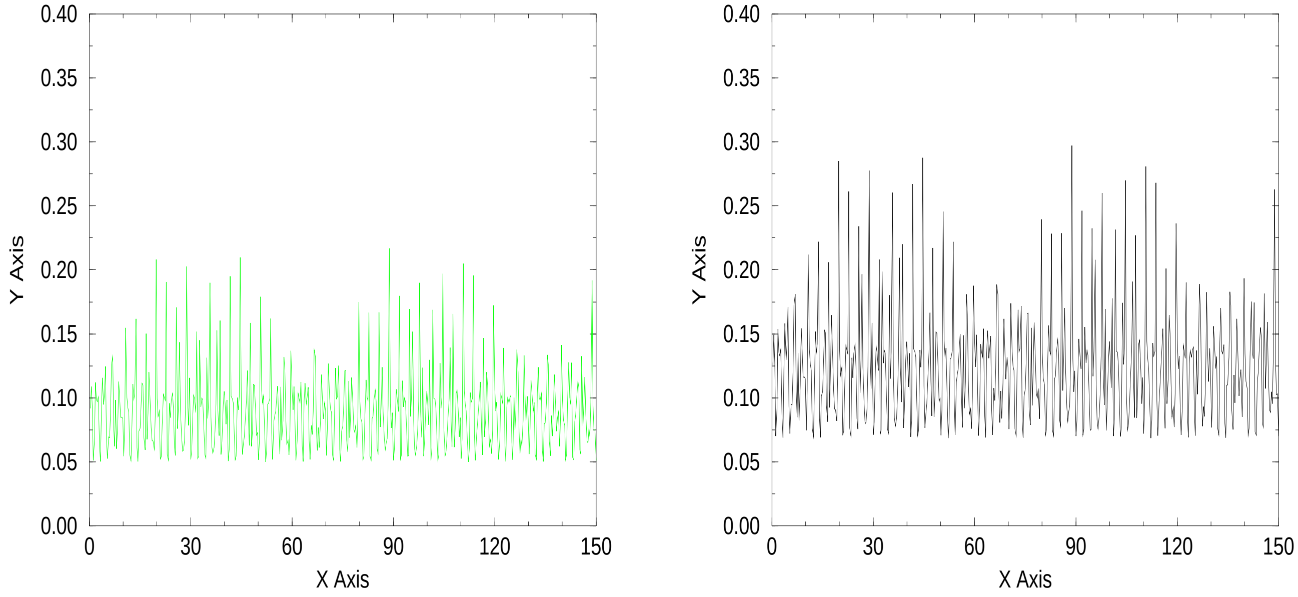


FIG. 1.2 – *Evolution en temps des normes* $\mathbf{L}^2$ *(gauche) et* $\mathbf{L}^\infty$ *(droite) de la solution.*

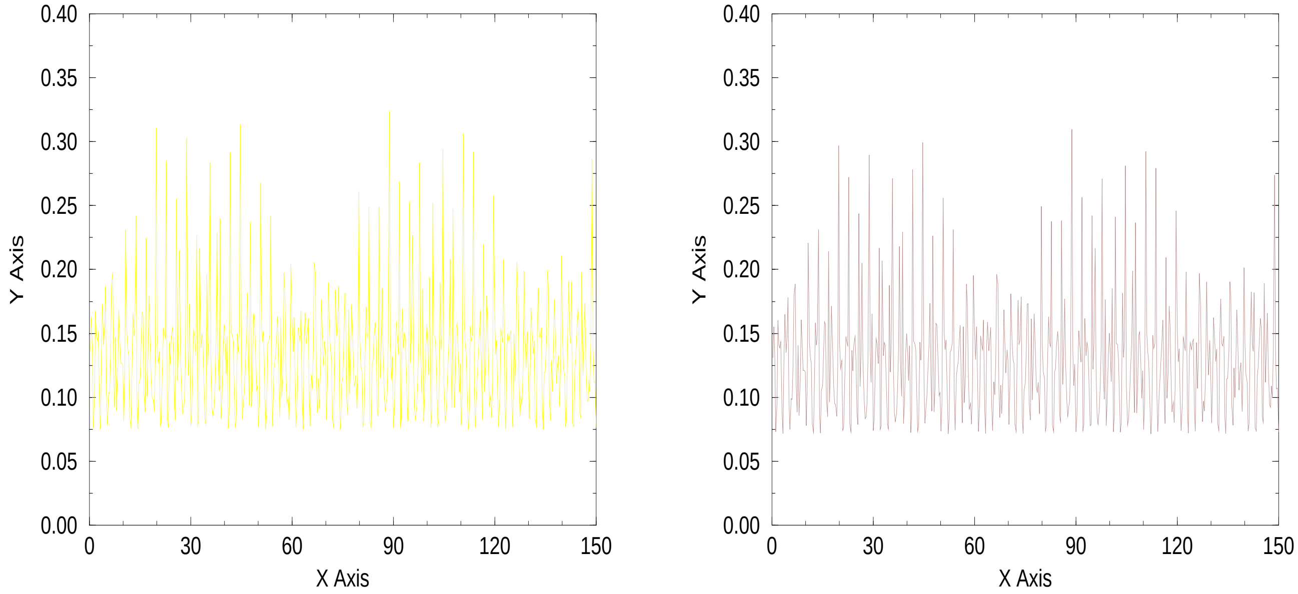


FIG. 1.3 – *Evolution en temps des quantités* $|y_{N_1}|_{L^2(0,2\pi)}$ *(gauche) et* $|z_{N_1}|_{L^2(0,2\pi)}$ *(droite) pour les valeurs* $N_1 = 16$ *(1), 32 (2), 48 (3), 64 (4) avec* $N = 96$.

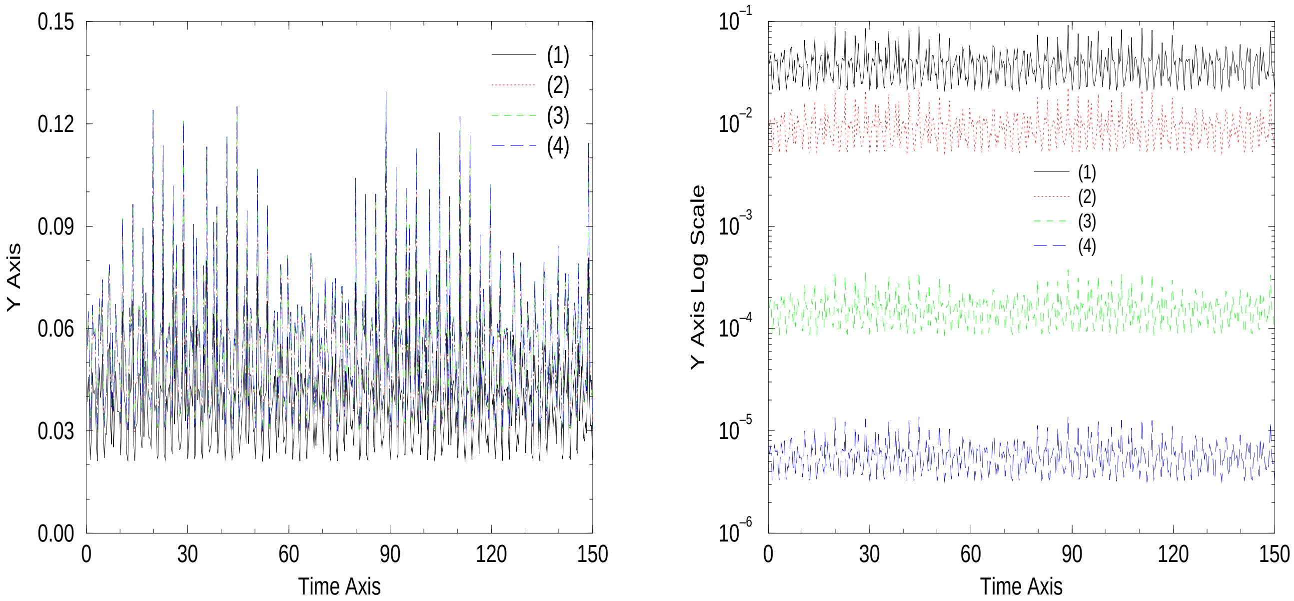
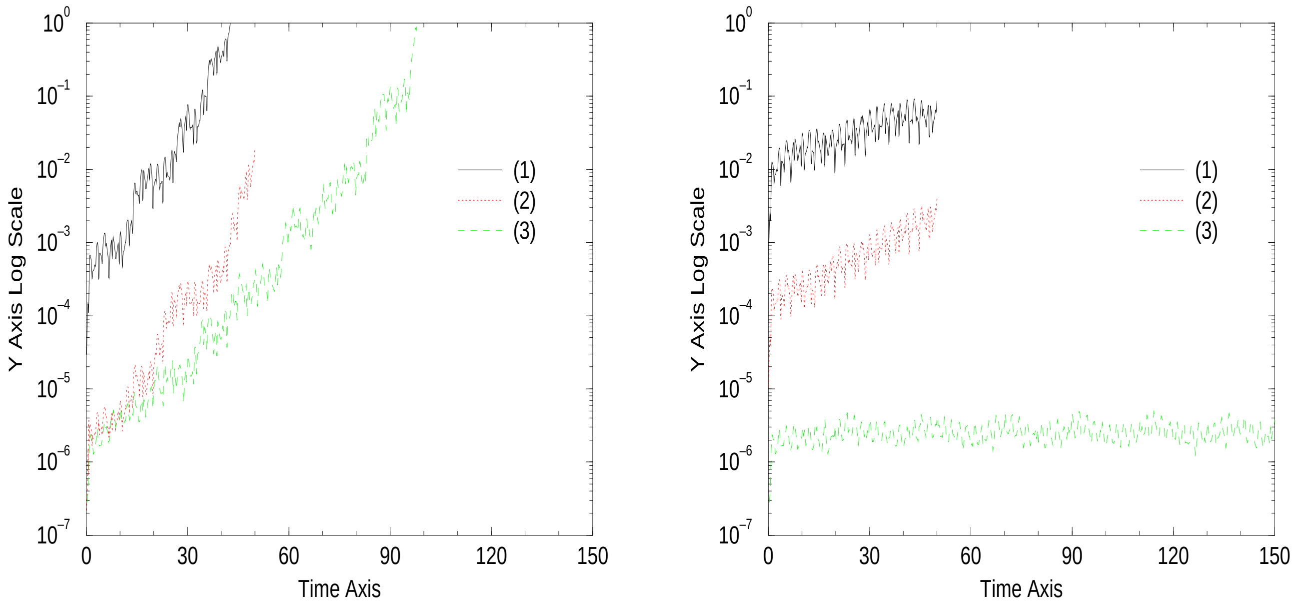


FIG. 1.4 – *Evolution en temps de la norme* L^2 *de l'erreur relative pour les schémas aliasé (gauche) et déaliasé (droite) pour* $N = 48$ *(1), 64 (2), 96 (3).*



1.4 Méthode Tau-Chebyshev.

Nous allons maintenant considérer une autre méthode spectrale : la méthode Tau-Chebyshev appliquée à l'équation de Burgers déterministe.

1.4.1 Discrétisation en espace.

On considère l'équation de Burgers en une dimension d'espace

$$\frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} + u \frac{\partial u}{\partial x} = f \ , \quad \text{dans } (-1,1) \times \mathbb{R}_+ \tag{1.47}$$

munie de conditions aux limites de type Dirichlet homogène

$$u(-1,\, t) = u(1,\, t) = 0 \ , \quad t \in \mathbb{R}_+ \tag{1.48}$$

et de la condition initiale

$$u(x,0) = u_0(x) \ , \quad x \in (-1,1) \tag{1.49}$$

Soit $\{T_k\}_{k\in\mathbb{N}}$ la famille des polynômes de Chebyshev orthogonaux pour le produit scalaire

$$(u,v)_w = \int_{-1}^{1} u(x)\, v(x)\, w(x)\, \mathrm{d}x$$

où $w(x) = (1-x^2)^{\frac{1}{2}}$ est la fonction de poids associée aux polynômes de Chebyshev.
La solution approchée de u, u_N, est un polynôme de degré N égal à la troncature de la série de Chebyshev de u à l'ordre N, pour N la fréquence de coupure.

$$u_N(x,t) = P_N\left(u(x,t)\right) = \sum_{k=0}^{N} \hat{u}_k(t) T_k(x)$$

dont les coefficients du développement dans cette base forment les inconnues du problème.
On demande l'orthogonalité du résidu de (1.47) à tous les polynômes de degré inférieur ou égal à $N-2$, i.e. $\mathcal{P}_{N-2}$.
On projette orthogonalement l'équation (1.47) dans l'espace des polynômes de degré inférieur ou ègal à $N-2$ à l'aide du produit scalaire $(.,.)_w$:

$$\int_{-1}^{1} \left\{ \frac{\partial u_N}{\partial t} - \nu \frac{\partial^2 u_N}{\partial x^2} + P_N\left(u_N \frac{\partial u_N}{\partial x} \right) \right\} T_w\, w\, \mathrm{d}x = \int_{-1}^{1} P_N f\, T_k\, w\, \mathrm{d}x$$

pour $k \in [\![0,\, N-2]\!]$.
Comme la famille $\{T_k\}_{k\in\mathbb{N}}$ forme un système orthogonal pour le produit scalaire $(.,.)_w$, nous obtenons

$$\frac{d\hat{u}_k(t)}{dt} - \nu \hat{u}_k^{(2)}(t) + \widehat{NL}_k(t) = \hat{f}_k(t) \ , \quad k \in [\![0,\, N-2]\!] \tag{1.50}$$

en notant

$$
\begin{aligned}
\frac{\partial^2 u_N}{\partial x^2} &= \sum_{k=0}^{N} \hat{u}_k^{(2)}(t)\, T_k(x) \;,\; \hat{u}_k^{(2)}(t) = \frac{2}{\pi c_k} \int_{-1}^{1} \frac{\partial^2 u_N}{\partial x^2}\, T_k\, w \,\mathrm{d}x \\
P_N\left(u_N \frac{\partial u_N}{\partial x}\right) &= \sum_{k=0}^{N} \widehat{NL}_k(t)\, T_k(x) \\
\widehat{NL}_k(t) &= \frac{2}{\pi c_k} \int_{-1}^{1} P_N\left(u_N \frac{\partial u_N}{\partial x}\right) T_k\, w \,\mathrm{d}x
\end{aligned}
$$

Les conditions aux limites (1.48) peuvent se mettre sous la forme

$$
\left\{ \begin{aligned} u_N(-1,t) &= 0 \\ u_N(\;\;1,t) &= 0 \end{aligned} \right. \quad \Longrightarrow \quad \left\{ \begin{aligned} \sum_{k=0}^{N/2} \hat{u}_{2k}(t) &= 0 \\ \sum_{k=0}^{N/2-1} \hat{u}_{2k+1}(t) &= 0 \end{aligned} \right. \tag{1.51}
$$

La condition initiale (1.47) s'écrit

$$
\hat{u}_k(0) = \frac{2}{\pi c_k} \int_{-1}^{1} u_0\, T_k\, w \,\mathrm{d}x \;,\;\; \forall k \in [\![0,\; N]\!] \tag{1.52}
$$

Les équations (1.50) - (1.52) forment un système complet d'équations différentielles ordinaires.

1.4.2 Discrétisation en temps.

Nous allons considérer maintenant l'intégration en temps de ce système d'équations. Nous écrivons l'équation de Burgers sous la forme

$$
\frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} = g \text{ avec } g = f - u \frac{\partial u}{\partial x} \tag{1.53}
$$

On l'intègre en temps à l'aide des deux schémas d'ordre 2 suivants ([32]) :
- le schéma d'Adams-Bashforth explicite pour le second membre g,
- le schéma de Crank-Nicholson semi-implicite pour la partie linéaire.

On intègre l'équation (1.53) sur $[t, t+\Delta t]$, $t \geqslant 0$ et $\Delta t > 0$, cela nous donne

$$
\begin{aligned}
u(t &+ \Delta t) - \frac{\nu \Delta t}{2} \left(\frac{\partial^2 u}{\partial x^2}\right)(t+\Delta t) \\
&= u(t) + \frac{\nu \Delta t}{2} \left(\frac{\partial^2 u}{\partial x^2}\right)(t) + \frac{\Delta t}{2} \{3g(t) - g(t-\Delta t)\}
\end{aligned}
$$

A chaque pas de temps, on est amené à résoudre le problème de Helmholtz suivant :

$$
u(\tilde{t}) - \lambda \left(\frac{\partial^2 u}{\partial x^2}\right)(\tilde{t}) = h(t)
$$

avec

$$\begin{aligned}
\tilde{t} &= t + \Delta t \\
\lambda &= \frac{\nu \Delta t}{2} \\
h(t) &= u(t) + \frac{\nu \Delta t}{2}\left(\frac{\partial^2 u}{\partial x^2}\right)(t) + \frac{\Delta t}{2}\left\{3g(t) - g(t-\Delta t)\right\}
\end{aligned}$$

1.4.3 Stabilité.

Les restrictions de précision temporelle et de stabilité proviennent seulement de l'intégration en temps. On a une condition de type (C.F.L.) sévère

$$N^2 \,\Delta t \ \ |u_N|_{L^\infty_w(-1,1)} \ < \ \text{constante}$$

Cette restriction provient de l'espacement des points de collocation de Chebyshev $\left\{\cos\left(\frac{j\pi}{N}\right)\right.$, $\left. j \in \mathbb{J}_N\right\}$ de l'ordre de $\dfrac{1}{N^2}$ près des extrémités de l'intervalle $[-1, 1]$. Pour une étude plus détaillée de cette restriction, on peut consulter ([13], [21]).

1.4.4 Résolution du problème de Helmholtz.

On s'intéresse au problème de Helmholtz précédent discrétisé par la méthode Tau-Chebyshev, i.e.

$$\left\{\begin{array}{l} u_N - \lambda \dfrac{d^2 u_N}{dx^2} = h_N \ , \quad |x| \leqslant 1 \ , \quad \lambda > 0 \\ u_N(\pm 1) = 0 \end{array}\right.$$

qui s'écrit :

$$\left\{\begin{array}{l} \hat{u}_k - \lambda \hat{u}_k^{(2)} = \hat{h}_k \ , \text{ pour } k \in \mathbb{I}_{N-2} \\ \displaystyle\sum_{k=0}^{N/2} \hat{u}_{2k}(t) = 0 \ , \quad \sum_{k=0}^{N/2-1} \hat{u}_{2k+1}(t) = 0 \end{array}\right. \tag{1.54}$$

en notant

$$\frac{d^2 u_N}{dx^2} = \sum_{k=0}^{N} \hat{u}_k^{(2)} \, T_k \text{ et } h_N = \sum_{k=0}^{N} \hat{h}_k \, T_k$$

Pour exprimer les $\left\{\hat{u}_k^{(2)}\right\}_k$ en fonction des $\{\hat{u}_k\}_k$, nous utilisons les relations (1.33) et (1.35) qui, combinées, nous donnent :

$$\hat{u}_k = \alpha_k \hat{u}_{k-2}^{(2)} + \beta_k \hat{u}_k^{(2)} + \gamma_k \hat{u}_{k+2}^{(2)} \ , \quad k \in [\![2, \, N]\!] \tag{1.55}$$

en posant

$$\alpha_k = \frac{c_{k-2}}{4k(k-1)} \ , \quad \beta_k = -\frac{e_{k+2}}{2(k^2-1)} \ , \quad \gamma_k = \frac{e_{k+4}}{4k(k+1)}$$

en reprenant les notations (1.34).

Nous allons faire disparaître les $\hat{u}_k^{(2)}$ de (1.54) en écrivant

$$\alpha_k (1.54)_{k-2} + \beta_k (1.54)_k + \gamma_k (1.54)_{k+2} \text{ pour } k \in [\![2, \, N]\!]$$

la somme nous donne

$$\alpha_k \hat{u}_{k-2} + (\beta_k - \lambda)\, \hat{u}_k + \gamma_k \hat{u}_{k+2} = b_k \tag{1.56}$$

avec

$$b_k = \alpha_k \hat{h}_{k-2} + \beta_k \hat{h}_k + \gamma_k \hat{h}_{k+2}$$

Les équations (1.54) et (1.56) forment un système de $(N+1)$ équations à $(N+1)$ inconnues. On peut découpler les indices pairs des indices impairs, nous obtenons ainsi deux systèmes quasi-tridiagonaux, l'un de dimension $(N/2+1)$ pour les modes pairs, l'autre de dimension $N/2$ pour les modes impairs. Ce que l'on écrit matriciellement

$$\left\{ \begin{array}{lll} A_{\text{pair}} \ \hat{U}_{\text{pair}} & = & B_{\text{pair}} \\ A_{\text{imp}} \ \hat{U}_{\text{imp}} & = & B_{\text{imp}} \end{array} \right. \tag{1.57}$$

Les coefficients des matrices A_{pair} et A_{imp} ne dépendent que de λ et de N; comme ces matrices sont creuses, les systèmes (1.57) peuvent être résolus efficacement en utilisant une décompositionLU. Pour le premier système, e.g., la décomposition s'écrit $A_{\text{pair}} = U_{\text{pair}}\, L_{\text{pair}}$:

$$L_{\text{pair}} = \begin{pmatrix} 1 & 0 & 0 & & & & & 0 \\ l_{1,0} & 1 & 0 & 0 & & & & \\ 0 & l_{2,1} & 1 & 0 & 0 & & & \\ & & & & & & & \\ & & & & & 1 & 0 & 0 \\ 0 & & & & & & 1 & 0 \\ 0 & 0 & 0 & & & 0 & l_{\frac{N}{2},\frac{N}{2}-1} & 1 \end{pmatrix}$$

$$U_{\text{pair}} = \begin{pmatrix} u_{0,0} & u_{0,1} & u_{0,2} & & & u_{0,\frac{N}{2}} \\ 0 & u_{1,1} & u_{1,2} & & & 0 \\ 0 & 0 & u_{2,2} & & & \\ & & & & & \\ 0 & & & & u_{\frac{N}{2}-1,\frac{N}{2}-1} & u_{\frac{N}{2}-1,\frac{N}{2}} \\ 0 & 0 & 0 & & 0 & u_{\frac{N}{2},\frac{N}{2}} \end{pmatrix}$$

1.4.5 Validation numérique.

On considère la fonction suivante $u_{ex}(x,t) = g(t)\, h(x)$ avec

$$\begin{aligned} g(t) &= \frac{1}{40} \left\{ \frac{1}{10} \exp\left[\sin(2t) - 3\sin(2\pi t)\right] - \cos(2\sqrt{2}t) \right\} + 0,075 \ , \\ h(x) &= \left\{ \alpha_1 \sin(\alpha_2 \pi x) \ + \ \beta_1 \sin\left(\beta_2 \frac{\pi}{2} x\right) \ \right\} \\ &\quad \exp\{ \ \alpha_1 \cos(\alpha_1 x) + \sin(\beta_1 x) \ \} \end{aligned}$$

avec les scalaires $\alpha_1 = \sqrt{5}$, $\alpha_2 = 2$; $\beta_1 = e^1$, $\beta_2 = 7$.
On travaille avec une viscosité allant de 10^{-2} à 10^{-4} .
Pour une viscosité ν égale à 10^{-2}, $N = 48$ ou 64 modes sont suffisants pour capter la solution, figure (1.5). Pour une viscosité ν égale à 10^{-3}, $N = 48$ et 64 modes sont insuffisants pour assurer la stabilité du calcul puisque les erreurs présentent une croissance

exponentielle, figure (1.5). Enfin, une viscosité $\nu = 10^{-4}$ nécessite au moins 96 modes pour que le calcul n'explose pas, figure (1.6).
L'augmentation du nombre de modes ou la diminution du pas de temps ne permettent pas d'obtenir la stabilité complète du calcul puisque même 128 ou 144 modes et des pas de temps très petits en 10^{-5} voire 10^{-6} n'apportent pas une réponse satisfaisante.
Cela peut être expliqué par le fait que Crank-Nicholson est un schéma incapable d'approcher convenablement les modes correspondants aux hautes fréquences ([22]).

1.5 Conclusion.

Dans ce chapitre, nous avons présenté deux méthodes spectrales : la méthode de Galerkin et la méthode Tau.
Nous les avons appliquées à l'équation de Burgers déterministe dans le cas de conditions aux limites périodiques pour la première methode et de conditions aux limites de non glissement pour la seconde.
Cela a mis en évidence, dans un cadre simple, les principes de ces méthodes, basées sur le développement en fonctions régulières des solutions d'équations aux dérivées partielles, ainsi que les difficultés pouvant survenir (l'aliasing, notamment).
La discrétisation en temps a été effectuée à l'aide de méthodes directes explicites ou semi-implicites. Les calculs pour une solution exacte ont démontré la grande précision en espace que l'on pouvait espérer.

FIG. 1.5 – *Evolution en temps de la norme L^2 de l'erreur pour $\Delta t = 10^{-4}$, $\nu = 10^{-2}$ (gauche) et $\Delta t = 5.10^{-5}$, $\nu = 10^{-3}$ (droite).*

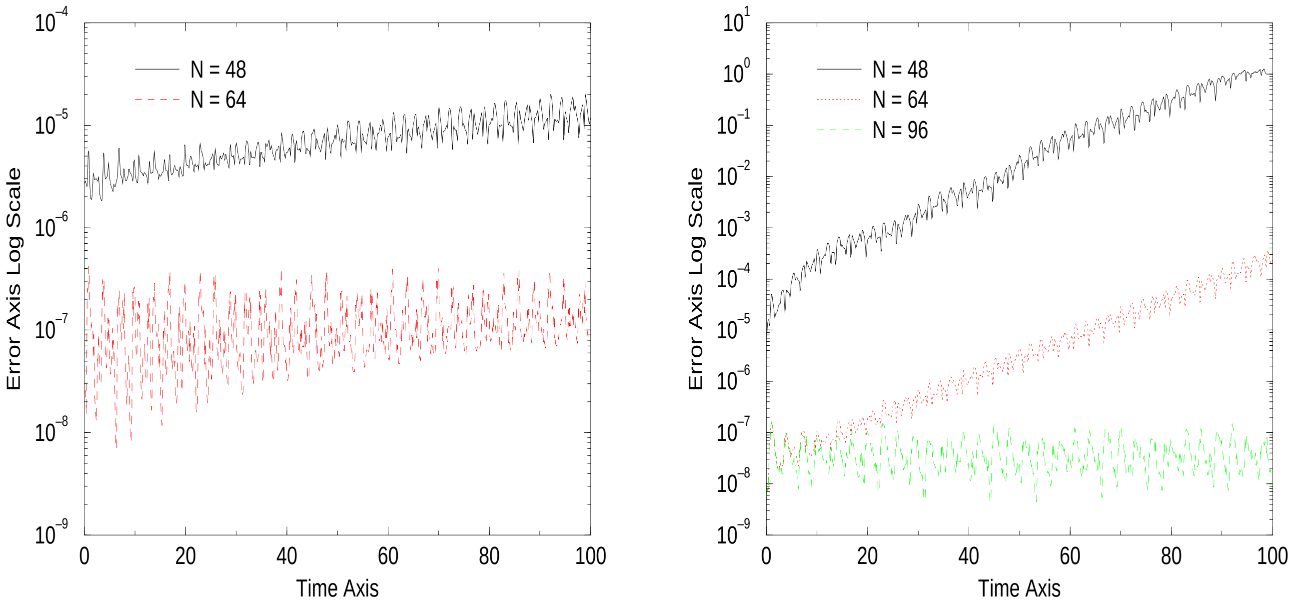
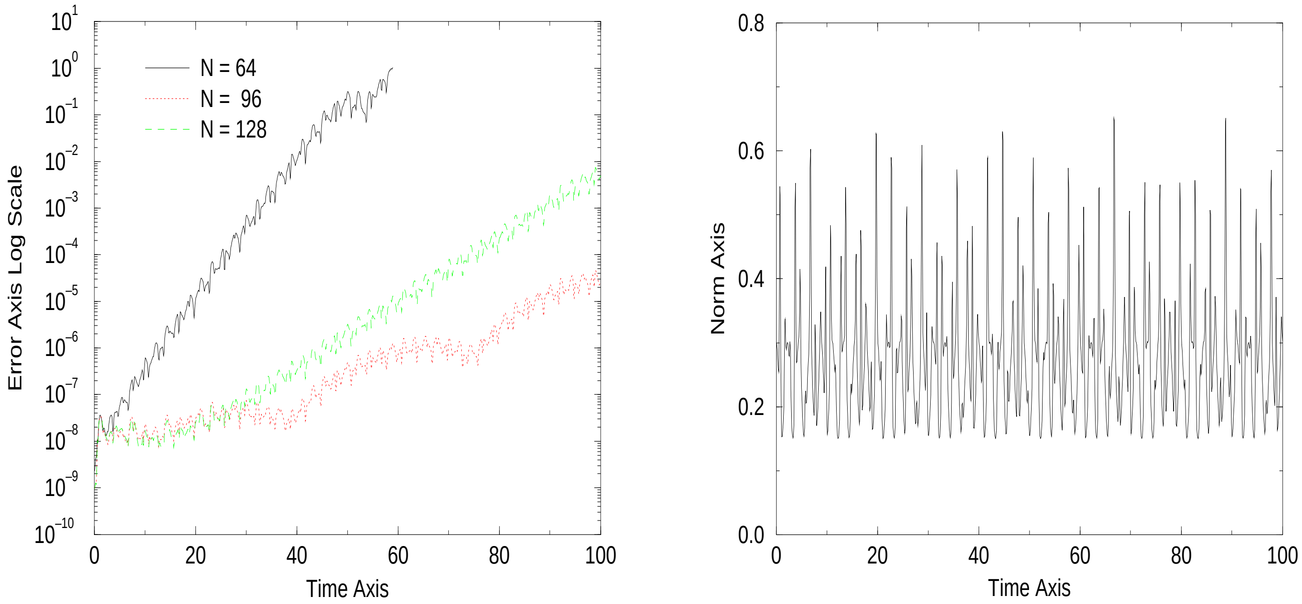


FIG. 1.6 – *Evolution en temps des normes L^2 de l'erreur (gauche) et de la solution (droite) pour $\Delta t = 2.10^{-5}$, $\nu = 10^{-4}$.*



Chapitre 2

Étude de l'équation de Schrödinger non linéaire.

2.1 Introduction

Nous nous intéressons au problème de transmission de signaux binaires par solitons via une fibre optique. Cette propagation peut être modélisée par l'équation de Schroedinger non linéaire faiblement amortie qui s'écrit en version non dimensionnalisée:

$$i\frac{\partial u}{\partial z} + \frac{1}{2}\frac{\partial^2 u}{\partial t^2} + |u|^2 u + i\alpha u = 0 \qquad (2.1)$$

où z représente l'abscisse le long de la fibre, t désigne le temps, α le coefficient d'atténuation linéique est un paramètre réel positif et u est une fonction de t et de z à valeurs complexes. Contrairement à la notation usuelle en théorie des équations aux dérivées partielles, t est la variable de "type spatial" et z la variable de "type temporel".

Le signal propagé dans la fibre est soumis à la dispersion (terme linéaire $\frac{1}{2}\frac{\partial^2 u}{\partial t^2}$) et à l'effet Kerr qui le déphase en fonction de sa puissance (terme non linéaire $|u|^2 u$).
Le soliton est la solution à la recherche d'un équilibre stable dans ce régime de propagation. C'est une impulsion de forme générique $u(t,z) = 1/ch(t)$.
Tout le long de la fibre, le signal est périodiquement amplifié et chaque opération d'amplification entraine l'ajout d'un bruit parasite de type bruit blanc.

Dans un premier temps, nous considérerons la transmission du signal sans amortissement : i.e. $\alpha = 0$. Nous obtenons alors l'équation de Schroedinger non linéaire classique (NLS):

$$i\frac{\partial u}{\partial z} + \frac{1}{2}\frac{\partial^2 u}{\partial t^2} + |u|^2 u = 0$$

Après l'étude du problème de Cauchy, nous présenterons les méthodes de discrétisation en temps et en espace et quelques résultats numériques. Dans un deuxième temps, nous nous placerons dans le cas de l'équation de Schroedinger non linéaire faiblement amortie (WDNLS) avec amplification périodique du signal. Enfin, dans un troisième et dernier temps, nous considèrerons l'amplification bruitée du signal.
L'équation de Schroedinger non linéaire (NLS) a fait l'objet de nombreuses études numériques afin de construire et de comparer différentes méthodes conservant ou non les

invariants de celle-ci (masse, énergie) ([28, 29, 40, 46]).
L'équation de Schroedinger non linéaire faiblement amortie (WDNLS), quant à elle, n'a fait l'objet que de peu d'études ([37]).

2.2 Présentation du problème.

On considère l'équation de Schrödinger non linéaire classique

$$i\frac{\partial u}{\partial z}+\frac{1}{2}\frac{\partial^2 u}{\partial t^2}+|u|^2\,u=0\quad t\in\mathbb{R}\,,\qquad z\geqslant 0 \tag{2.2}$$

munie de la condition initiale

$$u(t,0)=u_0(t)\quad t\in\mathbb{R} \tag{2.3}$$

t la variable temporelle, $t\in(0,T)$, avec $T=1024$;
z la variable spatiale, $z\in(0,13)$

Le signal est localisé en temps comme une succession finie d'informations binaires élémentaires (i.e. une succession de 0 et de 1). Nous pouvons alors nous restreindre à un intervalle borné de $\mathbb{R}$, ici $t\in(0,T)$ dont le choix sera justifié plus loin.
D'autre part, le signal étant nul aux limites de cet intervalle, par convenance pour la discrétisation en temps, nous prendrons des conditions limites périodiques :

$$u(t+T,z)=u(t,z),\qquad \forall t\in(0,T)$$

On note $u(t,0)$ la donnée initiale, i.e. le signal à transmettre.

2.3 Résultats théoriques.

Dans cette section, nous présentons deux quantités conservées relatives aux solutions de l'équation (NLS) avant d'énoncer les résultats d'existence, d'unicité et de régularité.

2.3.1 Invariants.

L'invariance mène aux lois de conservation.
La plus importante étant la conservation de la masse

$$\frac{d}{dz}\left(|u|_{L^2}\right)^2=\frac{d}{dz}\left\{\int_{\mathbb{R}}|u|^2\ \mathrm{d}t\right\}=0 \tag{2.4}$$

Nous considérons un autre invariant : la conservation de l'énergie.
Notons $E(u)$ l'énergie

$$E(u)=\left(\left|\frac{\partial u}{\partial t}\right|_{L^2}\right)^2-\left(|u|_{L^4}\right)^4=\int_{\mathbb{R}}\left|\frac{\partial u}{\partial t}\right|^2\ \mathrm{d}t-\int_{\mathbb{R}}|u|^4\ \mathrm{d}t \tag{2.5}$$

Nous avons alors

$$\frac{d}{dz}E(u)=0 \tag{2.6}$$

2.3.2 Conservation des invariants.

Nous sommes dans le cas d'un domaine borné de $\mathbb{R}$, de la forme $(0, T)$ avec la solution u périodique sur ce domaine.

<u>Conservation de la masse :</u>
On veut montrer que

$$\frac{d}{dz}\int_0^T |u|^2 \, \mathrm{d}t = 0$$

Pour cela, on multiplie (NLS) par $-i\overline{u}$ et on intègre sur $(0,\, T)$:

$$\int_0^T \frac{\partial u}{\partial z}\overline{u} \, \mathrm{d}t = \frac{i}{2}\int_0^T \frac{\partial^2 u}{\partial t^2}\overline{u} \, \mathrm{d}t + i\int_0^T |u|^2 u\overline{u} \, \mathrm{d}t$$

et $\displaystyle\int_0^T |u|^2 u\overline{u} \, \mathrm{d}t = \int_0^T |u|^4 \, \mathrm{d}t$
Le terme de bord des intégrations par parties disparait, d'où :

$$\int_0^T \frac{\partial^2 u}{\partial t^2}\overline{u} \, \mathrm{d}t = \left[\frac{\partial u}{\partial t}\overline{u}\right]_0^T - \int_0^T \frac{\partial u}{\partial t}\overline{\frac{\partial u}{\partial t}} \, \mathrm{d}t = -\int_0^T \left|\frac{\partial u}{\partial t}\right|^2 \, \mathrm{d}t$$

Ce qui entraine :

$$\int_0^T \frac{\partial u}{\partial z}\overline{u} \, \mathrm{d}t = -\frac{i}{2}\int_0^T \left|\frac{\partial u}{\partial t}\right|^2 \, \mathrm{d}t + i\int_0^T |u|^4 \, \mathrm{d}t$$

On prend la partie réelle de cette dernière égalité :

$$\Re\left(\int_0^T \frac{\partial u}{\partial z}\overline{u} \, \mathrm{d}t\right) = 0$$

lemme 1 *Pour v une fonction régulière, nous avons :*

$$\frac{d}{dz}|v|^2 = 2\Re\left(v\overline{\frac{\partial v}{\partial z}}\right).$$

Or, par ce lemme avec $v = u$, on en déduit que

$$2\Re\left(\int_0^T \frac{\partial u}{\partial z}\overline{u} \, \mathrm{d}t\right) = \frac{d}{dz}\left(\int_0^T |u|^2 \, \mathrm{d}t\right)$$

d'où

$$\frac{d}{dz}\left(|u|_{L^2}\right)^2 = \frac{d}{dz}\int_0^T |u|^2 \, \mathrm{d}t = 0$$

<u>Conservation de l'énergie :</u>
On veut montrer que

$$\frac{d}{dz}\left(\int_0^T \left|\frac{\partial u}{\partial t}\right|^2 \, \mathrm{d}t\right) - \frac{d}{dz}\left(\int_0^T |u|^4 \, \mathrm{d}t\right) = 0$$

On multiplie (NLS) par $\overline{\frac{\partial u}{\partial z}}$:

$$i\frac{\partial u}{\partial z}\overline{\frac{\partial u}{\partial z}} = -\frac{1}{2}\frac{\partial^2 u}{\partial t^2}\overline{\frac{\partial u}{\partial z}} - |u|^2 u\overline{\frac{\partial u}{\partial z}}$$

On prend la partie réelle de cette relation et on intègre en temps :

$$\frac{1}{2}\int_0^T \Re\left(\frac{\partial^2 u}{\partial t^2}\overline{\frac{\partial u}{\partial z}}\right)\, \mathrm{d}t + \int_0^T \Re\left(|u|^2 u\overline{\frac{\partial u}{\partial z}}\right)\, \mathrm{d}t = 0$$

Le terme de bord disparait, d'où :

$$\Re\left(\int_0^T \frac{\partial^2 u}{\partial t^2}\overline{\frac{\partial u}{\partial z}}\, \mathrm{d}t\right) = -\Re\left(\int_0^T \frac{\partial u}{\partial t}\overline{\frac{\partial^2 u}{\partial t\partial z}}\, \mathrm{d}t\right)$$

De là,

$$-\frac{1}{2}\Re\left(\int_0^T \frac{\partial u}{\partial t}\overline{\frac{\partial^2 u}{\partial t\partial z}}\, \mathrm{d}t\right) = -\frac{1}{4}\frac{d}{dz}\left(\int_0^T \left|\frac{\partial u}{\partial t}\right|^2\, \mathrm{d}t\right)$$

par le lemme précédent avec $v = \frac{\partial u}{\partial t}$.

De plus,

$$|u|^2\Re\left(u\overline{\frac{\partial u}{\partial z}}\right) = \frac{1}{2}|u|^2\frac{d}{dz}|u|^2 = \frac{1}{2}\frac{d}{dz}\left(\frac{(|u|^2)^2}{2}\right) = \frac{1}{4}\frac{d}{dz}|u|^4$$

Finalement,

$$-\frac{1}{4}\frac{d}{dz}\left(\int_0^T \left|\frac{\partial u}{\partial t}\right|^2\, \mathrm{d}t\right) + \frac{1}{4}\int_0^T \frac{d}{dz}|u|^4\, \mathrm{d}t = 0$$

i.e.

$$\frac{d}{dz}\left(\int_0^T \left|\frac{\partial u}{\partial t}\right|^2\, \mathrm{d}t\right) - \frac{d}{dz}\left(\int_0^T |u|^4\, \mathrm{d}t\right) = 0$$

2.3.3 Problème de Cauchy.

Nous rappelons ici les résultats d'existence, d'unicité et de régularité des solutions pour l'équation (NLS) classique.
Pour plus de détails, on peut se référer à ([44]).
Le problème

$$i\frac{\partial u}{\partial z} + \frac{1}{2}\frac{\partial^2 u}{\partial t^2} + |\,u\,|^2\, u = 0 \quad t \in \mathbb{R}\,, \quad z \geqslant 0 \tag{2.7}$$

muni de la condition initiale

$$u(t,0) = u_0(t) \quad t \in \mathbb{R} \tag{2.8}$$

($u(t,z)$ est une fonction à valeurs complexes et α un scalaire réel) se ramène, au problème plus général

$$i\frac{\partial u}{\partial z} - \frac{\partial^2 u}{\partial t^2} + f(u) = 0 \quad t \in \mathbb{R}\,, \quad z \geqslant 0 \tag{2.9}$$

quitte à résoudre l'équation conjuguée de (2.7) multipliée par -1.
f vérifie les hypothèses suivantes :

$$f: \mathbb{C} \to \mathbb{C} \quad ; \quad f(0) = 0$$

$$f \text{ s'écrit } f(u) = g(|u|^2)u \text{ où } g \text{ est une fonction réelle.}$$

Nous allons énoncer des résultats d'existence, d'unicité et de régularité pour le problème (2.9) et vérifier que leurs hypothèses sont satisfaites par le problème initial (2.7) - (2.8).

Soit $F(u) = G(|u|^2)/2$ et G la primitive de g pour laquelle $G(0) = 0$.
Nous notons $E(u(z))$ l'énergie, définie par

$$E(u(z)) = \int_{\mathbb{R}} \left\{ \frac{1}{2} \left| \frac{\partial u}{\partial t} \right|^2 + F(u) \right\} \, \mathrm{d}t \tag{2.10}$$

Commençons par les solutions faibles.

Théorème *2.4*

Supposons

$$F(u) \geqslant -c_1 |u|^2 - c_2 |u|^{q+1} \quad \textit{pour } q,\ 1 < q < 5 \quad \textit{et des constantes } c_1, c_2 \tag{2.11}$$

et

$$\frac{|F(u)|}{|f(u)|} \longrightarrow +\infty \quad \textit{quand} \quad |u| \to +\infty \tag{2.12}$$

Alors, pour toute donnée initiale vérifiant $E(u(0)) < +\infty$ et $u(0) \in L^2$, il existe une solution faiblement continue $u : \mathbb{R} \to H^1$ telle que $E(u(z)) \leqslant E(u(0))$ pour $z \in \mathbb{R}$.

Remarque *2*

(2.11) est une condition de répulsivité de la nonlinéarité.
(2.12) impose à F une croissance moindre que l'exponentielle lorsque $|u| \to +\infty$.

Pour $f(u) = g(|u|^2)u = -|u|^2 u$, nous obtenons $g(s) = -s$, d'où $G(s) = -\dfrac{s^2}{2}$ et $F(u) = -\dfrac{|u|^4}{4}$. Finalement

$$E(u(z)) = \int \left\{ \frac{1}{2} \left| \frac{\partial u}{\partial t} \right|^2 - \frac{|u|^4}{4} \right\} \, \mathrm{d}t \tag{2.13}$$

La condition (2.11) est vérifiée pour $q = 3$ et $c_1 = 0, c_2 \geqslant \frac{1}{4}$.

$$\frac{|F(u)|}{|f(u)|} = \frac{|u|}{4} \longrightarrow +\infty \quad \text{quand} \quad |u| \to +\infty$$

La condition (2.12) est aussi satisfaite pour le problème initial (2.7) - (2.8).

Nous avons le résultat suivant pour les solutions fortes.

Théorème *2.5*

Supposons (2.11) et que f soit une fonction de classe C^1 telle que

$$|f'(u)| \leqslant c_3 \left(1 + |u|^{p-1}\right) \quad \text{où } c_3 > 0,\ 1 < p < +\infty \tag{2.14}$$

Alors la solution du théorème (4) est unique. $u : \mathbb{R} \to H^1$ est une fonction fortement continue qui satisfait $E(u(z)) = E(u(0))$ pour $z \in \mathbb{R}$.
De plus, nous avons $|u(z)|_{L^2} = |u(0)|_{L^2}$ pour $z \in \mathbb{R}$.

Remarque *3*

La notation f, fonction de classe C^1, ne doit pas être vue dans $\mathbb{C}$ mais dans $\mathbb{R}^2$ en écrivant $\mathbb{C} \approx \mathbb{R}^2$. f est alors une fonction vectorielle, à deux variables dérivable.

Nous obtenons la majoration (2.14) pour $c_3 = 1$ et $p = 3$.
Enfin, voici un résultat de régularité pour les solutions.

Théorème *2.6*

Supposons que f vérifie les mêmes hypothèses que pour le théorème (5).
Si la donnée initiale appartient à H^2 alors l'unique solution est une fonction fortement continue à valeurs dans H^2.

2.4 Discrétisation du problème.

Contrairement à la notation usuelle en théorie des équations aux dérivées partielles, t est la variable de "type spatial" et z la variable de "type temporel". On fera donc attention lorsque l'on parlera de schéma en temps ou en espace.
La condition initiale est de la forme suivante

$$u(t, z=0) = \mathcal{K} \sum_{j=0}^{127} \frac{a_j}{ch(t - 8j - 4)} \tag{2.15}$$

$a_j = 0$ ou 1 avec une probabilité $1/2$ et $\mathcal{K}$ une constante réelle.
La séquence des a_j est utilisée pour coder de l'information numérique binaire ($a_j = 1$ ou $a_j = 0$). Un tel symbole est émis dans chaque intervalle de temps de longueur $T_b = 8$ dans notre exemple (d'où $T = 8.128 = 1024$, la dimension de notre domaine temporel). A chacun de ces symboles est associé un soliton, dont le profil est conservé durant la propagation.

2.4.1 Discrétisation en temps.

Puisque le signal est nul aux frontières du domaine temporel, par convenance nous considérons le domaine $t \in (0, T)$ avec des conditions limites périodiques. En effet, on choisira toujours $a_0 = 0$ et $a_{127} = 0$.
Nous écrivons alors naturellement la solution comme une série de Fourier

$$u(t,z) = \sum_{k\in\mathbb{Z}} \hat{u}_k(z)\Phi_k(t) \tag{2.16}$$

où

$$\Phi_k(t) = exp\left(ik\frac{2\pi}{T}t\right)$$

Pour $N > 0$, soient $j \in [\![0,\ N-1]\!]$ et $t_j = \dfrac{jT}{N}$. La solution u_N est représentée par ses valeurs aux noeuds t_j de la grille. Pour la méthode pseudo-spectrale Fourier-Collocation ([13, 26]), nous demandons que l'équation (2.7) soit satisfaite en ces points, i.e.

$$\left(i\frac{\partial u_N}{\partial z} + \frac{1}{2}\frac{\partial^2 u_N}{\partial t^2} + |\,u_N\,|^2\, u_N\right)_{t=t_j} = 0 \qquad \forall j \in [\![0,\ N-1]\!] \tag{2.17}$$

Et les conditions initiales sont données par

$$u_N(t_j, 0) = u_0(t_j) \qquad \forall j \in [\![0,\ N-1]\!]$$

N est le nombre de points et nous verrons ultérieurement qu'il devra être pris très grand : $N = 32768$.

Montrons que le problème semi-discrétisé en temps est bien posé.
Par la suite nous considérerons cette équation faiblement amortie, i.e.

$$i\frac{\partial u}{\partial z} + \frac{1}{2}\frac{\partial^2 u}{\partial t^2} + |\,u\,|^2\, u + i\alpha u = 0$$

avec α un scalaire réel positif. Nous traitons ainsi ici les cas où α est positif ou nul. Nous rappelons les notations suivantes :

$$(u,v) = \int_0^{2\pi} u(t)\,\overline{v(t)}\,\mathrm{d}t$$

et

$$(u,v)_N = \frac{2\pi}{N}\sum_{j=0}^{N-1} u(t_j)\,\overline{v(t_j)}$$

Pour u et v des fonctions de $S_N = span\{\Phi_k(t), k \in \mathbb{I}_N\}$, nous avons le résultat :

$$(u,v) = (u,v)_N$$

Nous multiplions l'équation (2.17) par $u_N(t_j)$ et faisons la sommation pour j allant de 0 à $N-1$:

$$\frac{2\pi}{N}\sum_{j=0}^{N-1}\left\{i\frac{\partial u_N(t_j)}{\partial z}\,\overline{u_N(t_j)} + \frac{1}{2}\frac{\partial^2 u_N(t_j)}{\partial t^2}\,\overline{u_N(t_j)} + |\,u_N(t_j)\,|^2\, u_N(t_j)\,\overline{u_N(t_j)} + i\alpha\, u_N(t_j)\,\overline{u_N(t_j)}\right\} = 0$$

$$\begin{aligned}
&\Rightarrow\ i\left(\frac{\partial u_N}{\partial z},u_N\right)_N+\frac{1}{2}\left(\frac{\partial^2 u_N}{\partial t^2},u_N\right)_N+\left(|\,u_N\,|^2\,u_N,u_N\right)_N+i\alpha\left(u_N,u_N\right)_N=0\\
&\Rightarrow\ \left(\frac{\partial u_N}{\partial z},u_N\right)-\frac{i}{2}\left(\frac{\partial^2 u_N}{\partial t^2},u_N\right)-i\left(|\,u_N\,|^2\,u_N,u_N\right)_N+\alpha\,|\,u_N\,|^2_{L^2(0,2\pi)}=0
\end{aligned}$$

Nous calculons les termes suivants :

$$\left(\frac{\partial^2 u_N}{\partial t^2},u_N\right)=\int_0^T\frac{\partial^2 u}{\partial t^2}\overline{u}\ \mathrm{d}t=\left[\frac{\partial u}{\partial t}\overline{u}\right]_0^T-\int_0^T\frac{\partial u}{\partial t}\frac{\overline{\partial u}}{\partial t}\ \mathrm{d}t=-\int_0^T\left|\frac{\partial u}{\partial t}\right|^2\ \mathrm{d}t\ \in\mathbb{R}$$

$$\left(|\,u_N\,|^2\,u_N,u_N\right)_N=\frac{2\pi}{N}\sum_{j=0}^{N-1}|\,u_N(t_j)\,|^4\in\mathbb{R}$$

Pour appliquer le lemme précédent, nous prenons la partie réelle de cette relation.

$$\begin{aligned}
\Re\left(\frac{\partial u_N}{\partial z},u_N\right)\ &+\ \alpha\,|\,u_N\,|^2_{L^2(0,2\pi)}=0\\
\Rightarrow\frac{1}{2}\frac{d}{dz}\,|\,u_N\,|^2_{L^2(0,2\pi)}\ &+\ \alpha\,|\,u_N\,|^2_{L^2(0,2\pi)}=0\\
\Rightarrow\frac{d}{dz}\,|\,u_N\,|^2_{L^2(0,2\pi)}\ &=\ -2\alpha\,|\,u_N\,|^2_{L^2(0,2\pi)}\leqslant 0
\end{aligned}$$

Donc le problème semi-discrétisé en temps est bien posé.

Dans notre premier exemple, un test avec une solution exacte, le nombre de points est trop important. En effet la précision spectrale ne nécessite pas un tel nombre de points, loin de là.
Par contre, par la suite, le nombre de points est nécessaire $N = 32768$. Ainsi il nous semble important d'en tenir compte dès le début, pour pouvoir tester les méthodes dans les conditions réelles d'emploi.
La discrétisation de l'opérateur de dérivation double temporelle $\frac{\partial^2 u}{\partial t^2}$ dans l'espace physique est une matrice pleine de dimension NxN. Par contre elle est diagonale dans l'espace spectral. L'évaluation de termes de la forme $\frac{\partial^2}{\partial t^2}$ sera donc effectuée dans l'espace spectral par le biais de Transformées de Fourier Rapides (FFT),voir [16] et [9]. Le nombre de points, $N = 2^{15}$, permettra de la faire de manière optimale.

2.4.2 Discrétisation en espace.

L'équation (NLS) a la particularité de posséder une infinité de lois de conservation. Nous en avons énoncer deux, précédemment, parmi les plus importantes : la conservation de la masse et la conservation de l'énergie.
Il nous semble important d'avoir des schémas d'avancement en espace qui préservent ces invariants.
D'autre part, de part la nature stochastique de la solution et par conséquent du grand nombre de points nécessaire pour la discrétisation temporelle, nous choisissons les méthodes *Split Step classique* et *Split Step Agrawal*. Ces deux méthodes (dont la seconde est une version plus précise de la première) sont basées sur le splitting d'opérateurs et sont

totalement explicites.

Nous écrivons l'équation (NLS) classique sous la forme :

$$\frac{\partial u}{\partial z} = \mathcal{L}u + i\,\mathcal{N}(u)u \tag{2.18}$$

où les opérateurs $\mathcal{L}$ et $\mathcal{N}$ sont définis par :

$$\begin{aligned} \mathcal{L}u &= \frac{i}{2}\frac{\partial^2 u}{\partial t^2} \\ \mathcal{N}(u) &= |\,u\,|^2 \end{aligned}$$

$\mathcal{L}$ est l'opérateur différentiel qui prend en compte la dispersion et l'absorption dans un milieu linéaire.
$\mathcal{N}$ est un opérateur non linéaire qui gouverne les effets des non linéarités de la fibre optique sur la propagation.
En général, la dispersion et la non linéarité agissent de pair tout le long de la fibre. La méthode *Split Step* obtient une approximation de la solution en supposant que, dans la propagation du signal sur une petite distance h, les effets de la dispersion et de la non linéarité peuvent être considérés comme agissant indépendemment. Plus précisement, la propagation entre z et $z+h$ est effectuée en deux étapes. Dans un premier temps, la non linéarité agit seule et $\mathcal{L} \equiv 0$ dans (2.18). Dans un second temps, la dispersion agit seule et $\mathcal{N} \equiv 0$ dans (2.18).
Nous écrivons cela

$$u(t, z+h) = exp\,[h\,\mathcal{L}]\,exp\,[ih\,\mathcal{N}(u(z))]\,u(t,z) \tag{2.19}$$

Pour estimer la précision de la méthode *Split Step*, nous remarquons qu'une solution exacte formelle de (2.18) est donnée par

$$u(t, z+h) = exp\,[h\,(\,\mathcal{L} + i\,\mathcal{N}(u))]\,u(t,z) \tag{2.20}$$

si $\mathcal{N}$ est supposé indépendant de z.
Nous rappelons maintenant la formule de Baker-Hausdorff pour deux opérateurs non commutatifs a et b :

lemme 2 *Formule de Baker-Hausdorff*
Soient a et b deux opérateurs non commutatifs, nous avons :

$$exp(a)exp(b) = exp\left\{a + b + \frac{1}{2}\,[a,b] + \frac{1}{12}\,[a-b,[a,b]] + ...\right\} \tag{2.21}$$

où $[a,b] = ab - ba$.

Une comparaison des équations (2.19) et (2.20) montre que la méthode *Split Step* ignore la non-commutativité des opérateurs $\mathcal{L}$ et $\mathcal{N}$. En utilisant la formule (2.21) avec $a = h\,\mathcal{L}$ et $b = ih\,\mathcal{N}$, le terme dominant de l'erreur provient du commutateur $\frac{1}{2}h^2\,[\mathcal{L},\,\mathcal{N}]$. De là, la méthode *Split Step* est précise à l'ordre 2 en h, soit un schéma d'ordre 1.

La précision de la méthode *Split Step* peut être améliorée en adoptant une procédure différente pour propager le signal sur un segment de z à $(z+h)$.
Dans cette procédure, l'équation (2.19) est remplacée par

$$u(t,z+h) = exp\left(\frac{h}{2}\,\mathcal{L}\right) exp\left[\int_z^{z+h} i\,\mathcal{N}(u(z'))\,\mathrm{d}z'\right] exp\left(\frac{h}{2}\,\mathcal{L}\right) u(t,z) \tag{2.22}$$

La différence principale réside dans le fait que l'effet de la non linéarité est incluse au milieu du segment plutôt qu'au bord de celui-ci.
De part la forme symétrique de l'exponentielle des opérateurs dans (2.22), cette méthode est connue comme la méthode *Split Step* symétrisée ([24]). Il est important que l'intégrale de l'exponentielle du milieu tienne compte de la dépendance en z de l'opérateur non linéaire $\mathcal{N}$. Si le pas d'espace h est suffisamment petit, l'intégrale peut être approchée par $exp\,[ih\ \mathcal{N}(u(z))]$, tout comme pour (2.19).
L'avantage le plus important d'utiliser la forme symétrique de (2.22) est que le terme dominant pour l'erreur provient du double commutateur dans la formule (2.21) et est à l'ordre 3 en h, soit une méthode d'ordre 2. Cela se vérifie en appliquant deux fois (2.21) dans (2.22) :

$$exp\left(\frac{h}{2}\,\mathcal{L}\right) exp\,[ih\ \mathcal{N}(u(z))]\, exp\left(\frac{h}{2}\,\mathcal{L}\right) = exp\left\{h\,\mathcal{L} + ih\,\mathcal{N} - i\frac{h^3}{2}\left[\,\mathcal{L}^2\,\mathcal{N} - \ \mathcal{N}\,\mathcal{L}^2\right] + ...\right\}$$

Nous obtenons ainsi la méthode *Split Step classique*, ([43]).

La précision de la méthode *Split Step classique* peut être encore améliorée en évaluant l'intégrale de l'équation (2.22) plus finement que par la valeur approchée $h\,\mathcal{N}(u(z))$.
Une approche simple consiste à appliquer la méthode des trapèzes et ainsi approcher l'intégrale par

$$\int_z^{z+h} i\,\mathcal{N}(u(z'))\,\mathrm{d}z' = \frac{ih}{2}\left\{\,\mathcal{N}(u(z)) + \ \mathcal{N}(u(z+h))\right\} \tag{2.23}$$

Cependant l'application de (2.23) n'est pas simple puisque $\mathcal{N}(u(z+h))$ est inconnue à la moitié du segment, en $\left(z+\frac{h}{2}\right)$. Pour remédier à cela nous appliquons alors le schéma bien connu de Heun (Runge-Kutta d'ordre 2 explicite) :
Intégrer l'équation différentielle $\frac{dy}{dz} = f(y,z)$ sur $[z, z+h]$ s'écrit

$$y(z+h) = y(z) + \frac{h}{2}\left\{f(y,z) + f(y+hf(y,z), z+h)\right\}$$

Cela revient à calculer une valeur approchée de y en $(z+h)$: $\tilde{y}(z+h)$, puis à faire la moyenne des contributions :

$$y(z+h) = y(z) + \frac{h}{2}\left\{f(y,z) + f(\tilde{y}, z+h)\right\}$$

Nous obtenons ainsi la méthode *Split Step Agrawal* (voir [37], [2]).

La mise en œuvre de ces méthodes est directe.
La longueur de la fibre est divisée en un grand nombre de segments qui ne sont pas nécessairement de même longueur. Le signal est propagé de segment en segment en utilisant l'équation (2.22).

Remarque 4

La méthode Split Step classique nécessite 4 FFT alors que la méthode Split Step Agrawal en utilise 6 (on peut réduire le coût théorique de 8 FFT de 2 unités en conservant les valeurs du signal après application de la première moitié de l'opérateur dispersif).

2.4.3 Discrétisation complète.

On note $u_0(t_j)$ les valeurs nodales de la condition initiale et $u_N(t_j, z^n)$ les valeurs nodales de la solution approchée en $t = j\Delta t$ et en $z = z^n = nh$ avec Δt le pas de temps et h le pas d'espace.

Méthode *Split-Step classique.*

On suppose que l'on dispose du vecteur $U_N^n = \{u_N(t_j, z^n)\}_j$.

$$U_N^{n+1} = \mathcal{F}^{-1} exp\left(\frac{h}{2}\widehat{\mathcal{L}}\right) \mathcal{F} \exp\left[ih\,|U_N^n|^2\right] \mathcal{F}^{-1} \exp\left(\frac{h}{2}\widehat{\mathcal{L}}\right) \mathcal{F}\, U_N^n$$

où $\widehat{\mathcal{L}}$ désigne l'opérateur de dérivation double dans l'espace de Fourier et $\mathcal{F}$ et $\mathcal{F}^{-1}$ désignent les transformées de Fourier directe et inverse.
Le coût du passage de U_N^n à U_N^{n+1} est en 4 $\mathrm{O}(N\log_2 N)$ + 6 $\mathrm{O}(N)$.

Méthode *Split-Step Agrawal.*

On suppose que l'on dispose du vecteur $U_N^n = \{u_N(t_j, z^n)\}_j$.
On calcule une valeur intermédiaire $\tilde{U}_N^{n+1}$ par la méthode *Split-Step classique*

$$\tilde{U}_N^{n+1} = \mathcal{F}^{-1} \exp\left(\frac{h}{2}\widehat{\mathcal{L}}\right) \mathcal{F} \exp\left[ih\,|U_N^n|^2\right] \mathcal{F}^{-1} \exp\left(\frac{h}{2}\widehat{\mathcal{L}}\right) \mathcal{F}\, U_N^n$$

pour construire U_N^{n+1} :

$$U_N^{n+1} = \mathcal{F}^{-1} \exp\left(\frac{h}{2}\widehat{\mathcal{L}}\right) \mathcal{F} \exp\left\{\frac{ih}{2}\left[|U_N^n|^2 + \left|\tilde{U}_N^{n+1}\right|^2\right]\right\} \mathcal{F}^{-1} \exp\left(\frac{h}{2}\widehat{\mathcal{L}}\right) \mathcal{F}\, U_N^n$$

Le coût du passage de U_N^n à U_N^{n+1} est en 6 $\mathrm{O}(N\log_2 N)$ + 10 $\mathrm{O}(N)$.

D'autre part, pour ces deux discrétisations, la programmation de ces algorithmes peut être grandement optimisée si l'on dispose de calculateurs vectoriels (longues boucles vectorisables) ou alors parallèles pour les FFT (bien que n'étant que monodimensionnelles, on peut les décomposer en paquets).

2.4.4 Stabilité des schémas.

En plus d'être explicites, ces deux schémas en espace sont inconditionnellement stables. Montrons le pour la méthode *Split-Step Agrawal.* Il suffit d'obtenir :

$$\left|U_N^{n+1}\right|_{L^2} = |U_N^n|_{L^2} \quad \forall n \in \mathbb{N}$$

Pour simplifier nous introduisons les notations suivantes, seulement pour cette section :

$$\begin{array}{c} u_N(t_j, z^n) \text{ sera noté } u_j^n \\ v_N(t_j, z^n) \text{ sera noté } v_j^n \\ w_N(t_j, z^n) \text{ sera noté } w_j^n \end{array}$$

Ecrivons le schéma sous la forme

$$\begin{aligned} V_N^n &= \mathcal{F}^{-1} \exp\left(\frac{h}{2}\widehat{\mathcal{L}}\right)\, \mathcal{F}\, U_N^n \\ W_N^n &= \exp\left\{\frac{ih}{2}\left[|U_N^n|^2 + \left|\tilde{U}_N^{n+1}\right|^2\right]\right\} V_N^n \\ U_N^{n+1} &= \mathcal{F}^{-1} \exp\left(\frac{h}{2}\widehat{\mathcal{L}}\right)\, \mathcal{F}\, W_N^n \end{aligned}$$

V_N^n et W_N^n ne sont pas des solutions intermédiaires mais seulement des valeurs intermédiaires. Nous allons successivement montrer que

$$\begin{aligned} \text{(i)} \quad \sum_j |u_j^n|^2 &= \sum_j |v_j^n|^2 \\ \text{(ii)} \quad \sum_j |v_j^n|^2 &= \sum_j |w_j^n|^2 \\ \text{(iii)} \quad \sum_j |w_j^n|^2 &= \sum_j |u_j^{n+1}|^2 \end{aligned}$$

Pour (i), nous employons l'identité de Parseval discrète :

$$\frac{1}{N}\sum_j |u_j^n|^2 = \sum_k |\hat{u}_k^n|^2 \tag{2.24}$$

De là le spectre de Fourier de V_N^n s'écrit

$$\hat{v}_k^n = \exp\left\{-\frac{ih}{2}\left(\frac{2\pi k}{L}\right)^2\right\} \hat{u}_k^n \quad ; \quad \forall k \in \left[\!\!\left[1 - \frac{N}{2},\, \frac{N}{2}\right]\!\!\right]$$

soit en prenant le module

$$|\hat{v}_k^n| = |\hat{u}_k^n| \quad \forall k \in \left[\!\!\left[1 - \frac{N}{2},\, \frac{N}{2}\right]\!\!\right]$$

D'où la première égalité (i) par une seconde application de (2.24).
Comme

$$w_j^n = \exp\left\{\frac{ih}{2}\left[\left|u_j^n\right|^2 + \left|\tilde{u}_j^{n+1}\right|^2\right]\right\} v_j^n \quad \forall j \in [\![0,\, N-1]\!]$$

alors en prenant le module et sommant sur j on obtient la seconde partie (ii).
Enfin la dernière, (iii), s'obtient de manière analogue à (i).

On procède de même pour le schéma *Split Step classique.*

2.5 Conservation des invariants discrets.

En notant $U_N^n = \{u_N(t_j, z^n)\}_j$ les valeurs aux noeuds t_j de la solution approchée, u_N, les deux invariants discrets, analogues des invariants du problème continu, sont :

Conservation de la masse :

$$|u_N|_{L^2} = \text{ constante} \tag{2.25}$$

Par Parseval, cela s'écrit

$$|u_N|_{L^2} = \sqrt{T}\left(\sum_{k=-N/2}^{N/2-1} |\hat{u}_k|^2\right)^{1/2} \tag{2.26}$$

Conservation de l'énergie :

$$\left(\left|\frac{\partial u_N}{\partial t}\right|_{L^2}\right)^2 - \left(|u_N|_{L^4}\right)^4 = \text{ constante} \tag{2.27}$$

Par Parseval, nous avons :

$$\left|\frac{\partial u_N}{\partial t}\right|_{L^2} = \sqrt{T}\left(\sum_{k=-N/2}^{N/2-1} \left|\frac{2\pi}{T} k\hat{u}_k\right|^2\right)^{1/2} \tag{2.28}$$

Ne pouvant calculer exactement $\displaystyle\int_0^T |u_N(t,z)|^4 \ dt$, nous sommes obligés de l'approcher par quadrature.
Nous avons porté notre choix sur la méthode des trapèzes. La très grande régularité des solutions (la solution exacte est analytique avec des amplitudes raisonnables) permet de se limiter à cette méthode d'ordre 2 : des tests effectués avec des méthodes d'ordre 4 ou 6 ne donnant pas de résultats meilleurs, sachant que la solution est approchée à O (Δz^2) par les schémas en espace pour $\Delta z = 10^{-2}$ et 10^{-3}.
Ainsi

$$\left(|u_N|_{L^4}\right)^4 = \frac{\Delta t}{2}\left\{|u_N(t=0,z)|^4 + |u_N(t=T,z)|^4\right\} + \Delta t \sum_{j=1}^{N-1} |u_N(t=t_j,z)|^4 \tag{2.29}$$

Nous calculons les quantités des membres de gauche des relations (2.25) et (2.27) grâce à (2.26), (2.28) et (2.29) et regardons leur évolution en fonction de z.

2.6 Validation des méthodes.

L'équation (NLS) classique

$$i\frac{\partial u}{\partial z}+\frac{1}{2}\frac{\partial^2 u}{\partial t^2}+|u|^2\,u=0 \tag{2.30}$$

possède des solutions analytiques de la forme

$$v(t,z)=\frac{1}{ch(t)}e^{iz/2}$$

Nous choisissons ainsi

$$u_0(t)=u(t,z=0)=\sum_{j=0}^{127}\frac{a_j}{ch(t-8j-4)}\quad \text{pour } t\in(0,T) \tag{2.31}$$

avec $a_{65}=1$ et $a_j=0$ pour $j\in[\![0,\,127]\!]\setminus\{65\}$. Nous prenons $j=65$ pour être au milieu du domaine.
$u(t,z)=u(t,0)e^{iz/2}$ est une solution du problème (2.30)-(2.31).

2.6.1 Analyse des résultats.

Les calculs ont été réalisés avec différents pas de temps (i.e. différentes valeurs de N) et différents pas d'espace Δz :
$N \;=\; 32768,\ 16384,\ 8192,\ 4096,\ 2048;$
$\Delta z \;=\; 10^{-3},\ 10^{-2}.$
Par convenance pour les FFT, les valeurs de N sont uniquement des puissances de 2 (bien que cela puisse aussi s'appliquer avec des entiers N de la forme $N=2^p3^q5^r$).
Les résultats obtenus sont présentés dans les tableaux suivants.
Nous commençons par la méthode *Split Step classique*.

TAB. 2.1 – *Erreurs relatives obtenues avec la méthode Split Step classique.*

Δz	N	$\lvert.\rvert_{L^\infty}$ à $z=\Delta z$	$\lvert.\rvert_{L^\infty}$ à $z=13$	$\lvert.\rvert_{L^2}$ à $z=\Delta z$	$\lvert.\rvert_{L^2}$ à $z=13$
10^{-2}	32768	$0,9779.10^{-6}$	$0,2923.10^{-4}$	$0,5741.10^{-6}$	$0,2863.10^{-4}$
	16384	$0,9779.10^{-6}$	$0,2923.10^{-4}$	$0,5741.10^{-6}$	$0,2863.10^{-4}$
	8192	$0,9779.10^{-6}$	$0,2923.10^{-4}$	$0,5741.10^{-6}$	$0,2863.10^{-4}$
	4096	$0,9788.10^{-6}$	$0,2923.10^{-4}$	$0,5741.10^{-6}$	$0,2863.10^{-4}$
	2048	$0,2173.10^{-4}$	$0,3936.10^{-3}$	$0,1681.10^{-4}$	$0,3859.10^{-3}$
10^{-3}	32768	$0,9791.10^{-8}$	$0,2838.10^{-6}$	$0,5746.10^{-8}$	$0,2781.10^{-6}$
	16384	$0,9791.10^{-8}$	$0,2842.10^{-6}$	$0,5746.10^{-8}$	$0,2784.10^{-6}$
	8192	$0,9791.10^{-8}$	$0,2848.10^{-6}$	$0,5746.10^{-8}$	$0,2890.10^{-6}$
	4096	$0,1074.10^{-7}$	$0,2861.10^{-6}$	$0,5895.10^{-8}$	$0,2800.10^{-6}$
	2048	$0,2089.10^{-4}$	$0,3647.10^{-3}$	$0,1641.10^{-4}$	$0,3579.10^{-3}$

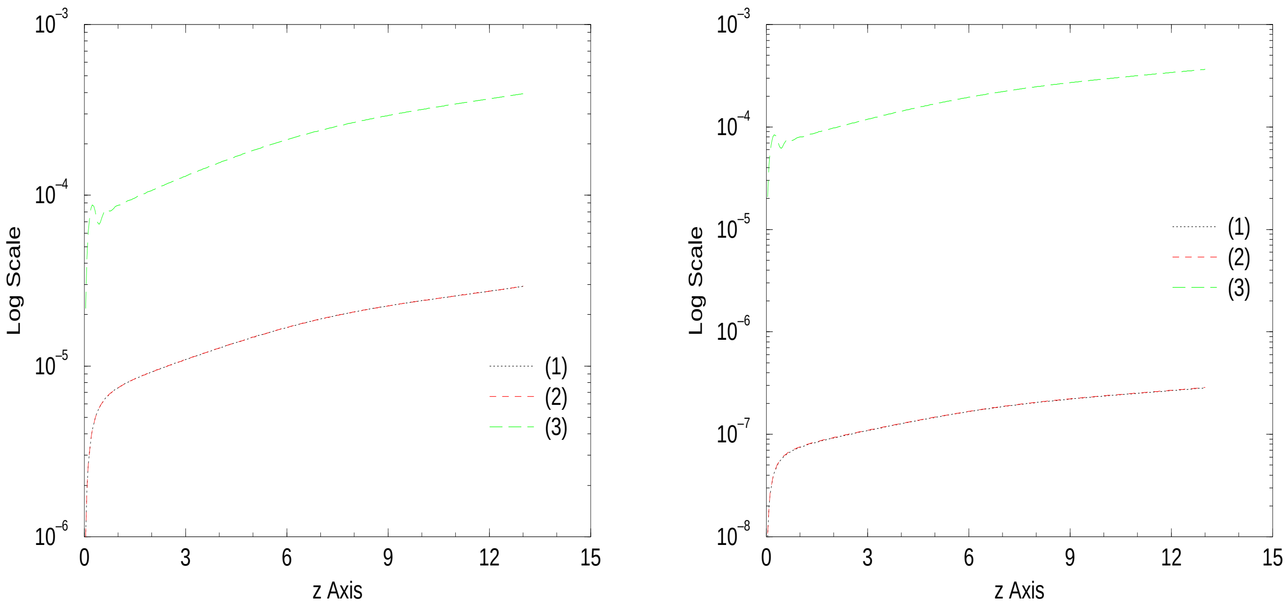
Pour la méthode *Split Step Agrawal*, nous obtenons

TAB. 2.2 – *Erreurs relatives obtenues avec la méthode Split Step Agrawal.*

Δz	N	$\lvert . \rvert_{L^\infty}$ à $z = \Delta z$	$\lvert . \rvert_{L^\infty}$ à $z = 13$	$\lvert . \rvert_{L^2}$ à $z = \Delta z$	$\lvert . \rvert_{L^2}$ à $z = 13$
10^{-2}	32768	$0,9800.10^{-6}$	$0,2926.10^{-4}$	$0,5749.10^{-6}$	$0,2866.10^{-4}$
	16384	$0,9800.10^{-6}$	$0,2926.10^{-4}$	$0,5749.10^{-6}$	$0,2866.10^{-4}$
	8192	$0,9800.10^{-6}$	$0,2926.10^{-4}$	$0,5749.10^{-6}$	$0,2866.10^{-4}$
	4096	$0,9809.10^{-6}$	$0,2926.10^{-4}$	$0,5749.10^{-6}$	$0,2866.10^{-4}$
	2048	$0,2178.10^{-4}$	$0,3938.10^{-3}$	$0,1684.10^{-4}$	$0,3862.10^{-3}$
10^{-3}	32768	$0,9793.10^{-8}$	$0,2838.10^{-6}$	$0,5747.10^{-8}$	$0,2781.10^{-6}$
	16384	$0,9793.10^{-8}$	$0,2843.10^{-6}$	$0,5747.10^{-8}$	$0,2785.10^{-6}$
	8192	$0,9793.10^{-8}$	$0,2848.10^{-6}$	$0,5747.10^{-8}$	$0,2890.10^{-6}$
	4096	$0,1074.10^{-7}$	$0,2861.10^{-6}$	$0,5896.10^{-8}$	$0,2800.10^{-6}$
	2048	$0,2090.10^{-4}$	$0,3647.10^{-3}$	$0,1642.10^{-4}$	$0,3579.10^{-3}$

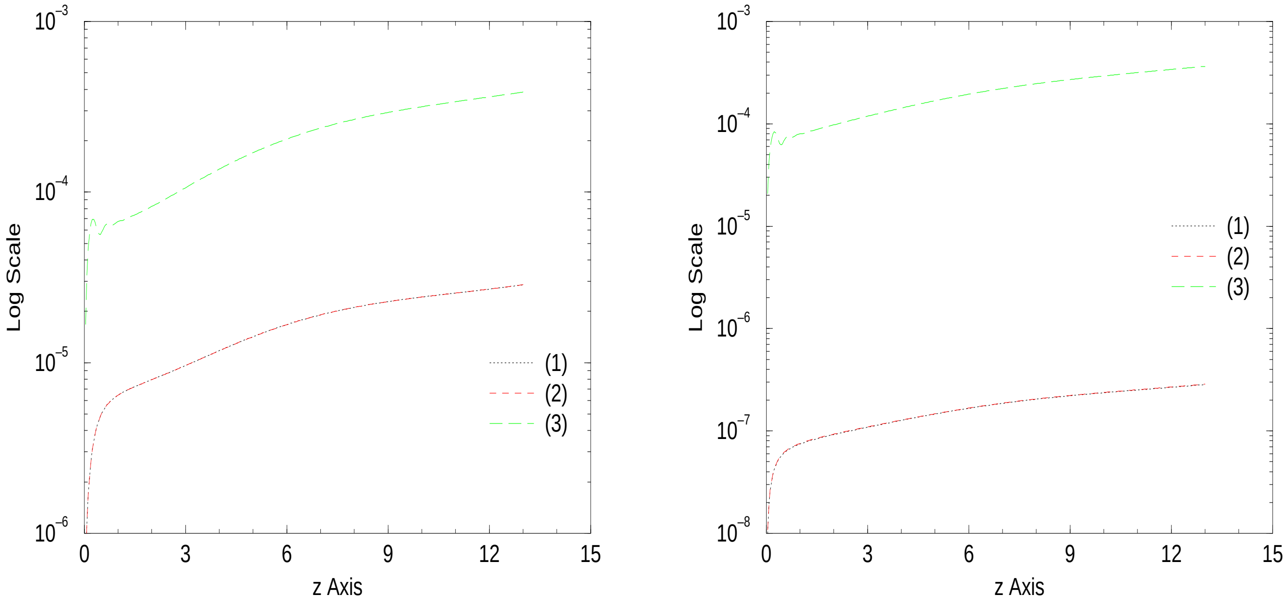
Le nombre de points, $N = 32768$, est trop grand puisque $N = 4096$ apporte la même précision mais il nous semblait important de s'assurer de la stabilité numérique des méthodes dans le cadre d'une solution exacte. D'autre part, nous vérifions numériquement que ces deux schémas sont d'ordre 2.

FIG. 2.1 – *Evolution en espace de l'erreur relative en norme L^∞ pour la méthode Split-Step Agrawal.*



Pour les figures (2.1) et (2.2) les légendes (1), (2) et (3) correspondent respectivement à

FIG. 2.2 – *Evolution en espace de l'erreur relative en norme* $\mathbf{L}^{\infty}$ *pour la méthode Split-Step classique.*



$N = 32768$, 4096 et 2048 points. A gauche nous avons les tracés pour $\Delta z = 10^{-2}$ et à droite pour $\Delta z = 10^{-3}$. Seules les courbes pour $N = 2048$ se distinguent des autres : il n'y a pas assez de modes pour une bonne résolution temporelle.

2.6.2 Conservation des invariants.

Nous représentons les valeurs des 2 invariants considéres avant et après la transmission pour les 2 méthodes considérées.
Pour mieux apprécier leur évolution, nous précisons les résultats (10 chiffres significatifs). Nous commençons par la méthode *Split Step classique.*

TAB. 2.3 – *Conservation des invariants avec la méthode Split Step classique.*

Δz	N	Masse à $z = 0$	Masse à $z = 13$	Energie à $z = 0$	Energie à $z = 13$
10^{-2}	32768	1,4142135624	1,4142135623	-1,3326822917	-1,3326944439
	16384	1,4142135624	1,4142135623	-1,3326822917	-1,3326944440
	8192	1,4142135624	1,4142135623	-1,3326822917	-1,3326944440
	4096	1,4142135624	1,4142135623	-1,3326822917	-1,3326944441
	2048	1,4142137117	1,4142137117	-1,3326936868	-1,3327851739
10^{-3}	32768	1,4142135624	1,4142135614	-1,3326822917	-1,3326824079
	16384	1,4142135624	1,4142135614	-1,3326822917	-1,3326824082
	8192	1,4142135624	1,4142135615	-1,3326822917	-1,3326824085
	4096	1,4142135624	1,4142135616	-1,3326822917	-1,3326824091
	2048	1,4142137117	1,4142137110	-1,3326936868	-1,3327731778

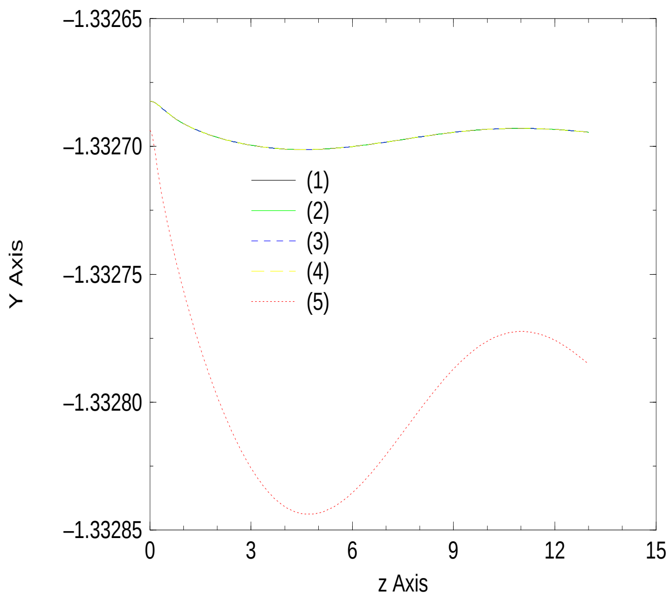
Les valeurs obtenues pour les invariants par la méthode *Split Step Agrawal* sont détaillées dans le tableau ci-dessous.

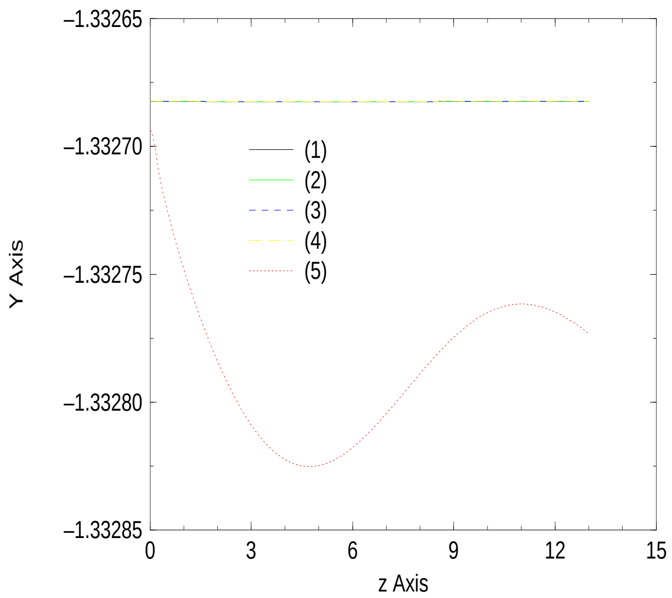
TAB. 2.4 – *Conservation des invariants avec la méthode Split Step Agrawal.*

Δz	N	Masse à $z=0$	Masse à $z=13$	Energie à $z=0$	Energie à $z=13$
10^{-2}	32768	1,4142135624	1,4142135623	-1,3326822917	-1,3326944596
	16384	1,4142135624	1,4142135623	-1,3326822917	-1,3326944596
	8192	1,4142135624	1,4142135623	-1,3326822917	-1,3326944596
	4096	1,4142135624	1,4142135623	-1,3326822917	-1,3326944597
	2048	1,4142137117	1,4142137117	-1,3326936868	-1,3327852893
10^{-3}	32768	1,4142135624	1,4142135614	-1,3326822917	-1,3326824079
	16384	1,4142135624	1,4142135614	-1,3326822917	-1,3326824082
	8192	1,4142135624	1,4142135615	-1,3326822917	-1,3326824085
	4096	1,4142135624	1,4142135616	-1,3326822917	-1,3326824092
	2048	1,4142137117	1,4142137110	-1,3326936868	-1,3327731880

Pour les figures (2.3) et (2.4) les légendes (1), (2), (3), (4) et (5) correspondent respectivement à $N = 32768$, 16384, 8192, 4096 et 2048 points.

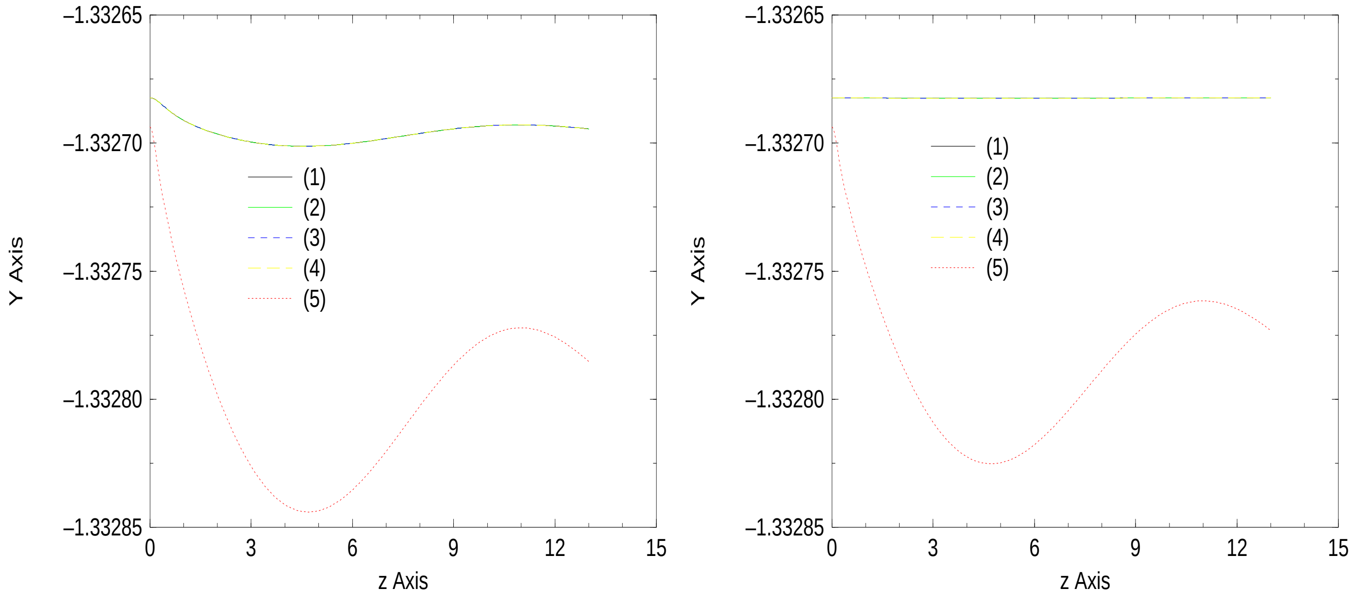
FIG. 2.3 – *Evolution en espace du second invariant $E(u)$ pour la méthode Split-Step classique.*





La masse est bien conservée par les deux méthodes puisque les variations sont de l'ordre de 10^{-8} à 10^{-9} avec une erreur en 10^{-4} à 10^{-6}. L'énergie, quant à elle, n'est préservée qu'à la précision des schémas.

FIG. 2.4 – *Evolution en espace du second invariant $E(u)$ pour la méthode Split-Step Agrawal.*



2.6.3 Conclusions.

Nous avons utilisé une solution exacte pour valider les méthodes de discrétisation employées.
Les méthodes spectrales, ici la méthode Fourier-Collocation, ont démontré leur grande précision puisque $N = 4096$ points suffisent (i.e. $\Delta t = \dfrac{T}{N} = 0,25$) pour des erreurs relatives finales de l'ordre de l'erreur théorique des schémas en espace, *Split Step classique* et *Split Step Agrawal*, soit 10^{-4} et 10^{-6}. Les valeurs importantes des erreurs pour le cas $N = 2048$ proviennent de la précision en temps inférieure à la précision des schémas d'espace.
Nous avons aussi pu vérifier la stabilité numérique de ces méthodes pour $N = 32768$.
La conservation des invariants est effectuée à l'ordre des schémas employés. La conservation de la masse est mieux préservée que la conservation de l'énergie.
Les deux méthodes sont aussi performantes dans le cas d'une solution exacte.

2.7 Amplification exacte.

L'équation de Schroedinger non linéaire s'écrit alors

$$i\frac{\partial u}{\partial z} + \frac{1}{2}\frac{\partial^2 u}{\partial t^2} + |u|^2\, u = -i\alpha u \quad , \alpha = 17 \tag{2.32}$$

Le terme ajouté $-i\alpha u$ sera traité numériquement avec le terme linéaire $\mathcal{L} = \dfrac{i}{2}\dfrac{\partial^2 u}{\partial t^2}$ qui devient alors

$$\mathcal{L} = \frac{i}{2}\frac{\partial^2 u}{\partial t^2} - \alpha u$$

après multiplication de l'équation par $-i$.

Ce terme modélise l'amortissement du signal pendant sa propagation le long de la fibre. Cet amortissement étant linéaire, on parlera alors de l'équation de Schroedinger non linéaire faiblement amortie (WDNLS).
Les pertes ($\alpha = 17$) sont compensées par un gain périodique localisé aux distances z multiples de $z_a = 0{,}04$. La valeur du gain est

$$e^{\alpha z_a} \approx 1,97$$

Pour l'instant, nous supposons que le gain est exact, i.e. sans bruit parasite.

Si nous voulons que le module de la solution u conserve sa forme quel que soit z, la condition suivante doit être vérifiée

$$\frac{\mathcal{K}^2}{z_a} \int_0^{z_a} e^{-2\alpha z} \, \mathrm{d}z = 1 \ , \text{ soit } \ \mathcal{K} = \sqrt{\frac{2\alpha z_a}{1 - e^{-2\alpha z_a}}} \approx 1,35 \tag{2.33}$$

En effet, la valeur de $\mathcal{K}$ est calculée telle que la puissance crête de l'impulsion, moyennée sur une période d'amplification, corresponde à la puissance du cas idéal sans perte ($\alpha = 0$), pour lequel l'impulsion se propage sans déformation. Cette puissance vaut 1 en unité normalisée.
La condition initiale s'écrit alors

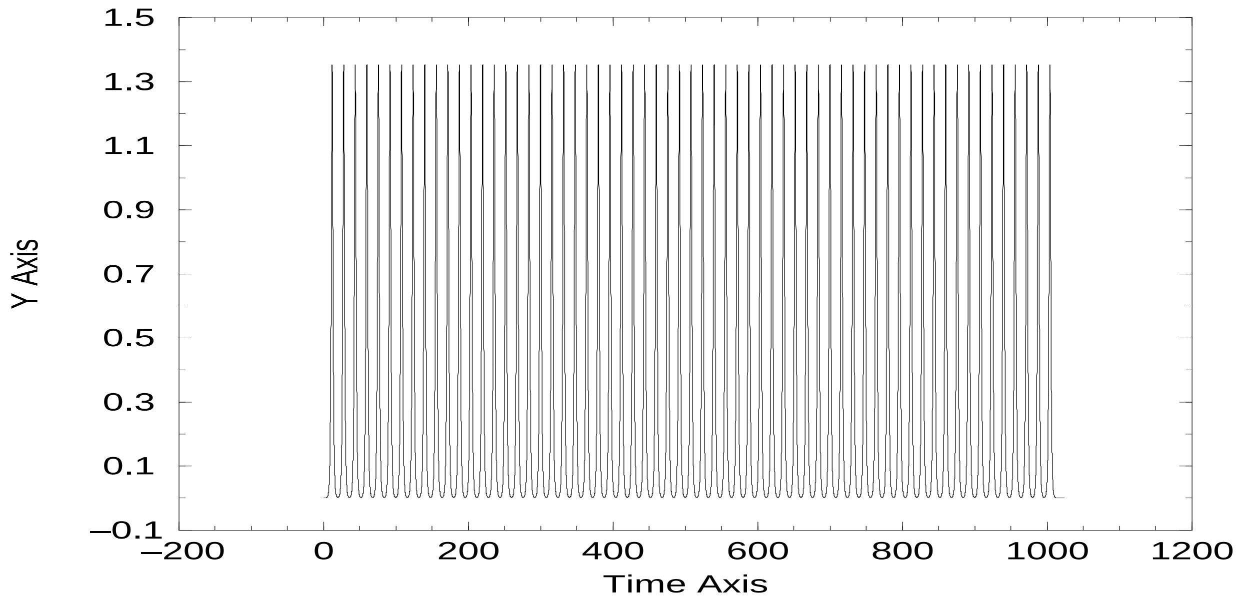
$$u(t, z = 0) = \ \mathcal{K} \sum_{j=0}^{127} \frac{1}{ch(t - 8j - 4)} \tag{2.34}$$

($a_j = 1$ si j est impair et $a_j = 0$ pour j pair en plus de $a_0 = a_{127} = 0$).
Nous n'avons plus de solution exacte. De même le second invariant n'est théoriquement plus conservé puisque le problème diffère du problème précédent. Seule la masse sera théoriquement préservée aux positions d'amplification grâce au gain $^{\alpha z_a}$.
Comme précédemment nous prenons, pour les deux méthodes, différents pas de temps Δt (i.e. différentes valeurs de N) et différents pas d'espace Δz :
N = 32768, 16384, 8192, 4096;
Δz = 10^{-3}, 10^{-2}.

2.7.1 Présentation des résultats.

Les résultats obtenus sont présentés dans les tableaux (2.5) et (2.6).
Pour les deux méthodes considérées, $|u|_{L^\infty}$ croît. Mais cette augmentation est beaucoup plus forte pour la méthode *Split-Step classique* que la méthode *Split-Step Agrawal.*
La figure (2.6) représente l'amplitude maximale d'une onde extraite du signal transmis (2.34) pour les deux méthodes avec deux pas d'espace différents $\Delta z = 10^{-3}$ et 10^{-2}.
(1) et (2) correspondent à la méthode *Split Step classique* pour $\Delta z = 10^{-2}$ et $\Delta z = 10^{-3}$.
(3) et (4) correspondent à la méthode *Split Step Agrawal* pour $\Delta z = 10^{-2}$ et $\Delta z = 10^{-3}$.

FIG. 2.5 – *Valeurs nodales de la condition initiale $u(t, z = 0)$.*



TAB. 2.5 – *Normes de la solution obtenue avec la méthode Split Step classique.*

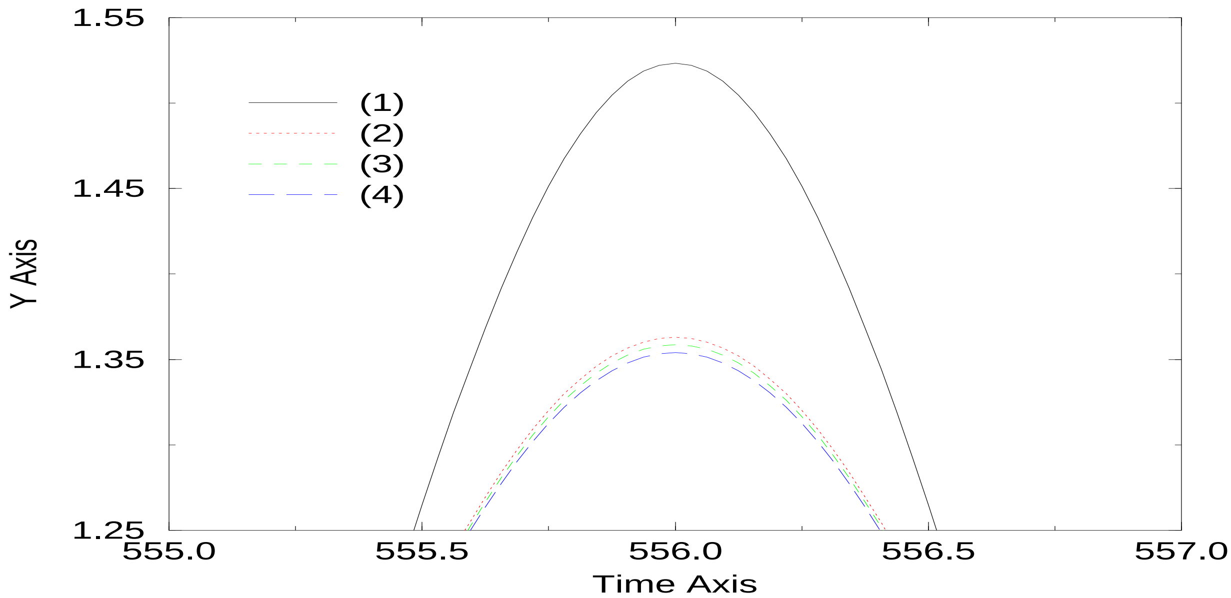
Δz	N	$\|u\|_{L^\infty}$ à $z = 0$	$\|u\|_{L^\infty}$ à $z = 13$	$\|u\|_{L^2}$ à $z = 0$	$\|u\|_{L^2}$ à $z = 13$
10^{-2}	32768	1.3526209976	1.5235730306	15.183180016	15.183180015
	16384	1.3526209976	1.5235729331	15.183180016	15.183180015
	8192	1.3526209976	1.5235728538	15.183180016	15.183180015
	4096	1.3526209976	1.5235728429	15.183180016	15.183180015
10^{-3}	32768	1.3526209976	1.3629467293	15.183180016	15.183180015
	16384	1.3526209976	1.3629467088	15.183180016	15.183180015
	8192	1.3526209976	1.3629466792	15.183180016	15.183180015
	4096	1.3526209976	1.3629465900	15.183180016	15.183180015

TAB. 2.6 – *Normes de la solution obtenue avec la méthode Split Step Agrawal.*

Δz	N	$\|u\|_{L^\infty}$ à $z = 0$	$\|u\|_{L^\infty}$ à $z = 13$	$\|u\|_{L^2}$ à $z = 0$	$\|u\|_{L^2}$ à $z = 13$
10^{-2}	32768	1,3526209976	1,3586254614	15,183180016	15,183180015
	16384	1,3526209976	1,3586254013	15,183180016	15,183180015
	8192	1,3526209976	1,3586253681	15,183180016	15,183180015
	4096	1,3526209976	1,3586252769	15,183180016	15,183180015
10^{-3}	32768	1,3526209976	1,3539180469	15,183180016	15,183180007
	16384	1,3526209976	1,3539180776	15,183180016	15,183180007
	8192	1,3526209976	1,3539180828	15,183180016	15,183180008
	4096	1,3526209976	1,3539180808	15,183180016	15,183180009

La seconde méthode donne visiblement des résultats plus précis que la première pour un même pas d'espace.

FIG. 2.6 – *Valeurs nodales de la solution $u(t, z)$ à $z = 13$, i.e. $|u(t, 13)|^2$.*



2.7.2 Conservation des invariants.

Nous présentons ici l'évolution des deux invariants considérés : la masse et l'énergie. Seule la première de ces deux quantités est préservée théoriquement. Mais nous calculons les deux. Pour la méthode *Split Step classique* cela donne

TAB. 2.7 – *Conservation des invariants avec la méthode Split Step classique.*

Δz	N	Masse à $z = 0$	Masse à $z = 13$	Energie à $z = 0$	Energie à $z = 13$
10^{-2}	32768	15,183180016	15,183180015	-281,10478628	-357,93649329
	16384	15,183180016	15,183180015	-281,10478628	-357,93649331
	8192	15,183180016	15,183180015	-281,10478628	-357,93649335
	4096	15,183180016	15,183180015	-281,10478628	-357,93649458
10^{-3}	32768	15,183180016	15,183180007	-281,10478628	-285,22634843
	16384	15,183180016	15,183180007	-281,10478628	-285,22634869
	8192	15,183180016	15,183180008	-281,10478628	-285,22634919
	4096	15,183180016	15,183180009	-281,10478628	-285,22635008

Les deux invariants sont conservés différemment. La masse qui est théoriquement ponctuellement conservée l'est effectivement à 10^{-8},10^{-9}, i.e. bien au-delà de la précision des méthodes.

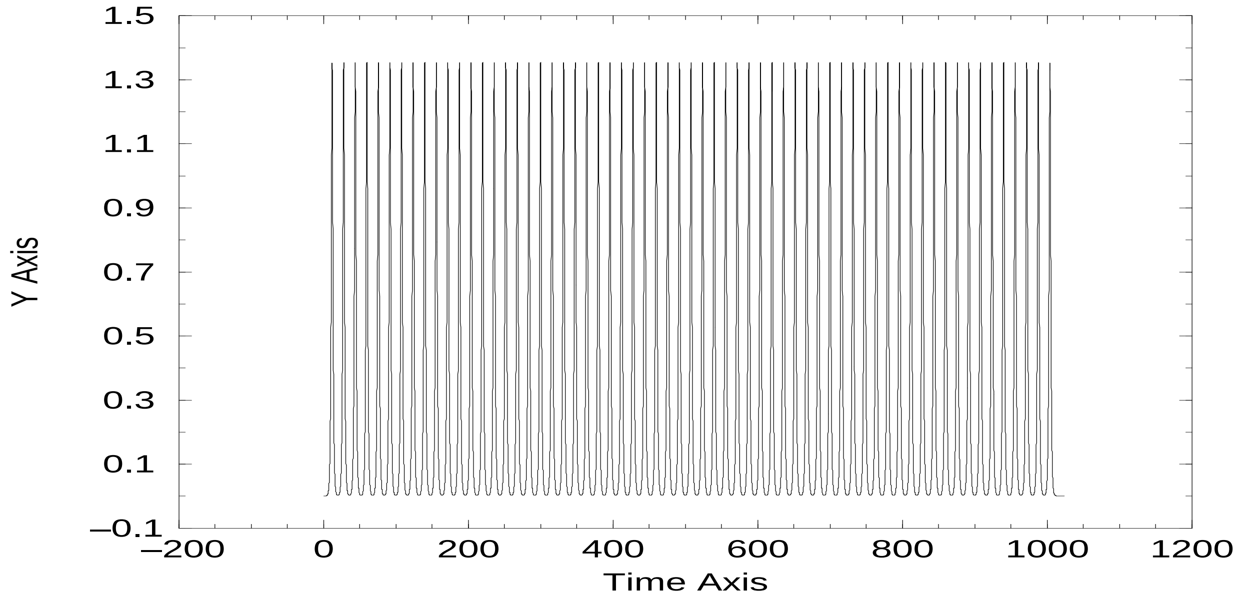
Pour la méthode *Split-Step Agrawal*

TAB. 2.8 – *Conservation des invariants avec la méthode Split Step Agrawal.*

Δz	N	Masse à $z=0$	Masse à $z=13$	Energie à $z=0$	Energie à $z=13$
10^{-2}	32768	15,183180016	15,183180015	-281,10478628	-283,47853447
	16384	15,183180016	15,183180015	-281,10478628	-283,47853468
	8192	15,183180016	15,183180015	-281,10478628	-283,47853510
	4096	15,183180016	15,183180015	-281,10478628	-283,47853590
10^{-3}	32768	15,183180016	15,183180007	-281,10478628	-281,59972139
	16384	15,183180016	15,183180007	-281,10478628	-281,59972163
	8192	15,183180016	15,183180008	-281,10478628	-281,59972209
	4096	15,183180016	15,183180009	-281,10478628	-281,59972289

L'énergie, quant à elle, est à peu près conservée par la méthode *Split-Step Agrawal* puisqu'elle enregistre une augmentation comprise entre 0,2% et 0,8% de sa valeur initiale alors qu'elle croît de plus de 27,5% pour la méthode *Split-Step classique* avec $\Delta z = 10^{-2}$.

FIG. 2.7 – *Valeurs nodales de la solution $u(t,z)$ à $z=13$, i.e. $|u(t_j, z=13)|^2$.*



2.7.3 Conclusions.

Nous considérons ici un problème de transmission exacte de signaux par une fibre optique modélisé par l'équation de Schroedinger non linéaire faiblement amortie.
Nous comparons les deux méthodes présentées précédemment pour ce problème. Leurs résultats sont de même qualité bien que l'on note une première différence qui est la forte augmentation de l'énergie pour la méthode *Split-Step classique.*
Cette transmission est exacte car les amplifications sont supposées sans parasitage, i.e. sans "bruit". Nous seront attentifs aux différences entre ces deux méthodes lorsque nous prendrons le cas plus réaliste d'une transmission bruitée.

2.8 Amplification bruitée.

On se place dans le cadre de la transmission amortie/amplifiée, i.e. l'équation de Schroedinger non linéaire faiblement amortie (WDNLS)

$$i\frac{\partial u}{\partial z} + \frac{1}{2}\frac{\partial^2 u}{\partial t^2} + |u|^2\, u = -i\alpha u \quad , \alpha = 17 \tag{2.35}$$

Les pertes sont compensées par un gain périodique localisé aux distances $z_a = 0{,}04$. La valeur du gain est $e^{\alpha z_a}$.
Nous avons la contrainte sur le module de la solution

$$\frac{\mathcal{K}^2}{z_a}\int_0^{z_a} e^{-2\alpha z}\,\mathrm{d}z = 1 \tag{2.36}$$

qui est prise en compte dans la condition initiale :

$$u(t, z = 0) = \mathcal{K}\sum_{j=0}^{127} \frac{a_j}{ch(t - 8j - 4)} \tag{2.37}$$

$a_j = 0$ ou 1 avec une probabilité $1/2$ mais on impose $a_0 = a_{127} = 0$.
A chaque amplification, le gain est parasité avec l'ajout d'un bruit $\mathcal{B}$, défini dans l'espace de Fourier, qui touche l'ensemble des modes spectraux

$$\mathcal{B}(t) = \sum_{k=-N/2}^{N/2-1} \hat{w}_k e^{i\Phi_k(t)} \tag{2.38}$$

Ce bruit a une puissance $\mathrm{P}_{\mathrm{bruit}} = 3,7.10^{-4}$ (unité normalisée). Cela signifie que tous les modes de Fourier s'écrivent sous la forme

$$\hat{b}_k = re^{i\Psi_k}$$

où r est une constante égale à

$$r = \sqrt{\frac{\mathrm{P}_{\mathrm{bruit}}}{N}} \quad (\approx 1,51.10^{-4} \text{ pour } N = 32768)$$

grâce à l'égalité de Parseval.

Remarque 5

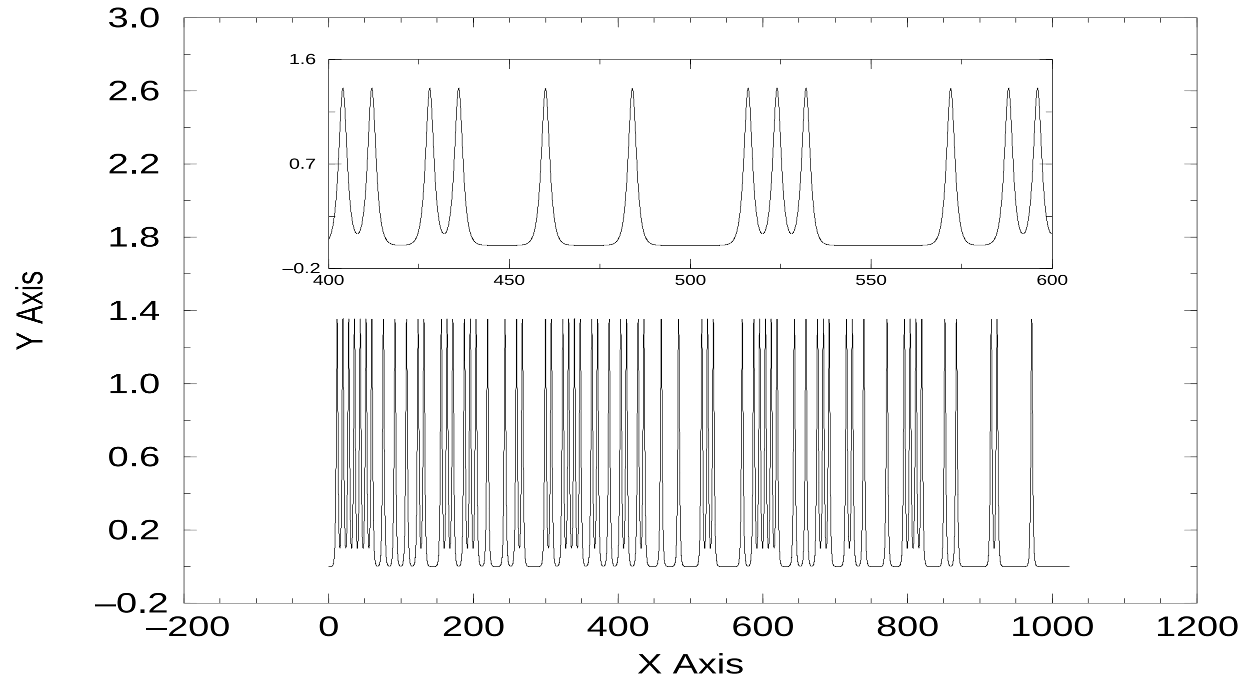
Nous verrons que le problème de la transmission amortie/amplifiée-bruitée sera toujours numériquement traité avec $N = 32768$ points. Cette valeur de N est imposée par l'ajout du bruit.

La phase de chaque mode, $\hat{b}_k$, à savoir Ψ_k, est un nombre aléatoire de loi uniforme, compris entre 0 et 2π.

Remarque 6

Les phases aléatoires du bruit sont générées numériquement par un algorithme, une suite, et sont donc pseudo-aléatoires. Cette suite est complètement déterminée par la donnée d'une "clé". A chaque amplification, toutes les phases sont regénérées.

FIG. 2.8 – *Valeurs nodales de la condition initiale* $u(t, z = 0)$.



2.8.1 Caractérisation de la qualité d'une transmission.

En fin de liaison, il nous faut décoder l'information, donc interpréter si le symbole reçu dans chaque intervalle de temps de longueur $\mathrm{T_b}$ correspond à un "1" ou à un "0".
Pour cela on mesure l'énergie du signal reçu dans chaque intervalle de temps $[j\mathrm{T_b},\ (j+1)\mathrm{T_b}]$. Si elle est inférieure à un seuil (que l'on ajuste), on considère que l'on a affaire à un "0" et si elle est supérieure au seuil, il s'agit d'un "1". On appelle alors *Taux d'erreur* (TEB) la probabilité que l'on a de faire une erreur de décodage, c'est-à-dire de prendre un "0" pour un "1" ou l'inverse.

Nous caractérisons la qualité d'une transmission par les mesures statistiques suivantes : on note $v(t) = |u(t, z = 13)|^2$.
Soit c_j le barycentre de chaque impulsion, i.e. le barycentre de l'énergie reçue dans l'intervalle de longueur $\mathrm{T_b}$, défini pour tout j tel que $a_j = 1$ par

$$c_j = \frac{\int_{t=8j}^{t=8(j+1)} (t - 8j - 4) v(t)\ \mathrm{d}t}{\int_{t=8j}^{t=8(j+1)} v(t)\ \mathrm{d}t} \tag{2.39}$$

Nous définissons alors l'écart-type des c_j, σ_{gt}, i.e. l'écart-type du temps d'arrivée, lui-même égal au barycentre de l'énergie reçue dans l'intervalle de longueur $\mathrm{T_b}$, par :

$$\sigma_{gt} = \sqrt{< c_j^2 > - < c_j >^2}$$

où $< . >$ désigne une moyenne sur les j tel que $a_j = 1$.
Nous appelons σ_{gt} la "gigue temporelle".

On pose

$$Q_t = \frac{0,7\mathrm{T_b}}{2\sigma_{gt}} \tag{2.40}$$

Soit $A_j = \frac{1}{8}\int_{t=8j}^{t=8(j+1)} v(t)\ dt$ défini pour tout $j \in [\![0,\ 127]\!]$.
Nous définissons le facteur de qualité Q par

$$Q = \frac{\mu_1 - \mu_0}{\sigma_1 + \sigma_0} \tag{2.41}$$

où :
μ_1 et σ_1 sont respectivement la valeur moyenne et l'écart-type des A_j calculés sur les j tels que $a_j = 1$;
μ_0 et σ_0 sont respectivement la valeur moyenne et l'écart-type des A_j calculés sur les j tels que $a_j = 0$.
Q_t et Q sont les deux mesures quantitatives de qualité d'une liaison.

2.8.2 Appréciation de la qualité d'une transmission.

Deux phénomènes engendrés par le bruit d'amplification sont à l'origine des erreurs :
• après transmission, l'amplitude de chaque symbole transmis (pour $a_j = 0$ ou $a_j = 1$) est aléatoire et est donc susceptible de passer au-delà du seuil pour $a_j = 0$ ou en deça pour $a_j = 1$, provoquant une erreur. C'est de ce phénomène que rend compte le facteur Q.
On peut en effet montrer que si l'amplitude des "1" et des "0" suit une distribution gaussienne, ce qui est généralement une bonne approximation, le taux d'erreur vaut

$$TEB = \frac{1}{2} erfc\left(\frac{Q}{\sqrt{2}}\right) \tag{2.42}$$

où $erfc$ est la fonction "erreur complémentaire" :

$$erfc(x) = 1 - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2}\ \mathrm{d}y \tag{2.43}$$

• La seconde source d'erreur vient du fait que le "temps d'arrivée" de chaque symbole "1" dans son intervalle de longueur $\mathrm{T_b}$ est aléatoire et qu'il y a alors un risque pour qu'il "déborde" sur le symbole voisin, ce qui provoque une erreur.
C'est ce phénomène que mesure la "gigue temporelle", σ_{gt}.
On peut montrer qu'avec un récepteur standard, le taux d'erreur correspondant vaut

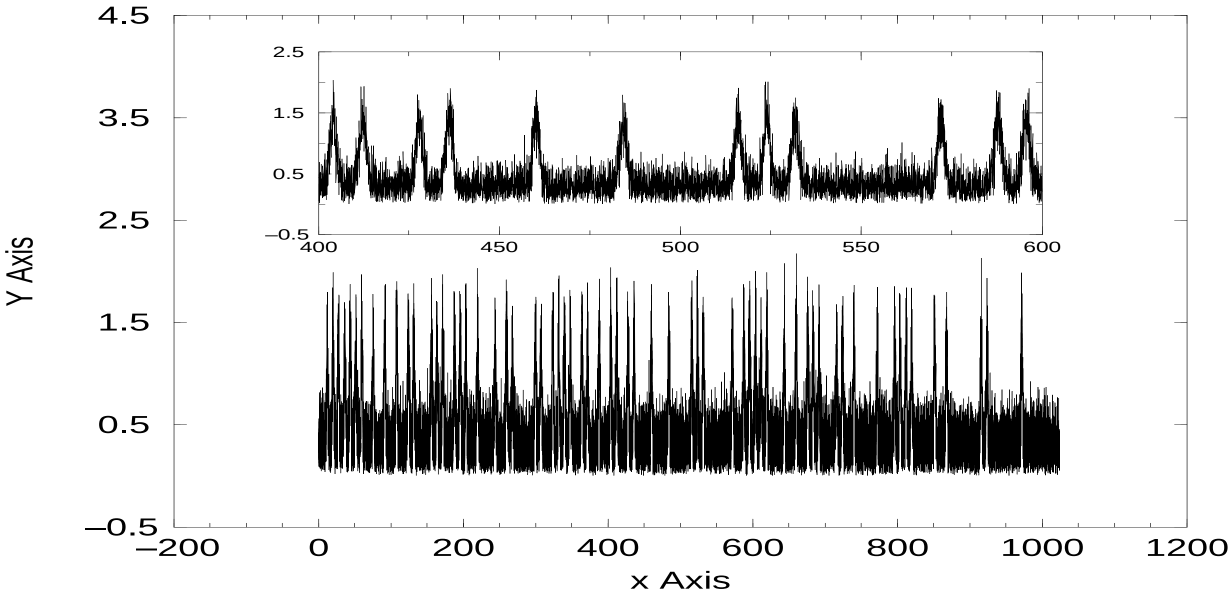
$$TEB = erfc\left(\frac{Q_t}{\sqrt{2}}\right) \tag{2.44}$$

Plus Q et Q_t sont élevés, plus le taux d'erreur est faible. Au cours de nos simulations, on considère qu'un système de transmission est viable si Q et Q_t sont supérieurs à 10 ou 12. Les facteurs de qualités sont significativement supérieurs aux minima précédents. Ce sont des grandeurs statistiques, il existe une barre d'erreur que nous estimons à deux ou trois unités; i.e. Q doit être supérieur à $16,5 \pm 2$ ou 3 et Q_t à $20,5 \pm 2$ ou 3.

2.8.3 Analyse des résultats.

Jusqu'à présent, nous prenions $N = 32768$ points plus pour vérifier la stabilité numérique dans le cas de solution exacte que par réel besoin, puisque comme nous l'avons vu précédemment $N = 4096$ points fournissait des résultats d'aussi bonne qualité.
Par conséquent le temps calcul nécessaire à simuler la transmission d'un signal n'était pas énorme.
Maintenant que nous devons prendre $N = 32768$ points, le temps calcul est forcément beaucoup plus important (on peut l'estimer à environ un facteur 10). Une alternative est d'augmenter le pas d'espace Δz, ce que nous avons testé.
Dans cet optique, nous précisons dorénavant le temps calcul de chaque simulation; les calculs, sauf mention contraire étant effectués sur le super calculateur CRAY YMP du Campus d'Orsay. Les temps calcul sont déterminés par la machine. Ces données peuvent varier suivant l'occupation de la machine. Nous avons effectué plusieurs simulations pour certaines combinaisons des paramètres et les temps calcul varient au maximum de 10%.
Pour les deux méthodes précédentes, nous avons effectué différents calculs : d'une part nous prenons $N = 32768$ et faisons varier Δz de 10^{-3} à 10^{-2} en prenant bien soin d'avoir toujours un sous-multiple de $z_a = 0,04$ pour pouvoir amplifier le signal aux bons endroits.
D'autre part, nous prenons $\Delta z = 10^{-3}$ fixé et diminuons le nombre de points d'un facteur 2 à chaque fois.

FIG. 2.9 – *Valeurs nodales de la solution $u(t,z)$ à $z = 13$, i.e. $|u(t,z = 13)|^2$ pour $N = 32768$, $\Delta z = 10^{-3}$.*



Les résultats sont détaillés dans les tableaux suivants.
Nous voyons aisément à l'aide du tableau 2.9 que la méthode *Split Step classique* ne fournit de bons résultats que pour les paramètres $(N, \Delta z) = (32768, 10^{-3})$. Pour les autres couples, les facteurs de qualité sont bien inférieurs aux minima fixés.

TAB. 2.9 – *Résultats obtenus avec la méthode Split Step classique (clé1).*

N	Δz	Q_t	Q	Temps CPU
32768	10^{-3}	17.5692	17.1006	4149 s
32768	$2\ 10^{-3}$	20.5283	7.7817	1998 s
32768	$4\ 10^{-3}$	18.5560	7.4388	1053 s
32768	$8\ 10^{-3}$	23.1968	5.1183	580 s
32768	10^{-2}	18.6440	4.2446	490 s
16384	10^{-3}	11.8136	13.8131	2019 s
8192	10^{-3}	8.1099	8.1569	1059 s
4096	10^{-3}	5.2981	5.2340	543 s

TAB. 2.10 – *Résultats obtenus avec la méthode Split Step Agrawal (clé1).*

N	Δz	Q_t	Q	Temps CPU
32768	10^{-3}	20.3891	15.5669	6977 s
32768	$2\ 10^{-3}$	20.7823	15.8193	3520 s
32768	$4\ 10^{-3}$	20.7361	15.6696	1859 s
32768	$8\ 10^{-3}$	21.2304	15.1917	996 s
32768	10^{-2}	20.6165	15.1563	817 s
16384	10^{-3}	12.6154	12.1220	3658 s
8192	10^{-3}	8.3478	7.9001	1928 s
4096	10^{-3}	5.3694	5.1409	1015 s

Remarque 7

Pour générer les bruits aléatoires des simulations précédentes, la "clé" a été la même partout : cela permet de voir les seules conséquences du changement des paramètres N et Δz.

Remarque 8

En pratique, le pas d'espace Δz est plus près de 10^{-3} que de 10^{-2}, généralement de l'ordre de $1,5.10^{-3}$.

TAB. 2.11 – *Résultats obtenus avec la méthode Split Step Agrawal (clé2).*

N	Δz	Q_t	Q	Temps CPU
32768	10^{-3}	15.2963	14.1205	6977 s
32768	$2\ 10^{-3}$	15.3327	13.9429	3351 s
32768	$4\ 10^{-3}$	15.1351	14.3026	1795 s
32768	$8\ 10^{-3}$	15.3626	14.0868	946 s
32768	10^{-2}	15.4526	14.2758	800 s

La méthode *Split Step Agrawal*, quant à elle, fournit des résultats plus satisfaisants. Pour cette seconde méthode nous avons testé différentes clés qui génèrent donc des bruits différents. Les tableaux 2.10 et 2.11 présentent les résultats pour deux clés distinctes (clé1 = 1010, clé2 = 4500). Cela met en évidence la sensibilité des résultats au bruit existant. De plus il est clair que le nombre de points $N = 32768$ est nécessaire.

Nous allons examiner la solution pour voir si nous ne pouvons pas la calculer à moindre coût.

2.9 La méthode multi-niveaux Split Step Agrawal (MLSSA).

Dans cette partie, nous appelons cycle l'intervalle d'espace de longueur $z_a = 0,04$ qui commence à une amplification/bruitage et qui finit juste avant l'amplification/bruitage qui suit.

2.9.1 Idée de la méthode.

Pour examiner la solution, nous avons considéré l'évolution du spectre de Fourier de la solution au cours d'un cycle.
Au bout de 10 cycles, où l'influence de la condition initiale n'est pas trop forte, nous avons représenté l'allure du spectre à différents endroits, c'est-à-dire pour différentes valeurs de z.
Le spectre de Fourier de u est plat à partir du nombre d'onde $k = 1000$.
Nous remarquons que lorsque l'on avance dans le cycle, d'une part l'allure du spectre est conservée et d'autre part il présente une diminution uniforme de l'amplitude des modes. Les figures précédentes montrent bien ce phénomène, le spectre est "translaté" vers le bas mais son aspect est le même.
L'idée est donc de découpler la partie du spectre située en deça d'un niveau N_1 (à déterminer), que nous appellerons $\mathbf{v}$, de la partie située au-delà que nous appellerons $\mathbf{w}$.
Cette décomposition a lieu dans l'espace de Fourier

$$u = \mathbf{v} + \mathbf{w} \tag{2.45}$$

$$\mathbf{v} = \sum_{|k| \leqslant N_1/2} \hat{u}_k e^{\Phi_k(t)} \text{ et } \mathbf{w} = \sum_{N_1/2 < |k| < N/2} \hat{u}_k e^{\Phi_k(t)} \tag{2.46}$$

avec

$$\Phi_k(t) = ik\frac{2\pi}{T}t$$

Pour évaluer l'action des différents opérateurs sur les $\mathbf{w}$, nous avons effectué le test suivant : nous nous plaçons au début du cycle 11 (c'est-à-dire au point d'abscisse $z = 10z_a$).

FIG. 2.10 – *Spectre d'énergie de $u(t, z)$ à $z = 0.400$ (gauche) et à $z = 0.413$ (droite).*

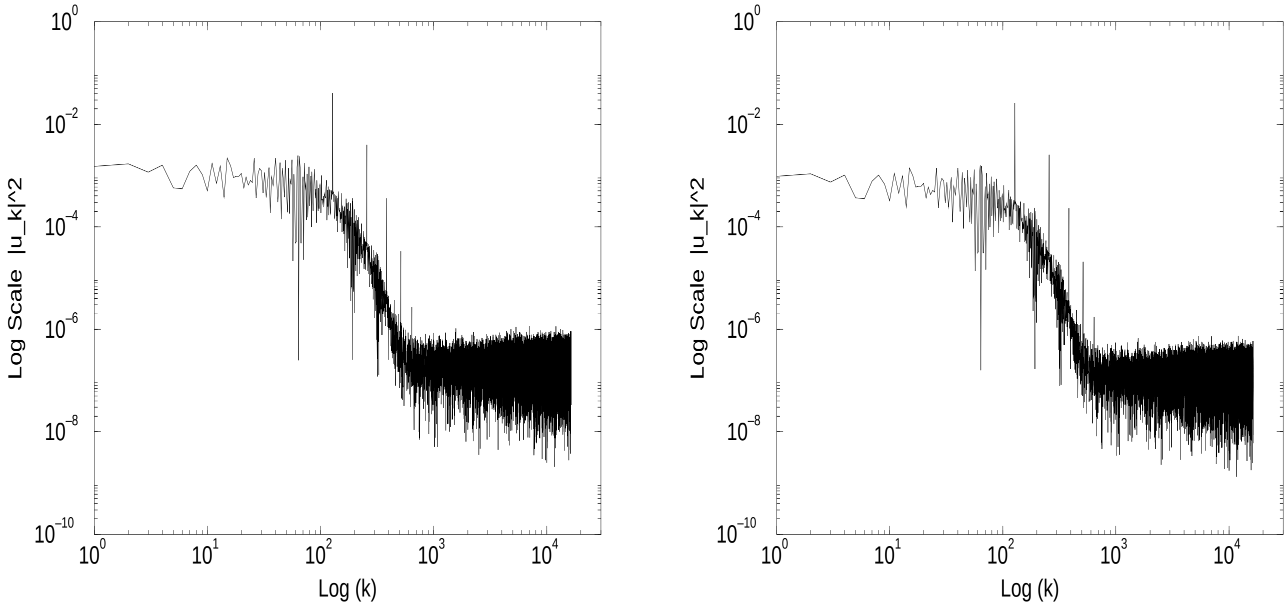
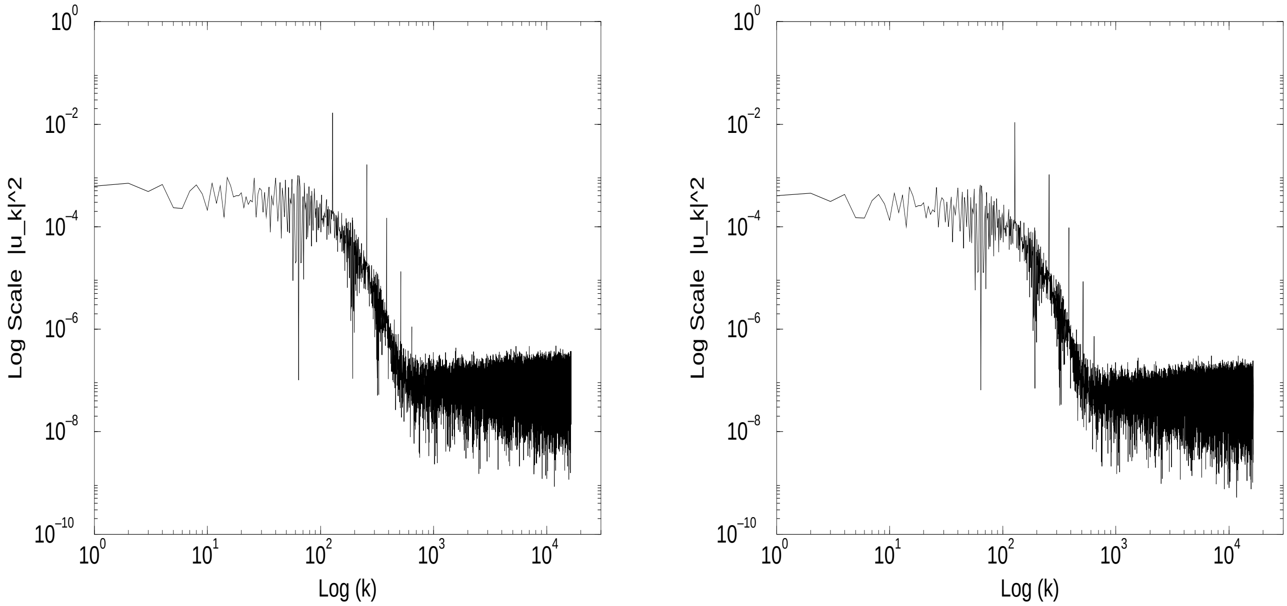


FIG. 2.11 – *Spectre d'énergie de $u(t, z)$ à $z = 0.426$ (gauche) et à $z = 0.439$ (droite).*



On considère l'équation de Schroedinger linéaire :

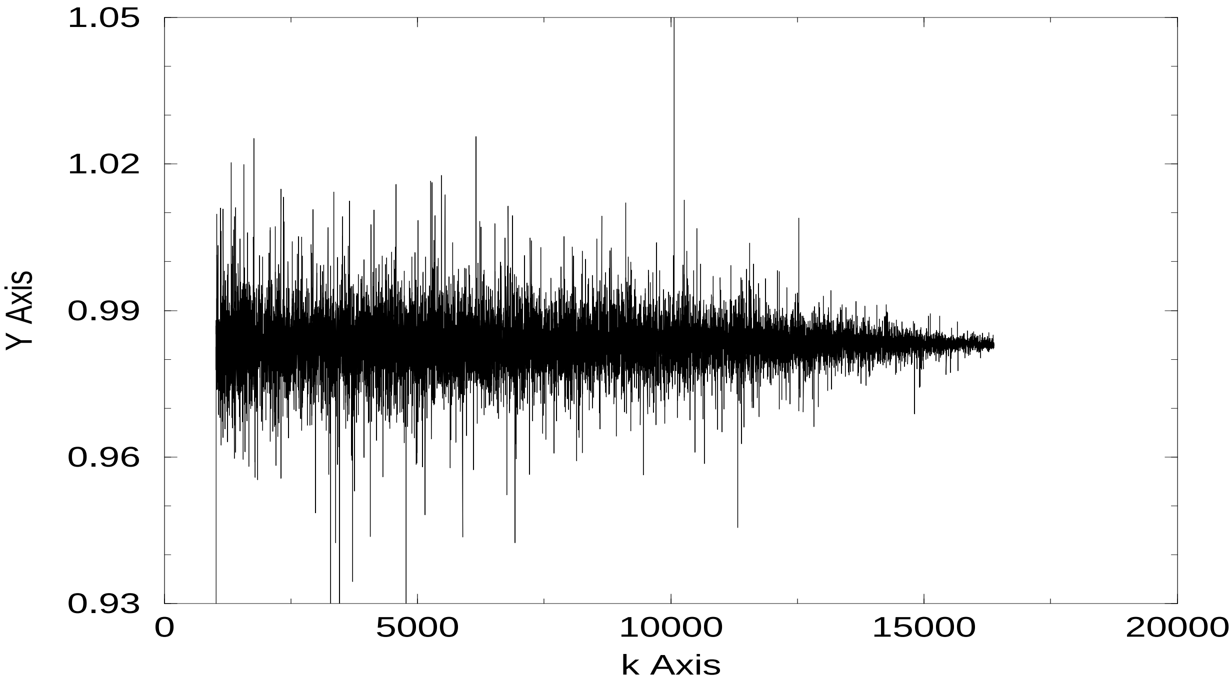
$$i\frac{\partial u}{\partial z} + \frac{1}{2}\frac{\partial^2 u}{\partial t^2} = i\alpha u \qquad \text{avec} \quad \alpha = 17 \tag{2.47}$$

La comparaison de l'évolution au cours du cycle des $\mathbf{w}$ régis uniquement par l'équation linéaire d'une part et l'évolution des $\mathbf{w}$ provenant de l'équation complète nous a amené à la conclusion suivante :

$$\left|\widehat{\mathbf{w}}_{\mathrm{lin}}(k)\right| \approx 0,9832 \left|\widehat{\mathbf{w}}_{\mathrm{nlin}}(k)\right| \ \text{ pour } N_1/2 < |k| < N/2 \tag{2.48}$$

à la fin du cycle.

FIG. 2.12 – *Quotient des modules des modes de Fourier de* $\mathbf{w}$ *:* $\left|\widehat{\mathbf{w}}_{lin}(k)\right| / \left|\widehat{\mathbf{w}}_{nlin}(k)\right|$.



$\widehat{\mathbf{w}}_{\mathrm{lin}}(k)$ désigne le mode de fréquence k pour les $\mathbf{w}$ calculés par l'équation linéaire (2.47); $\widehat{\mathbf{w}}_{\mathrm{nlin}}(k)$ désigne le mode de fréquence k pour les $\mathbf{w}$ calculés par l'équation non linéaire, i.e. l'équation complète.

Cela signifie que le module d'un mode k pris dans le spectre des $\mathbf{w}$ est donné presque totalement par l'opérateur linéaire et donc que l'opérateur non linéaire n'a qu'une action très limitée.

Ce test est effectué avec différents "bruits" (i.e. différentes "clés") et les résultats obtenus sont très proches: le facteur dans la relation (2.48) oscille entre 0,9820 et 0,9840.

Nous avons ainsi décidé de faire avancer les $\mathbf{v}$ par l'équation complète projetée sur les N_1 premiers modes et les $\mathbf{w}$ par l'équation linéaire (2.47) projetée sur les $N - N_1$ derniers modes.

2.9.2 Les équations de la méthode (MLSSA).

La recherche des nouvelles équations à résoudre s'effectue dans l'espace spectral où la séparation des $\mathbf{v}$ et des $\mathbf{w}$ est plus maniable.

Pour déterminer l'équation vérifiée par les $\mathbf{v}$ nous partons de l'équation (WDNLS) dont u est la solution écrite sous la forme

$$\frac{\partial u}{\partial z} = \mathcal{L}u + i\,\mathcal{N}(u)u \tag{2.49}$$

où les opérateurs $\mathcal{L}$ et $\mathcal{N}$ sont définis par:

$$\begin{aligned} \mathcal{L}u &= \frac{i}{2}\frac{\partial^2 u}{\partial t^2} - \alpha u \\ \mathcal{N}(u) &= |u|^2 \end{aligned}$$

Définition *1*

On note :

$\mathbf{P}_{N_1}$ l'opérateur de projection sur les N_1 premiers modes, dans l'espace spectral :

$$\mathbf{v} = \mathbf{P}_{N_1} u = \sum_{|k| \leqslant N_1/2} \hat{u}_k e^{\Phi_k(t)}$$

$\mathbf{Q}_{N_1}$ l'opérateur de projection sur les $N - N_1$ derniers modes dans l'espace spectral :

$$\mathbf{w} = \mathbf{Q}_{N_1} u = \sum_{N_1/2 < |k| < N/2} \hat{u}_k e^{\Phi_k(t)}$$

$\mathbf{I}_{N_1}$ l'opérateur d'interpolation sur les N_1 noeuds $t_j = \dfrac{jT}{N_1}$, $j = 0, ..., N_1 - 1$, dans l'espace physique :

$$\mathbf{I}_{N_1}(u)(t_j) = u(t_j)$$

On part de (2.49) que l'on projette à l'aide de $\mathbf{P}_{N_1}$

$$\mathbf{P}_{N_1} \frac{\partial u}{\partial z} = \mathbf{P}_{N_1}\ \mathcal{L}u + \mathbf{P}_{N_1}\ \{i\ \mathcal{N}(u)u\}$$

D'où

$$\frac{\partial \mathbf{v}}{\partial z} = \ \mathcal{L}\mathbf{v} + i\mathbf{P}_{N_1}\ \{\ \mathcal{N}(u)\mathbf{v}\} \tag{2.50}$$

Pour l'opérateur non linéaire, nous considérons la partie composée des $\mathbf{v}$ et celle provenant de l'interaction des $\mathbf{v}$ et des $\mathbf{w}$, ce que l'on peut écrire :

$$\begin{aligned} \mathcal{N}(u) &= \mathcal{N}(\mathbf{v} + \mathbf{w}) = |\mathbf{v} + \mathbf{w}|^2 = (\mathbf{v} + \mathbf{w})(\overline{\mathbf{v}} + \overline{\mathbf{w}}) \\ &= |\mathbf{v}|^2 + \left(|u|^2 - |\mathbf{v}|^2\right) \\ &= \mathcal{N}(\mathbf{v}) + (\ \mathcal{N}(u) -\ \mathcal{N}(\mathbf{v})) \end{aligned} \tag{2.51}$$

En insérant (2.51) dans (2.50) nous obtenons

$$\mathbf{P}_{N_1}\ [\ (\ \mathbf{P}_{N_1}\ \mathcal{N}(\mathbf{v}) + \mathbf{P}_{N_1}\ (\ \mathcal{N}(u) -\ \mathcal{N}(\mathbf{v})\ )\ )\ \mathbf{v}\ ]$$

Pour la méthode Fourier-Collocation, où la résolution a lieu dans l'espace physique l'équation (2.50) s'écrit :

$$\mathbf{I}_{N_1} \frac{\partial \mathbf{v}}{\partial z} = \mathbf{I}_{N_1}\ \mathcal{L}\mathbf{v} + i\,\mathbf{I}_{N_1}\ (\mathbf{P}_{N_1}\ [\ \mathbf{P}_{N_1}\{\ \mathcal{N}(\mathbf{v})\} + i\,\mathbf{P}_{N_1}\{\ \mathcal{N}(u) -\ \mathcal{N}(\mathbf{v})\ \}\ ]\ \mathbf{v}\ ) \tag{2.52}$$

L'équation des $\mathbf{w}$ s'obtient en projetant l'équation de Schroedinger linéaire sur les $N - N_1$ derniers modes dans l'espace spectral.

$$\frac{\partial u}{\partial z} = \ \mathcal{L}u \text{ avec } \ \mathcal{L}u = \frac{i}{2}\frac{\partial^2 u}{\partial t^2} - \alpha u \tag{2.53}$$

On projette cette équation à l'aide de $\mathbf{Q}_{N_1}$:

$$\mathbf{Q}_{N_1} \frac{\partial u}{\partial z} = \mathbf{Q}_{N_1}\ \mathcal{L}u$$

d'où

$$\frac{\partial \mathbf{w}}{\partial z} = \mathcal{L}\mathbf{w} \tag{2.54}$$

L'équation des $\mathbf{w}$ étant totalement linéaire, elle s'intègre exactement en espace, en une seule étape, quelle que soit la longueur de l'intervalle spatial considéré. De surcroît il est plus intéressant d'intégrer cette équation dans l'espace spectral car l'opérateur $\mathcal{L}$ y est diagonal.

Le système projeté à résoudre, qui découle de (NLS), est donc

$$\begin{aligned} \mathbf{I}_{N_1} \frac{\partial \mathbf{v}}{\partial z} &= \mathbf{I}_{N_1}\, \mathcal{L}\mathbf{v} + i\, \mathbf{I}_{N_1}\, \left(\mathbf{P}_{N_1}\, \left[\, \mathbf{P}_{N_1} \{\, \mathcal{N}(\mathbf{v})\} + i\, \mathbf{P}_{N_1} \{\, \mathcal{N}(u) - \mathcal{N}(\mathbf{v})\, \}\, \right]\, \mathbf{v}\, \right) \quad \text{(a)} \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (2.55) \\ \frac{\partial \mathbf{w}}{\partial z} &= \mathcal{L}\mathbf{w} \quad \text{(b)} \end{aligned}$$

Traitement du terme non linéaire.

L'équation (2.55a) gouvernant $\mathbf{v}$ dépend aussi de $\mathbf{w}$ par le terme non linéaire d'interaction $\mathbf{P}_{N_1} \{\, \mathcal{N}(u) - \mathcal{N}(\mathbf{v})\, \}$. Nous avons remarqué précédemment, par la relation (2.48) que les modes de Fourier de $\mathbf{w}$ décroissaient uniformément à cause du coefficient d'amortissement linéique. Pour éliminer la dépendance en $\mathbf{w}$ dans (2.55a), nous recherchons un terme à substituer au véritable terme non linéaire. Nous prenons comme candidat le terme non linéaire pour un z_0 donné, soumis à l'action de α. Alors nous calculons

$$\mathcal{NLT}_1 = \mathbf{P}_{N_1} \{\, \mathcal{N}(u(t,z))\, \}, \tag{2.56}$$

et

$$\mathcal{NLT}_2 = \mathbf{P}_{N_1}\, \mathcal{N}(\mathbf{v}(t,z)) + e^{-2\alpha\delta_z} \mathbf{P}_{N_1} \left[\, \mathcal{N}(vw(t,z_0))\, \right], \tag{2.57}$$

où

$$\mathcal{N}(vw(t,z_0)) = \mathcal{N}(u(t,z_0)) - \mathcal{N}(\mathbf{v}(t,z_0)),$$

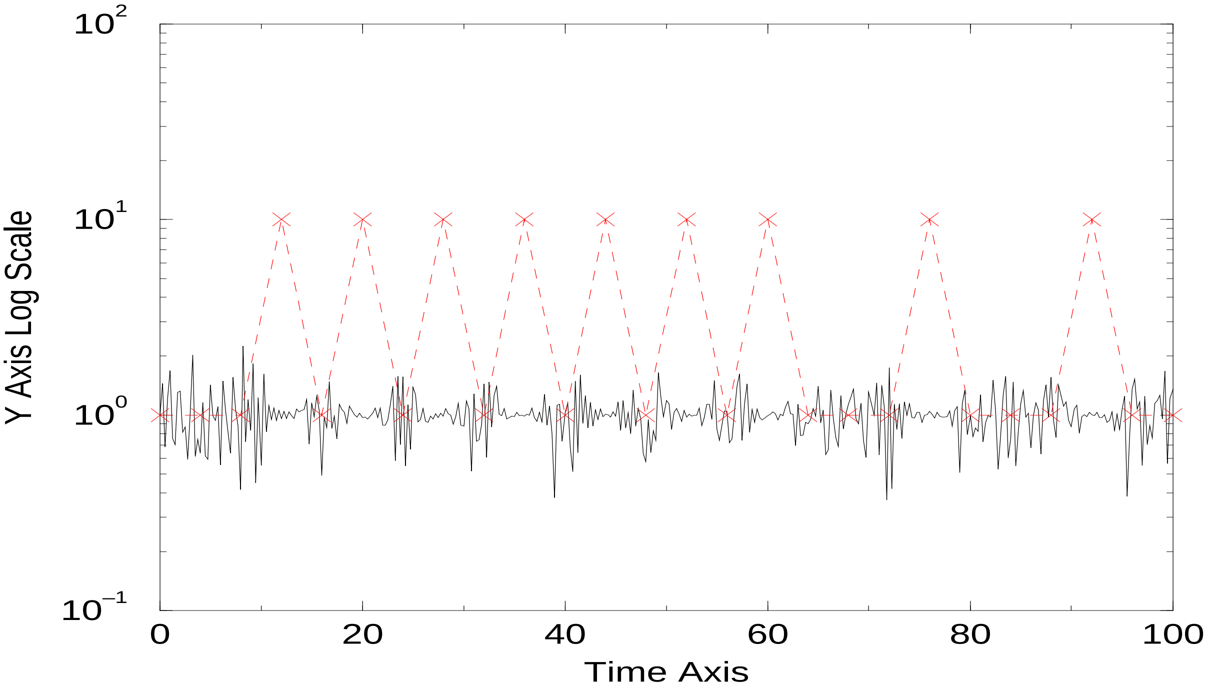
et $z = z_0 + \delta_z$, pour $\delta_z > 0$.
Pour estimer la précision de cette approximation, nous effectuons le test suivant : soit N_1 fixé (e.g. $N_1 = 4096$) et pour une valeur fixée de t_j ($j \in [\![0,\, N_1 - 1]\!]$) nous intégrons l'équation (WDNLS) (2.49) de z_0 à $z_0 + z_a$ où z_0 est un point d'amplification ($z_0 = l.z_a$, $l \in \mathbb{N}^*$). Nous évaluons alors les moyennes suivantes :

$$\begin{aligned} < \mathcal{NLT}_1 >_j &= < \mathbf{I}_{N_1} \{\, \mathbf{P}_{N_1}\, \mathcal{N}(u(t,z))\, \} >_j \\ &= \frac{1}{z_a} \int_{z_0}^{z_0+z_a} \mathbf{I}_{N_1} \{\, \mathbf{P}_{N_1}\, \mathcal{N}(u(t_j,z))\, \}\ \mathrm{d}z, \end{aligned}$$

et

$$\begin{aligned} < \mathcal{NLT}_2 >_j &= < \mathbf{I}_{N_1} \left\{\, \mathbf{P}_{N_1}\, \mathcal{N}(\mathbf{v}(t_j,z)) + e^{-2\alpha\delta_z} \mathbf{P}_{N_1}\, \mathcal{N}(vw(t,z_0))\, \right\} >_j \\ &= \frac{1}{z_a} \int_{z_0}^{z_0+z_a} \mathbf{I}_{N_1} \left\{\, \mathbf{P}_{N_1}\, \mathcal{N}(\mathbf{v}(t,z)) + e^{-2\alpha(z-z_0)} \mathbf{P}_{N_1}\, \mathcal{N}(vw(t_j,z_0))\, \right\}\ \mathrm{d}z. \end{aligned}$$

FIG. 2.13 – *Quotient des moyennes des termes non linéaires* $<\mathcal{NLT}_2>_j$ / $<\mathcal{NLT}_1>_j$



Nous avons considéré ces moyennes pour plusieurs valeurs de N_1 et les avons calculées sur différents cycles de l'intervalle de z, $[0, 13]$. Dans les figures 2.13 and 2.14 les lignes continues représentent le quotient $< \mathcal{NLT}_2 >_j$ / $< \mathcal{NLT}_1 >_j$ et les pics en pointillés la localisation des informations binaires et nous pouvons voir que le quotient reste de l'ordre de 1. Le fait que les deux quantités $< \mathcal{NLT}_1 >_j$ et $< \mathcal{NLT}_2 >_j$ soient du même ordre de grandeur est important puisque ce sont des quantités aléatoires. La précision est d'autant meilleure là où l'information est présente (symbole "1") plutôt que là où les parasites dominent (symbole "0").

Ainsi, dans la mise en œuvre de cette méthode, nous substituerons le terme $\mathcal{NLT}_2$ calculé à l'aide de la formule (2.57) au véritable terme non linéaire $\mathcal{NLT}_1$ tant que nous intégrons le système des deux équations.

2.9.3 Application de la méthode.

Nous allons maintenant considérer la manière d'utiliser ce découplage.
On se fixe les idées en prenant un pas d'espace $\Delta z = 10^{-3}$.
Le cycle représente alors 40 itérations pour $z_a = 0,04$.
Nous le décomposons à l'aide des 3 étapes suivantes.

Etape 1.
On effectue *nbiter1* itérations avec l'équation complète (2.49) en u.

Etape 2.
On sépare alors, dans l'espace spectral la solution $u = \mathbf{v} + \mathbf{w}$ à l'aide des opérateurs $\mathbf{P}_{N_1}$ et $\mathbf{Q}_{N_1}$.

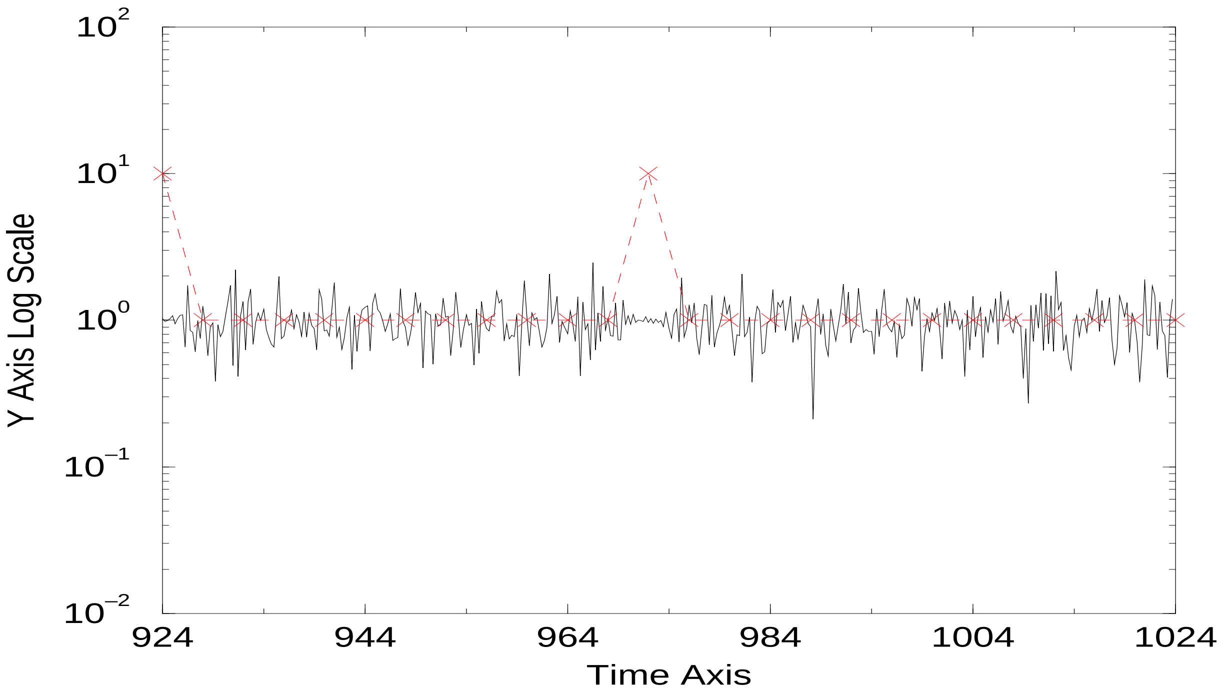
On évalue la partie du terme non linéaire précédent correspondant à l'interaction des $\mathbf{v}$ et des $\mathbf{w}$:

$$\mathbf{P}_{N_1}\left\{\,\mathcal{N}(u)-\ \mathcal{N}(\mathbf{v})\right\},$$

partie qui sera figée pour toute cette deuxième étape.
On effectue alors *nbiter2* avec le système découplé (2.55) pour $\mathbf{v}$ et $\mathbf{v}$.

Etape 3.
On reforme la solution complète, u, dans l'espace spectral : $u = \mathbf{v} + \mathbf{w}$ où $\mathbf{v}$ provient de l'équation (2.55 (a)) et $\mathbf{w}$ de l'équation linéaire (2.55 (b)).
On effectue alors *nbiter3* itérations avec l'équation (NLS) complète (2.49) comme à l'étape 1.

FIG. 2.14 – *Quotient des moyennes des termes non linéaires* $<\mathcal{NLT}_2>_j \,/\, <\mathcal{NLT}_1>_j$



On appelle *V-cycle* le couple formé par les étapes 2 et 3. L'algorithme consiste alors en l'étape 1 suivie d'un certain nombre de *V-cycle* de telle sorte que l'on termine par une étape 3 juste avant d'arriver à l'amplification/bruitage suivante :
i.e on travaille bien avec l'équation complète avant et après une amplification/bruitage suivante.
La condition à respecter pour le nombre d'itérations de chaque étape est la suivante :

$$\textit{nbiter1}\ + (\text{nombre de } \textit{V-cycle})\,(\textit{nbiter2} + \textit{nbiter3}\,) = 40$$

De manière plus générale, elle s'écrit :

$$\textit{nbiter1}\ + (\text{nombre de } \textit{V-cycle})\,(\textit{nbiter2} + \textit{nbiter3}\,) = \frac{z_a}{\Delta z} \tag{2.58}$$

2.9.4 Mise en œuvre de la méthode.

Nous avons constaté précédemment que la méthode *Split-Step classique* n'était pas adaptée pour résoudre ce problème, par conséquent nous avons greffé cette idée uniquement sur la méthode *Split-Step Agrawal*.
Cela signifie que la résolution numérique des deux équations non linéaires (2.49) et (2.55 (a)) sont réalisées par cette méthode (l'équation linéaire ne posant pas de problème).
De plus, la valeur $N = 32768$ est indispensable, elle est désormais fixée. Le pas d'espace Δz, quant à lui sera pris au début égal à 10^{-3} puis on essayera de l'augmenter.

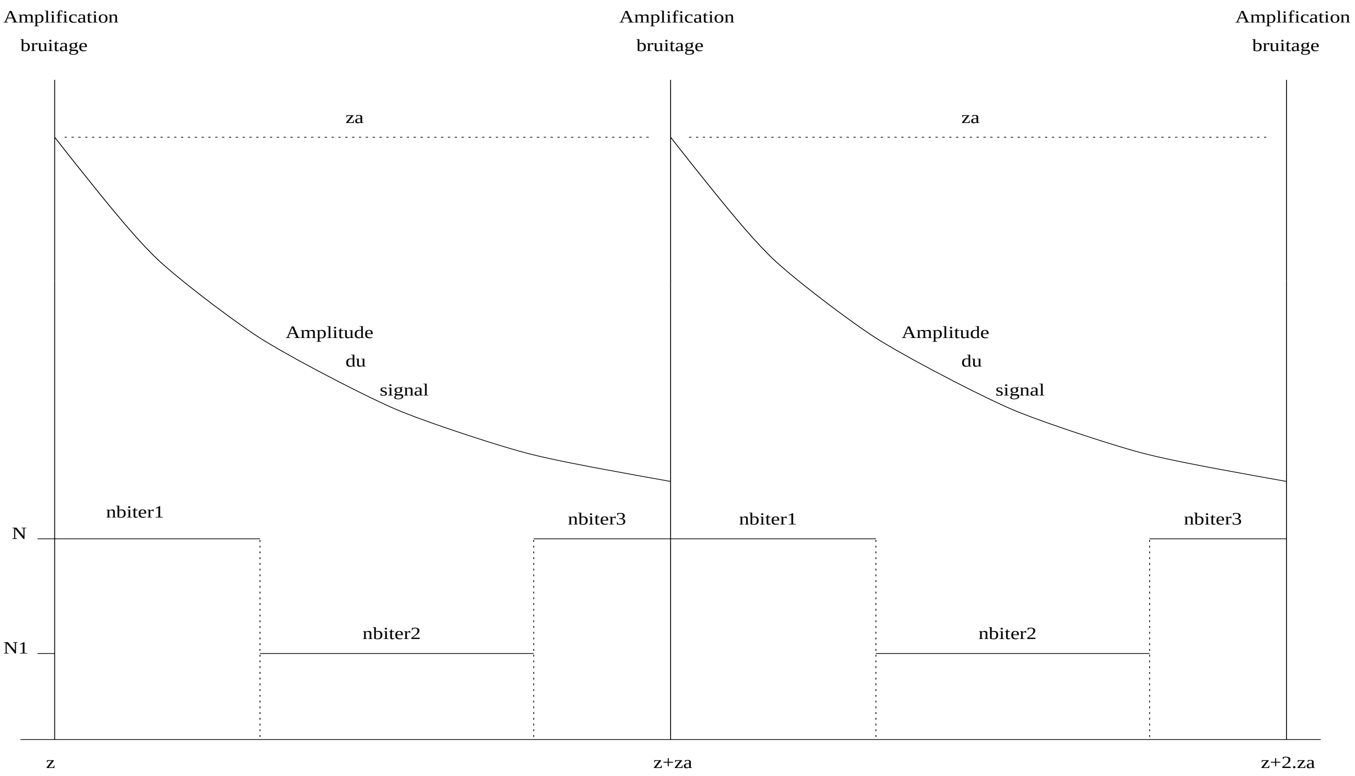
Les différentes simulations qui suivent sont entièrement basées sur le choix des cinq paramètres suivants :

N_1	:	la fréquence de coupure qui sépare u en $\mathbf{v}$ et $\mathbf{w}$;
nbiter1	:	le nombre d'itérations effectuées avec l'équation complète;
nbiter2	:	le nombre d'itérations effectuées avec la solution découplée en $\mathbf{v}$ et $\mathbf{w}$; Le
nbiter3	:	le nombre d'itérations effectuées avec la solution reconstituée;
V-cycle	:	le nombre de couples *nbiter2*, *nbiter3*.

but est bien évidemment de pouvoir effectuer le calcul de $z = 0$ à $z = 13$ au moindre coût calcul. Pour cela il est important que la part de l'étape 2 soit la plus grande possible. Par conséquent on peut envisager deux stratégies :
• On décide que le nombre de *V-cycle* sera égal à un, c'est-à-dire que la relation (2.58) s'écrit

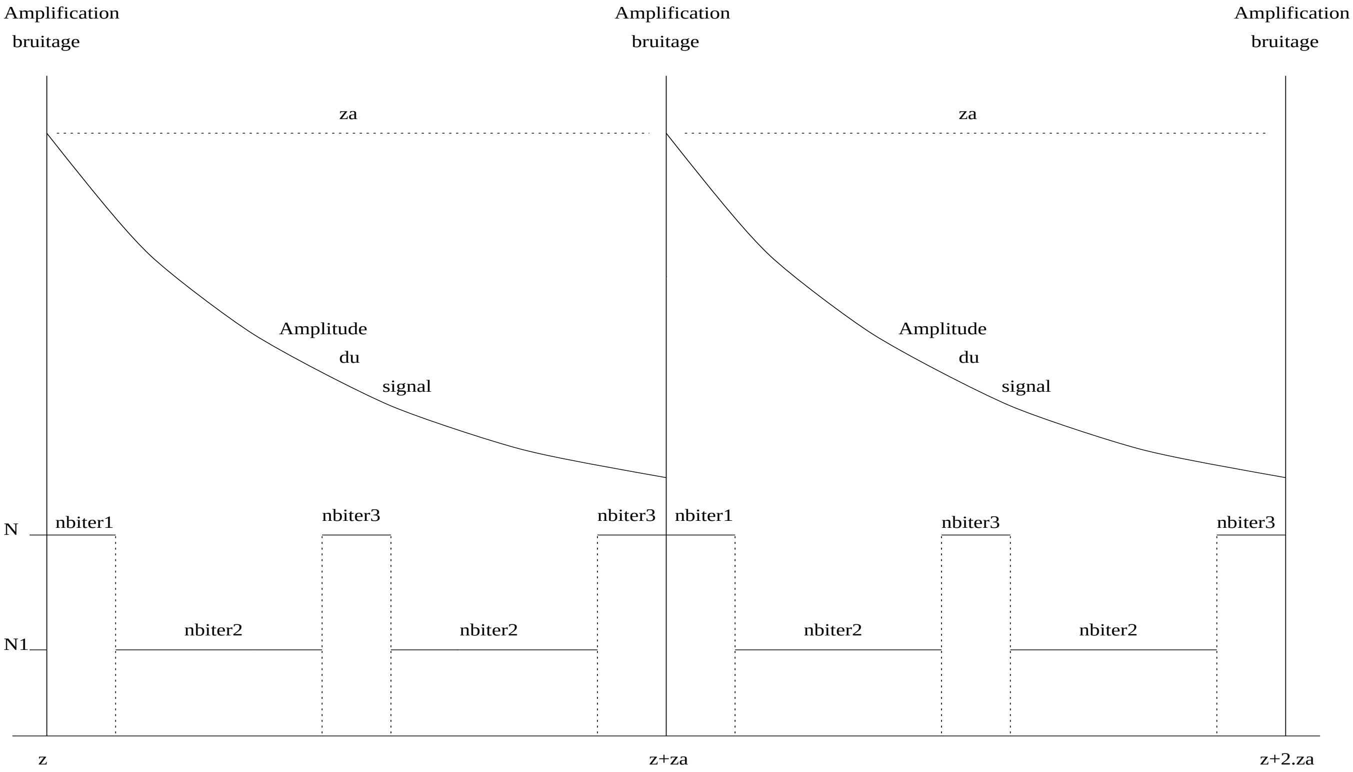
$$\textit{nbiter 1} \; + \; \textit{nbiter 2} \; + \; \textit{nbiter 3} \; = \frac{z_a}{\Delta z}$$

FIG. 2.15 – *Description d'un V-cycle.*



Pour que le gain soit significatif il faut que l'étape 2 soit très longue. Cela implique intégrer très longtemps l'équation des $\mathbf{v}$. Il faudra donc faire attention à ce que cela n'entraine pas d'erreur trop importante pour la solution complète u.

FIG. 2.16 – *Description d'un V-cycle.*



• On décide que le nombre de *V-cycle* est supérieur à un. On a alors plusieurs découplages et reformations dans un même cycle. Il ne faut pas alors en faire de trop pour que le gain en temps calcul réalisé grâce à l'étape 2 ne soit pas perdu par les opérations annexes de découplage et de reformation. En général, on prendra ce nombre égal à deux voire plus rarement trois.

2.9.5 Présentation des résultats.

Les tests ont été effectués avec différents niveaux de séparation N_1 : 16384, 8192, 4096, 2048 et même 1024.
Nous fixons Δz à 10^{-3} pour le moment : cela correspond à 40 itérations par cycle donc nous disposons d'une grande liberté dans le choix des combinaisons des paramètres décrivant la nouvelle méthode.
Par la suite nous augmenterons Δz jusqu'à la valeur 10^{-2} en prenant soin de choisir toujours des sous-multiples de z_a. Pour ces autres valeurs, cela permet d'avancer plus vite, mais en contrepartie nous perdons la grande liberté de choix dans les paramètres puisque pour $\Delta z = 10^{-2}$, le cycle est parcouru en 4 itérations seulement ce qui ne laisse que très peu de combinaisons avec l'étape 2 la plus longue possible sachant que les étapes 1 et 3 doivent compter au moins une itération.

Les niveaux 16384, 8192 ont donné des résultats qualitativement suffisants mais le gain en temps calcul n'était pas très important. D'autre part, des calculs ultérieurs avec des valeurs de N_1 inférieures aux deux valeurs précédentes ont fourni des résultats de qualité équivalente à moindre coût calcul.
Nous présentons donc ici des résultats obtenu uniquement pour les valeurs de N_1 égales à 4096, 2048 et $N_1 = 1024$.
La colonne intitulée *V-cycle* donne pour chaque calcul les valeurs dans l'ordre des quatre paramètres *nombre de V-cycle*, *nbiter1*, *nbiter2* et *nbiter3*.
Nous commençons par $N_1 = 4096$ et $N_1 = 2048$.

TAB. 2.12 – *Résultats obtenus avec la méthode (MLSSA) pour* $N_1 = 4096, 2048$ *et* $\Delta z = 10^{-3}$ *(clé1).*

	$N_1 = 4096$			$N_1 = 2048$		
V-cycle	Q_t	Q	Temps CPU	Q_t	Q	Temps CPU
(3,4,10,2)	20.8062	15.8087	2856 s	20.7506	15.9000	2593 s
(2,6,12,5)	20.8351	15.8274	3531 s	20.8021	15.8870	3450 s
(2,4,14,4)	20.8312	15.8185	2969 s	20.7808	15.9082	2900 s
(2,2,16,3)	20.7983	15.8322	2438 s	20.7310	15.9324	2254 s
(2,2,18,1)	20.7740	15.8333	1921 s	20.6966	15.9391	1616 s

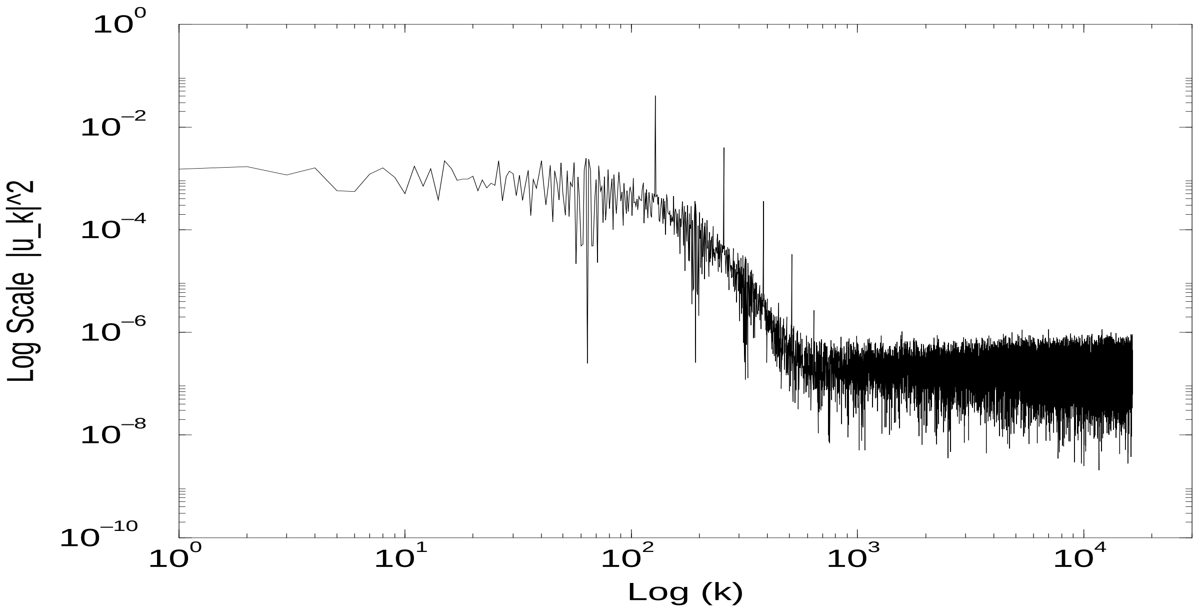
Avec $N_1 = 1024$, nous obtenons

TAB. 2.13 – *Résultats obtenus avec la méthode (MLSSA) pour* $N_1 = 1024$ *et* $\Delta z = 10^{-3}$ *(clé1).*

V-cycle	Q_t	Q	Temps CPU
(3,4,10,2)	47.4593	15.5150	2578 s
(2,6,12,5)	45.3512	15.5928	3394 s
(2,4,14,4)	47.1350	15.5542	2709 s
(2,2,16,3)	47.9079	15.4719	2086 s
(2,2,18,1)	47.9409	15.3749	1418 s

Les valeurs des facteurs de qualité pour $N_1 = 1024$ (en particulier Q_t) sont largement supérieures à celles obtenues pour $N_1 = 4096$ ou 2048. Si nous considérons le spectre de Fourier de la solution (figure 2.17), nous remarquons que $N_1 = 1024$ est placée entre la décroissance du spectre et la zone plane du spectre dominée par le bruit. Les effets du bruit blanc sont bien pris en compte dans la partie figée du terme non linéaire mais il semble que $N_1 = 1024$ agisse comme un filtre affaiblissant l'action du bruit. Nous conservons cette fréquence de coupure mais le niveau de référence pour estimer l'amélioration apportée par cette méthode sera plutôt $N_1 = 4096$ ou 2048 que 1024.
Prendre une valeur plus petite (comme par exemple $N_1 = 512$) nous placerait dans la partie du spectre correspondant à la solution mais où l'intensité du bruit est très petite par rapport au module des modes du spectre : le découplage reviendrait alors à résoudre linéairement une partie de la solution, ce qui entrainerait de grosses erreurs d'approximation.

FIG. 2.17 – *Spectre d'énergie de la solution $u(t, z)$.*



Jusqu'à présent, nous faisions toujours 2 *V-cycles* par cycle. Pour ces valeurs de N_1, nous essayons d'en faire un seul en augmentant peu à peu la part des itérations effectuées dans l'étape 2. Les résultats sont satisfaisants qualitativement et en temps calcul. Nous pouvons ainsi passer près de 90% des itérations formant un cycle avec $N_1 = 4096$, 2048 ou 1024 points pour résoudre l'équation (NLS) des $\mathbf{v}$. Les $\mathbf{w}$ sont résolus juste avant la première itération de l'étape 2.

TAB. 2.14 – *Résultats obtenus avec la méthode (MLSSA) pour $N_1 = 4096$, 2048 et $\Delta z = 10^{-3}$ (clé1).*

	$N_1 = 4096$			$N_1 = 2048$		
V-cycle	Q_t	Q	Temps CPU	Q_t	Q	Temps CPU
(1,9,22,9)	20.8013	15.8237	4003 s	20.7748	15.8639	3768 s
(1,8,24,8)	20.8119	15.8217	3743 s	20.7792	15.8789	3422 s
(1,7,26,7)	20.8182	15.8083	3438 s	20.7788	15.8828	3070 s
(1,6,28,6)	20.8243	15.8007	3137 s	20.7761	15.8925	2742 s
(1,5,30,5)	20.8232	15.7918	2831 s	20.7657	15.8958	2411 s
(1,4,32,4)	20.8263	15.7934	2533 s	20.7615	15.9013	2096 s
(1,3,34,3)	20.8157	15.8103	2231 s	20.7436	15.9183	1775 s
(1,2,36,2)	20.7991	15.8233	1929 s	20.7233	15.9259	1459 s
(1,1,38,1)	20.7756	15.8304	1595 s	20.6987	15.9325	1133 s

TAB. 2.15 – *Résultats obtenus avec la méthode (MLSSA) pour $N_1 = 1024$ et $\Delta z = 10^{-3}$ (clé1).*

V-cycle	Q_t	Q	Temps CPU
(1,9,22,9)	43.9714	15.5771	3244 s
(1,8,24,8)	45.2148	15.5757	2965 s
(1,7,26,7)	46.1789	15.5585	2651 s
(1,6,28,6)	46.9068	15.5368	2349 s
(1,5,30,5)	47.3758	15.5029	2042 s
(1,4,32,4)	47.6484	15.4629	1758 s
(1,3,34,3)	47.7091	15.4132	1450 s
(1,2,36,2)	47.6262	15.3531	1123 s
(1,1,38,1)	47.3615	15.2731	829 s

Comme pour la méthode *Split-Step Agrawal*, nous avons pris le pas d'espace compris entre 10^{-3} et 10^{-2} et les valeurs de N_1 précédentes à savoir 4096, 2048 et 1024. Pour les différents pas d'espace nous prenons la combinaison qui permette à l'étape 2 d'être la plus grande longue possible.
Nous commençons par $N_1 = 4096$.

TAB. 2.16 – *Résultats obtenus avec la méthode (MLSSA) pour $N_1 = 4096$ (clé1 et clé2).*

		clé1			clé2		
Δz	*V-cycle*	Q_t	Q	Temps CPU	Q_t	Q	Temps CPU
10^{-3}	(1,1,38,1)	20.7756	15.8304	1423 s	15.7472	13.7511	1418 s
$2\ 10^{-3}$	(1,1,18,1)	20.8329	15.8206	985 s	15.7304	13.7773	985 s
$4\ 10^{-3}$	(1,1,8,1)	20.7691	15.7147	767 s	15.6209	13.8411	766 s
$8\ 10^{-3}$	(1,1,3,1)	20.9388	15.6705	661 s	15.5460	13.9954	652 s
10^{-2}	(1,1,2,1)	20.8077	15.4705	641 s	15.5748	14.1707	631 s

Les résultats présentés dans ces tableaux correspondent aux 2 clés préalablement choisies, à savoir les clés *clé1* et *clé2* valant respectivement 2005 et 4500.
Avec $N_1 = 2048$, nous obtenons

TAB. 2.17 – *Résultats obtenus avec la méthode (MLSSA) pour $N_1 = 2048$ (clé1 et clé2).*

		clé1			clé2		
Δz	*V-cycle*	Q_t	Q	Temps CPU	Q_t	Q	Temps CPU
10^{-3}	(1,1,38,1)	20.6987	15.9325	1011 s	15.7277	13.7219	1018 s
$2\ 10^{-3}$	(1,1,18,1)	20.7569	15.9142	789 s	15.7197	13.7454	780 s
$4\ 10^{-3}$	(1,1,8,1)	20.7052	15.8200	756 s	15.6419	13.7700	667 s
$8\ 10^{-3}$	(1,1,3,1)	20.9366	15.6976	618 s	15.5836	13.9206	612 s
10^{-2}	(1,1,2,1)	20.8252	15.4421	606 s	15.6059	14.0798	600 s

Enfin, pour $N_1 = 1024$, cela donne

TAB. 2.18 – *Résultats obtenus avec la méthode (MLSSA) pour $N_1 = 1024$ (clé1 et clé2).*

		clé1			clé2		
Δz	*V-cycle*	Q_t	Q	Temps CPU	Q_t	Q	Temps CPU
10^{-3}	(1,1,38,1)	47.3615	15.2731	814 s	38.4756	14.6427	839 s
$2\ 10^{-3}$	(1,1,18,1)	47.3664	15.3523	697 s	38.5848	14.5975	699 s
$4\ 10^{-3}$	(1,1,8,1)	48.1714	15.4130	618 s	38.6401	14.4888	633 s
$8\ 10^{-3}$	(1,1,3,1)	46.6254	15.4451	591 s	36.7690	14.4986	590 s
10^{-2}	(1,1,2,1)	43.7295	15.2225	600 s	34.0451	14.6015	588 s

Nous pouvons faire un parallèle avec les résultats obtenus par la méthode *Split-Step Agrawal* pour ces mêmes clés.
Pour ces deux méthodes, les résultats obtenus avec la clé clé2 sont moins bons que ceux obtenus avec la clé clé1, les valeurs des facteurs de qualité enregistrent une baisse de 2 ou 5 unités. Cependant, nous remarquons que, malgré cette baisse, les valeurs obtenues pour la méthode modifiée sont que celles obtenues par la méthode *Split-Step Agrawal*.
Les facteurs de qualité des tableaux 2.16, 2.17 et 2.18 sont de l'ordre de grandeur souhaité avec des temps calcul inférieurs à ceux de la méthode de référence.
Pour $\Delta z = 10^{-3}$, nous avons 1011 s CPU par rapport à 6977 s pour la méthode de référence.
Pour $\Delta z = 10^{-2}$, nous avons 600 s par rapport à 800 s.
Soit un gain de 85 % pour le premier comparatif et 25 % pour le second.

2.10 Conclusion.

Dans le cadre de la modélisation de la transmission d'un signal par l'équation de Schroedinger non linéaire faiblement amortie (WDNLS), nous présentons une modification de la méthode *Split-Step Agrawal* qui permet un gain appréciable en temps CPU et en qualité de la transmission. Cette amélioration est fondée sur la décomposition dans l'espace de Fourier du spectre de la solution en deux parties :
- la première, notée $\mathbf{v}$ correspond à la partie principale du signal,
- la seconde, notée $\mathbf{w}$ correspond à la partie secondaire du signal, là où le bruit est relativement plus important en intensité que le signal lui-même.
La première est calculée par une projection de l'équation (NLS) agissant sur les premiers modes de la solution,
la seconde est calculée en considérant la projection sur la queue du spectre de l'équation de Schroedinger linéaire.
Les résultats, après analyses de la transmission, présentent une amélioration qualitative de la transmission doublée d'un gain en temps calcul.

Chapitre 3

Étude des équations de Maxwell.

3.1 Introduction

Dans ce chapitre, nous nous intéressons au traitement de conditions aux limites non périodiques pour des méthodes spectrales multi-niveaux. Nous considérons comme cadre de notre étude le problème d'électromagnétisme bidimensionnel appelé la *cavité résonnante*. Il s'agit d'un domaine $\Omega \subset \mathbb{R}^2$ d'une part où les champs électrique et magnétique ($\mathbf{E}$ et $\mathbf{H}$) ne dépendent pas de la troisième variable d'espace et d'autre part muni de conditions aux limites d'un conducteur parfait : des conditions aux limites de type Dirichlet homogène suivant la direction et l'inconnue considérées.
La discrétisation spatiale est effectuée par la méthode pseudo-spectrale Tau à l'aide des polynômes de Legendre, orthogonaux pour le produit scalaire usuel de l'espace $L^2(\Omega)$ (contrairement aux polynômes de Chebyshev).
Nous avions remarqué au chapitre 1 que pour les polynômes de Jacobi les opérateurs de projection ne commutent pas avec les opérateurs de dérivation spatiale. En projetant les équations de Maxwell ainsi semi-discrétisées en espace nous obtenons un système d'équations couplées qu'il nous faut résoudre, couplage renforcé par les conditions aux limites. En effet, pour la discrétisation spatiale considérée, l'imposition des conditions aux limites utilise toutes les valeurs de la solution et non pas seulement les valeurs aux bord du domaine.

3.2 Problème physique : équations de l'électromagnétisme.

Soient $\mathbf{x}$ la variable d'espace, $\mathbf{x} = (x, y, z) \in \mathbb{R}^d, d = 2$ ou 3 et t la variable temporelle. Nous introduisons les notations suivantes ([19]) :

• champ électromagnétique :

$\mathbf{E} = \mathbf{E}(\mathbf{x}, t)$; champ électrique
$\mathbf{D} = \mathbf{D}(\mathbf{x}, t)$; induction électrique
$\mathbf{H} = \mathbf{H}(\mathbf{x}, t)$; champ magnétique
$\mathbf{B} = \mathbf{B}(\mathbf{x}, t)$; induction magnétique

• charges et courants :

$\rho = \rho(\mathbf{x}, t)$; densité de charge électrique
$\mathbf{j} = \mathbf{j}(\mathbf{x}, t)$; densité de courant électrique

Une étude des phénomènes électromagnétiques consiste à déterminer les quatre champs de vecteurs $\mathbf{E}, \mathbf{B}, \mathbf{D}, \mathbf{H}$ vérifiant :
• le théorème d'Ampère, qui permet de calculer le champ magnétique engendré par un courant,
• la loi de Faraday, qui lie la force électromotrice à la variation de flux d'induction,
• la loi définissant la charge électrique,
• la loi de Gauss, postulant l'absence de charge magnétique.

On en déduit les expressions locales de ces lois physiques appelées équations de Maxwell :

$$\left\{\begin{array}{rcl} \dfrac{\partial \mathbf{D}}{\partial t} - \nabla \times \mathbf{H} & = & -\mathbf{j} \\ \dfrac{\partial \mathbf{B}}{\partial t} + \nabla \times \mathbf{E} & = & \mathbf{0} \\ div(\mathbf{D}) & = & \rho \\ div(\mathbf{B}) & = & 0 \end{array}\right. \tag{3.1}$$

On y ajoute la loi de conservation de la charge électrique

$$\frac{\partial \rho}{\partial t} + div(\mathbf{j}) = 0$$

Les relations entre champs et inductions sont données par des lois de comportement, caractéristiques du milieu considéré.
(i) proportionnalité de champs et inductions :
cette loi se traduit par les deux relations

$$\left\{\begin{array}{rcl} \mathbf{D} & = & \epsilon(\mathbf{x})\mathbf{E} \\ \mathbf{B} & = & \mu(\mathbf{x})\mathbf{H} \end{array}\right. \tag{3.2}$$

où $\epsilon(\mathbf{x})$ est la permittivité du vide
$\mu(\mathbf{x})$ est la perméabilité magnétique pour un milieu isotrope simple.
Ces scalaires sont supposés indépendants des phénomènes électromagnétiques dont le milieu est le siège. Ce sont des lois linéaires.

(ii) la loi d'Ohm en milieu conducteur et en l'absence de courant source

$$\mathbf{j} = \sigma \mathbf{E}$$

où σ est la conductivité.

A l'aide de la loi de conservation (3.2), on peut écrire les équations (3.1) sous forme conservative en variables $(\mathbf{D}, \mathbf{B})$.

$$\left\{\begin{array}{rcl} \dfrac{\partial \mathbf{D}}{\partial t} - \nabla \times \left(\dfrac{\mathbf{B}}{\mu(\mathbf{x})}\right) & = & -\mathbf{j} \\ \dfrac{\partial \mathbf{B}}{\partial t} + \nabla \times \left(\dfrac{\mathbf{D}}{\epsilon(\mathbf{x})}\right) & = & \mathbf{0} \\ div(\mathbf{D}) & = & \rho \\ div(\mathbf{B}) & = & 0 \end{array}\right. \tag{3.3}$$

La formulation en $(\mathbf{E},\mathbf{H})$ étant

$$\left\{\begin{array}{rrcl} \epsilon(\mathbf{x})\dfrac{\partial \mathbf{E}}{\partial t} & -\nabla\times\mathbf{H} & = & -\mathbf{j} \\ \mu(\mathbf{x})\dfrac{\partial \mathbf{H}}{\partial t} & +\nabla\times\mathbf{E} & = & \mathbf{0} \\ & div(\mu(\mathbf{x})\mathbf{H}) & = & \rho \\ & div(\epsilon(\mathbf{x})\mathbf{E}) & = & 0 \end{array}\right. \tag{3.4}$$

Dans le cas de milieux homogènes, les formulations (3.3) et (3.4) sont équivalentes.
Nous considérons l'étude du champ électromagnétique à l'intérieur d'une cavité Ω (domaine borné) $\Omega \subset \mathbb{R}^3$, occupée par le vide et limitée par un milieu considéré comme conducteur parfait.
Les conditions aux limites sont alors :

$$\left\{\begin{array}{rcl} \mathbf{E}\wedge\nu_{|_\Gamma} & = & \mathbf{0} \\ \mathbf{H}\cdot\nu_{|_\Gamma} & = & 0 \end{array}\right. \tag{3.5}$$

avec ν la normale extérieure à Γ, la frontière de Ω.
Physiquement, il s'agit d'un champ électromagnétique existant dans la cavité et ce champ évolue selon les équations de Maxwell. On suppose les phénomènes indépendants de la troisième dimension, il s'agit alors d'une cavité résonnante.

3.3 Problème mathématique étudié.

Nous choisissons d'étudier la formulation (3.4) en les variables $(\mathbf{E},\mathbf{H})$ avec les conditions aux limites d'un conducteur parfait (3.5).
Nous faisons maintenant des hypothèses sur le milieu de propagation.
On suppose celui-ci

- $\left.\begin{array}{l}\text{homogène}\\ \text{non polarisable}\\ \text{non magnétique}\end{array}\right\}$ ainsi $\epsilon(\mathbf{x})$ et $\mu(\mathbf{x})$ sont constantes et positives, soit $\epsilon(\mathbf{x})=\epsilon_0$ et $\mu(\mathbf{x})=\mu_0$.
- non conducteur : $\mathbf{j}=\mathbf{0}$ et $\rho=0$.

Sous ces hypothèses, le système (3.4) devient

$$\left\{\begin{array}{rcl} \dfrac{\partial\widetilde{\mathbf{E}}}{\partial\widetilde{t}} - \widetilde{\nabla}\times\left(\dfrac{\widetilde{\mathbf{H}}}{\epsilon_0}\right) & = & \mathbf{0} \\ \dfrac{\partial\widetilde{\mathbf{H}}}{\partial\widetilde{t}} + \widetilde{\nabla}\times\left(\dfrac{\widetilde{\mathbf{E}}}{\mu_0}\right) & = & \mathbf{0} \end{array}\right. \quad ; \ (\widetilde{\mathbf{x}},\widetilde{t})\in\widetilde{\Omega}\times\mathbb{R}_+ \tag{3.6}$$

$$+ \text{ conditions aux limites } \left\{\begin{array}{rcl} \widetilde{\mathbf{E}}\wedge\nu_{|_{\widetilde{\Gamma}}} & = & \mathbf{0} \\ \widetilde{\mathbf{H}}\cdot\nu_{|_{\widetilde{\Gamma}}} & = & 0 \end{array}\right. \qquad \widetilde{\Gamma}=\partial\widetilde{\Omega}$$

$$+ \text{ conditions initiales } \left\{\begin{array}{l} \widetilde{\mathbf{E}}(\mathbf{x},t=0)=\widetilde{\mathbf{E}_0} \\ \widetilde{\mathbf{H}}(\mathbf{x},t=0)=\widetilde{\mathbf{H}_0} \end{array}\right.$$

Les équations $div(\widetilde{\mathbf{H}})=0$ et $div(\widetilde{\mathbf{E}})=0$ sont en effet redondantes pour une condition initiale à divergence nulle.

Nous restreignons notre étude au cas où toutes les quantités ne dépendent que de x, y. Alors le système (3.6) peut se découpler en

$$\left\{\begin{array}{rcl} \dfrac{\partial \widetilde{\mathbf{E}_1}}{\partial \tilde{t}} - \widetilde{\mathbf{rot}}\left(\dfrac{\widetilde{H}_z}{\epsilon_0}\right) & = & \mathbf{0} \\ \dfrac{\partial \widetilde{H}_z}{\partial \tilde{t}} + \widetilde{rot}\left(\dfrac{\widetilde{\mathbf{E}_1}}{\mu_0}\right) & = & 0 \end{array}\right. \tag{3.7}$$

et

$$\left\{\begin{array}{rcl} \dfrac{\partial \widetilde{E}_z}{\partial \tilde{t}} - \widetilde{rot}\left(\dfrac{\widetilde{\mathbf{H}_1}}{\epsilon_0}\right) & = & 0 \\ \dfrac{\partial \widetilde{\mathbf{H}_1}}{\partial \tilde{t}} + \widetilde{\mathbf{rot}}\left(\dfrac{\widetilde{E}_z}{\mu_0}\right) & = & \mathbf{0} \end{array}\right. \tag{3.8}$$

$$\widetilde{\mathbf{E}_1} = \left\{\widetilde{E}_x, \widetilde{E}_y\right\} \quad ; \quad \widetilde{\mathbf{H}_1} = \left\{\widetilde{H}_x, \widetilde{H}_y\right\}$$

où $\widetilde{rot}(\mathbf{v}) = \dfrac{\partial v_y}{\partial \tilde{x}} - \dfrac{\partial v_x}{\partial \tilde{y}}$ pour $\mathbf{v} = \{\tilde{v}_x, \tilde{v}_y\}$

$$\widetilde{\mathbf{rot}}(\phi) = \left\{\frac{\partial \phi}{\partial \tilde{y}}, -\frac{\partial \phi}{\partial \tilde{x}}\right\}$$

Nous limiterons notre étude au système (3.7).
Dans la suite, $\widetilde{\mathbf{E}_1}$ sera noté $\widetilde{\mathbf{E}}$.

Ainsi nous considérons le problème suivant :

$$\left\{\begin{array}{rcl} \dfrac{\partial \widetilde{\mathbf{E}}}{\partial \tilde{t}} - \widetilde{\mathbf{rot}}\left(\dfrac{\widetilde{H}_z}{\epsilon_0}\right) & = & \mathbf{0} \\ \dfrac{\partial \widetilde{H}_z}{\partial \tilde{t}} + \widetilde{rot}\left(\dfrac{\widetilde{\mathbf{E}}}{\mu_0}\right) & = & 0 \end{array}\right. \quad \text{dans } \widetilde{\Omega} \times \mathbb{R}_+ \ \text{ avec } \ \widetilde{\Omega} = (-L, +L) \times (-L, +L) \ \text{ et } L > 0$$

On effectue le changement de variables suivant :

$$x = \frac{\tilde{x}}{L} \ ; \ y = \frac{\tilde{y}}{L} \ ; \ t = \frac{c\tilde{t}}{L} \text{ avec } c = \frac{1}{\sqrt{\epsilon_0 \mu_0}} \ ; \ H_z = \sqrt{\frac{\epsilon_0}{\mu_0}} \widetilde{H}_z \ ; \ E_x = \widetilde{E}_x \ ; \ E_y = \widetilde{E}_y$$

Ce qui donne

$$\left\{\begin{array}{rcl} \dfrac{\partial \mathbf{E}}{\partial t} - \mathbf{rot}\,(H_z) & = & \mathbf{0} \\ \dfrac{\partial H_z}{\partial t} + rot\,(\mathbf{E}) & = & 0 \end{array}\right. \quad \text{dans } \Omega \times \mathbb{R}_+ \tag{3.9}$$

On note dorénavant

$$\mathbf{U}(t) = \{\mathbf{E}(t); H_z(t)\} \ ; \ \mathcal{A} = \begin{pmatrix} 0 & \mathbf{rot} \\ rot & 0 \end{pmatrix}$$

Le système (3.9) s'écrit sous la forme classique suivante

$$\left\{\begin{array}{l} \dfrac{\partial \mathbf{U}}{\partial t} + \mathcal{A}\mathbf{U} = \mathbf{0} \\ \mathbf{U}(0) = \mathbf{U}_0 \end{array}\right. \tag{3.10}$$

Nous prenons comme conditions aux limites :

$$\mathbf{E} \wedge \nu_{|_\Gamma} = \mathbf{0}$$

avec ν la normale extérieure à Γ, la frontière de Ω.
En effet, nous avons intrinsèquement

$$\mathbf{H} \cdot \nu_{|_\Gamma} = 0 \quad \text{pour } \mathbf{H} = \begin{pmatrix} 0 \\ 0 \\ H_z \end{pmatrix}$$

3.3.1 Existence et unicité des solutions.

Choisissons $D(\mathcal{A})$ et $\mathcal{H}$ tels que $\mathcal{A}$ soit une application de $D(\mathcal{A})$ dans $\mathcal{H}$.

$$\mathcal{H} = \mathbf{H}_0(div^0, \Omega) \times L^2(\Omega)$$

avec
$$\begin{array}{rcl} \mathbf{H}_0(div^0, \Omega) & = & \{\mathbf{v} \in \mathbf{H}_0(div, \Omega) / div\, \mathbf{v} = 0\} \\ \mathbf{H}_0(div, \Omega) & = & \left\{\mathbf{v} \in (L^2(\Omega))^2 / div\, \mathbf{v} \in L^2(\Omega), \mathbf{v} \wedge \nu_{|_\Gamma} = 0\right\}; \end{array}$$

et
$$\begin{array}{rcl} D(\mathcal{A}) & = & \mathbf{D} \times H^1(\Omega) \\ \mathbf{D} & = & \left\{\mathbf{v} \in (H^1(\Omega))^2 / div\, \mathbf{v} = 0, \mathbf{v} \wedge \nu_{|_\Gamma} = 0\right\} \end{array}$$

Pour le problème de Cauchy du système (3.10), on peut facilement vérifier que l'opérateur $-\mathcal{A}$ est le générateur infinitésimal d'un semi-groupe de contraction $\mathcal{S}(t)$ et ainsi le problème (3.10) possède une unique solution $\mathbf{U} \in \mathcal{C}^1(0, T; \mathcal{H}) \cap \mathcal{C}^0(0, T; D(\mathcal{A}))$ donnée par $\mathbf{U}(t) = \mathcal{S}(t)\mathbf{U_0}$ (voir [1, 19]).

3.3.2 Formulation variationnelle.

Soit $\mathcal{V} = \mathbf{H}_0(div^0, \Omega) \times H^1(\Omega)$.
On note $((.,.))$ le produit scalaire usuel de l'espace $(L^2(\Omega))^3$.
La formulation variationnelle du problème est :

$$\left(\left(\frac{\partial \mathbf{U}}{\partial t}, \mathbf{U}^*\right)\right) + a\,(\mathbf{U}(t), \mathbf{U}^*) = 0 \quad \forall \mathbf{U} \in \mathcal{V}, \qquad \mathbf{U}(0) = \mathbf{U_0} \tag{3.11}$$

où $a(.,.)$ est la forme bilinéaire associée à l'opérateur $\mathcal{A}$:

$$a\,(\mathbf{U}, \mathbf{U}^*) = \int_\Omega rot\,(\mathbf{E})\; {H_z}^*\, \mathrm{d}\Omega - \int_\Omega \mathbf{rot}\,(H_z)\; \mathbf{E}^*\, \mathrm{d}\Omega$$

pour tout $\mathbf{U} = \{H_z, \mathbf{E}\}$, $\mathbf{U}^* = \{{H_z}^*, \mathbf{E}^*\} \in \mathcal{V}$.

3.3.3 Invariant.

Montrons que le problème continu (3.10) est bien posé. Nous appliquons la méthode de l'énergie à

$$\frac{\partial \mathbf{U}}{\partial t} + \mathcal{A}\mathbf{U} = \mathbf{0}$$

Alors

$$\frac{1}{2}\frac{d}{dt} \parallel \mathbf{U} \parallel^2 = \left(\left(\frac{\partial \mathbf{U}}{\partial t}, \mathbf{U}\right)\right) = -\left(\left(\mathcal{A}U, \mathbf{U}\right)\right)$$

On développe le dernier terme :

$$\begin{aligned}
((\mathcal{A}U, \mathbf{U})) &= \left(-\frac{\partial U_3}{\partial y}, U_1\right) + \left(\frac{\partial U_3}{\partial x}, U_2\right) + \left(-\frac{\partial U_1}{\partial y} + \frac{\partial U_2}{\partial x}, U_3\right) \\
&= -\int_\Omega \frac{\partial U_1 U_3}{\partial y}\, \mathrm{d}\Omega + \int_\Omega \frac{\partial U_2 U_3}{\partial x}\, \mathrm{d}\Omega \\
&= \int_\Omega \mathbf{d}iv \begin{pmatrix} U_2 U_3 \\ -U_1 U_3 \end{pmatrix} \mathrm{d}\Omega \\
&= \int_{x=\pm 1} U_2 U_3\, \mathrm{d}\Omega - \int_{y=\pm 1} U_1 U_3\, \mathrm{d}\Omega \\
&= 0
\end{aligned}$$

grâce aux conditions aux limites : $U_1(x, y = \pm 1, t) = U_2(x = \pm 1, y, t) = 0$.
Donc le problème continu est bien posé.
D'autre part,

$$\frac{d}{dt} \parallel \mathbf{U} \parallel^2 = 0$$

entraine

$$\parallel \mathbf{U}(t) \parallel^2 = \parallel \mathbf{U}(0) \parallel^2 = \parallel \mathbf{U}_0 \parallel^2$$

Ainsi la norme $L^2(\Omega)$ de la solution est un invariant du problème continu.

3.3.4 Solution analytique.

Les équations de Maxwell sous la forme précédente possède des solutions analytiques qui, outre les conditions aux limites, vérifient aussi la condition de divergence nulle :

$$\frac{\partial U_1}{\partial x} + \frac{\partial U_2}{\partial y} = 0$$

Notre choix se porte sur

$$\left\{\begin{array}{lll}
u_1(x,y,t) & = & \lambda_1 \cos(\omega_1 t) \quad \cos(k_1 \pi x) \quad \sin(k_2 \pi y) \\
u_2(x,y,t) & = & \lambda_2 \cos(\omega_1 t) \quad \sin(k_1 \pi x) \quad \cos(k_2 \pi y) \\
u_3(x,y,t) & = & \lambda_3 \sin(\omega_1 t) \quad \cos(k_1 \pi x) \quad \cos(k_2 \pi y)
\end{array}\right.$$

Les scalaires λ_1, λ_2, λ_3, le vecteur d'onde $\mathbf{k} = (k_1, k_2)^{\mathrm{T}}$ et la pulsation ω_1 sont déterminés de la manière suivante :
Pour $\mathbf{k}$ donné la relation de dispersion suivante permet de calculer ω_1 :

$${\omega_1}^2 = \pi^2 \,|\mathbf{k}|^2$$

On fixe λ_3, on détermine alors les scalaires λ_1 et λ_2 par

$$\left\{\begin{array}{lll}
\lambda_1 & = & \dfrac{\pi}{\omega_1} k_2 \lambda_3 \\
 & & \\
\lambda_2 & = & -\dfrac{\pi}{\omega_1} k_1 \lambda_3
\end{array}\right.$$

Cependant cette solution exacte possède des propriétés de parité selon la composante u_1, u_2, u_3 et la variable spatiale x, y considérées. Cela se reflète dans le spectre obtenu après transformation dans l'espace spectral. Pour demeurer dans un cadre d'étude plus général, nous allons former une solution analytique qui ne possède aucune propriété de cette nature. Nous utilisons le fait que notre problème est linéaire et homogène et ainsi une combinaison linéaire de solutions est encore solution. Ainsi, pour $\mathbf{k} = (k_1, k_2)^{\mathrm{T}}$ donné, le triplet u_4, u_5, u_6 suivant est aussi solution

$$\left\{\begin{array}{lcl} u_4(x,y,t) & = & \lambda_4 \cos(\omega_2 t) \quad \sin((k_1+\frac{1}{2})\pi x) \quad \cos((k_2+\frac{1}{2})\pi y) \\ u_5(x,y,t) & = & \lambda_5 \cos(\omega_2 t) \quad \cos((k_1+\frac{1}{2})\pi x) \quad \sin((k_2+\frac{1}{2})\pi y) \\ u_6(x,y,t) & = & \lambda_6 \sin(\omega_2 t) \quad \sin((k_1+\frac{1}{2})\pi x) \quad \sin((k_2+\frac{1}{2})\pi y) \end{array}\right.$$

On fixe λ_6, on détermine alors les scalaires λ_4 et λ_5 par

$$\left\{\begin{array}{lcl} \lambda_4 & = & \dfrac{\pi}{\omega_2}(k_2+\dfrac{1}{2})\lambda_6 \\ \lambda_5 & = & -\dfrac{\pi}{\omega_2}(k_1+\dfrac{1}{2})\lambda_6 \end{array}\right.$$

avec la pulsation

$${\omega_2}^2 = \pi^2\left((k_1+\frac{1}{2})^2 + (k_2+\frac{1}{2})^2\right)$$

Une troisième solution est

$$\left\{\begin{array}{lcl} u_7(x,y,t) & = & \lambda_7 \cos(\omega_3 t) \quad \cos(k_1\pi x) \quad \cos((k_2+\frac{1}{2})\pi y) \\ u_8(x,y,t) & = & \lambda_8 \cos(\omega_3 t) \quad \sin(k_1\pi x) \quad \sin((k_2+\frac{1}{2})\pi y) \\ u_9(x,y,t) & = & \lambda_9 \sin(\omega_3 t) \quad \cos(k_1\pi x) \quad \sin((k_2+\frac{1}{2})\pi y) \end{array}\right.$$

On fixe λ_9, on détermine alors les scalaires λ_7 et λ_8 par

$$\left\{\begin{array}{lcl} \lambda_7 & = & \dfrac{\pi}{\omega_3}(k_2+\dfrac{1}{2})\lambda_9 \\ \lambda_8 & = & -\dfrac{\pi}{\omega_3}k_1\lambda_9 \end{array}\right.$$

avec la pulsation associée

$${\omega_3}^2 = \pi^2\left(k_1^2 + (k_2+\frac{1}{2})^2\right)$$

Enfin une quatrième solution est

$$\left\{\begin{array}{lcl} u_{10}(x,y,t) & = & \lambda_{10} \cos(\omega_4 t) \quad \sin((k_1+\frac{1}{2})\pi x) \quad \sin(k_2\pi y) \\ u_{11}(x,y,t) & = & \lambda_{11} \cos(\omega_4 t) \quad \cos((k_1+\frac{1}{2})\pi x) \quad \cos(k_2\pi y) \\ u_{12}(x,y,t) & = & \lambda_{12} \sin(\omega_4 t) \quad \sin((k_1+\frac{1}{2})\pi x) \quad \cos(k_2\pi y) \end{array}\right.$$

On fixe λ_{12}, on détermine alors les scalaires λ_{10} et λ_{11} par

$$\left\{\begin{array}{rcl}\lambda_{10} & = & \dfrac{\pi}{\omega_4}k_2\lambda_{12} \\ \\ \lambda_{11} & = & -\dfrac{\pi}{\omega_4}(k_1+\dfrac{1}{2})\lambda_{12}\end{array}\right.$$

avec la pulsation

$$\omega_4{}^2 = \pi^2\left((k_1+\frac{1}{2})^2+k_2{}^2\right)$$

Finalement notre solution analytique est

$$\left\{\begin{array}{rcl}U_1(x,y,t) & = & (u_1+u_4+u_7+u_{10})\,(x,y,t) \\ U_2(x,y,t) & = & (u_2+u_5+u_8+u_{11})\,(x,y,t) \\ U_3(x,y,t) & = & (u_3+u_6+u_9+u_{12})\,(x,y,t)\end{array}\right. \tag{3.12}$$

Avec le vecteur d'onde $\mathbf{k}=(k_1,k_2)^{\mathrm{T}}$ et les scalaires $\lambda_3,\lambda_6,\lambda_9,\lambda_{12}$ nous déterminons tous les paramètres définissant cette solution analytique du problème considéré.
Nous remarquons que les composantes de la solution exacte sont des fonctions très régulières, la discrétisation spatiale par méthode spectrale est très intéressante.
De nombreuses études numériques ont été menées dans le cadre des équations de Maxwell. Citons par exemple ([52, 51, 4]).

3.4 Méthode classique.

Nous présentons maintenant les discrétisations spatiale et temporelle qui constitueront la méthode de référence dite méthode classique.

3.4.1 Discrétisation spatiale.

Les conditions aux limites de type Dirichlet homogène nous font choisir les polynômes de Legendre comme base polynomiale de décomposition des inconnues.
Bien qu'il n'existe pas de Transformée Rapide de Legendre pour passer de l'espace physique dans l'espace spectral, on peut construire une transformée à partir d'une Transformée de Chebyshev Rapide suivie d'un changement de base polynomiale ([3, 42]).
D'autre part, les polynômes de Legendre sont orthogonaux pour le produit scalaire usuel de $L^2(\Omega)$, ce que nous mettrons à profit ultérieurement.

Pour imposer les conditions limites, nous choisissons la méthode Tau. Afin d'obtenir une discrétisation spatiale Tau-Legendre menant à un problème semi-discrétisé en espace bien posé, il faut prendre quelques précautions. Pour cela, nous commençons par étudier le problème 1D correspondant.
Soit le problème suivant défini pour $x\in[-1,1]$:

$$\left\{\begin{array}{rcl}\dfrac{\partial v}{\partial t} & = & -\dfrac{\partial w}{\partial x} \\ \dfrac{\partial w}{\partial t} & = & -\dfrac{\partial v}{\partial x} \\ v(x=\pm1,t)=0 & & \end{array}\right.$$

On montre facilement que le problème continu est bien posé. On pose $\mathbf{V} = (v, w)^{\mathrm{T}}$, alors :

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\,\|\mathbf{V}\|^2 &= \left(\frac{\partial v}{\partial t}, v\right) + \left(\frac{\partial w}{\partial t}, w\right) = -\left(\frac{\partial w}{\partial x}, v\right) - \left(\frac{\partial v}{\partial x}, w\right) \\ &= \int \frac{\partial(v\,w)}{\partial x}\,\mathrm{d}x = [v\,w]_{-1}^{+1} = 0 \quad \text{car } v(\pm 1, t) = 0. \end{aligned}$$

On applique la méthode spectrale Tau-Legendre, d'où pour un entier $N > 0$

$$v_N = P_N v = \sum_{k=0}^{N} \hat{v}_k(t) L_k(x)$$

$$w_N = P_N w = \sum_{k=0}^{N} \hat{w}_k(t) L_k(x)$$

On note P_N l'opérateur de projection orthogonale sur $\{L_0(x), ..., L_N(x)\}$ l'espace vectoriel des polynômes de degré inférieur ou égal à N en x.
De même pour $0 < N_1 < N$, on note Q_{N_1} l'opérateur de projection orthogonale sur $\{L_{N_1+1}(x), ..., L_N(x)\}$ et on a $Q_{N_1} = I_N \backslash P_{N_1}$, pour I_N l'opérateur identité dans $\{L_0(x), ..., L_N(x)\}$.
On applique la méthode de l'énergie :

$$\begin{aligned} I &= \frac{1}{2}\frac{d}{dt}\left\{|\,P_{N-2}v_N\,|^2 + |\,w_N\,|^2\right\} = \left(\frac{\partial P_{N-2}v_N}{\partial t}, P_{N-2}v_N\right) + \left(\frac{\partial w_N}{\partial t}, w_N\right) \\ &= -\left(P_{N-2}\frac{\partial w_N}{\partial x}, P_{N-2}v_N\right) - \left(\frac{\partial v_N}{\partial x}, w_N\right) \\ &= -\left([P_{N-2} + Q_{N-2} - Q_{N-2}]\,\frac{\partial w_N}{\partial x}, [P_{N-2} + Q_{N-2} - Q_{N-2}]\,v_N\right) - \left(\frac{\partial v_N}{\partial x}, w_N\right) \\ &= -\left(P_N\frac{\partial w_N}{\partial x}, P_N v_N\right) + \left(P_N\frac{\partial w_N}{\partial x}, Q_{N-2}v_N\right) + \left(Q_{N-2}\frac{\partial w_N}{\partial x}, P_N v_N\right) \\ &\quad - \left(Q_{N-2}\frac{\partial w_N}{\partial x}, Q_{N-2}v_N\right) - \left(\frac{\partial v_N}{\partial x}, w_N\right) \\ &= -\left(\frac{\partial w_N}{\partial x}, v_N\right) - \left(\frac{\partial v_N}{\partial x}, w_N\right) \\ &\quad + \left(\frac{\partial w_N}{\partial x}, Q_{N-2}v_N\right) + \left(Q_{N-2}\frac{\partial w_N}{\partial x}, v_N\right) - \left(Q_{N-2}\frac{\partial w_N}{\partial x}, Q_{N-2}v_N\right) \end{aligned}$$

Comme pour le problème continu, on a

$$\left(\frac{\partial w_N}{\partial x}, v_N\right) + \left(\frac{\partial v_N}{\partial x}, w_N\right) = 0$$

On calcule alors les trois intégrales restantes que l'on appelle resp. I_1, I_2, I_3. On utilise la notation suivante $\hat{u}^{(p)}$ désigne le spectre de Legendre de $\dfrac{\partial^p u}{\partial x^p}$.

$$\begin{aligned} I_1 = \int_{-1}^{1} \frac{\partial w_N}{\partial x}\,Q_{N-2}\,(v_N)\;\mathrm{d}x &= \sum_{k=0}^{N} \hat{w}_k^{(1)} \sum_{p=N-1}^{N} \hat{v}_p\,\frac{2}{2k+1}\delta_{k,p} \\ &= \frac{2}{2(N-1)+1}\hat{w}_{N-1}^{(1)}\hat{v}_{N-1} = 2\,\hat{w}_N\,\hat{v}_{N-1} \end{aligned}$$

$$\begin{aligned}
I_2 = \int_{-1}^{1} Q_{N-2}\left(\frac{\partial w_N}{\partial x}\right) v_N \,\mathrm{d}x &= \sum_{k=N-1}^{N} \hat{w}_k^{(1)} \sum_{p=0}^{N} \hat{v}_p \,\frac{2}{2k+1}\delta_{k,p} \\
&= \frac{2}{2(N-1)+1}\hat{w}_{N-1}^{(1)}\hat{v}_{N-1} = 2\,\hat{w}_N\,\hat{v}_{N-1}
\end{aligned}$$

$$\begin{aligned}
I_3 = \int_{-1}^{1} Q_{N-2}\left(\frac{\partial w_N}{\partial x}\right) Q_{N-2}(v_N)\,\mathrm{d}x &= \sum_{k=N-1}^{N} \hat{w}_k^{(1)} \sum_{p=N-1}^{N} \hat{v}_p \,\frac{2}{2k+1}\delta_{k,p} \\
&= \frac{2}{2(N-1)+1}\hat{w}_{N-1}^{(1)}\hat{v}_{N-1} = 2\,\hat{w}_N\,\hat{v}_{N-1}
\end{aligned}$$

D'où finalement :

$$I = I_1 + I_2 - I_3 = 2\,\hat{w}_N\,\hat{v}_{N-1}$$

On ne peut pas conclure car on ne connait rien du signe de $\hat{w}_N\hat{v}_{N-1}$.

En appliquant une idée du Professeur Gottlieb ([27]), nous allons montrer que nous pouvons obtenir une discrétisation spatiale Tau-Legendre bien posée.
Comme v_N et w_N sont solution du problème, alors l'équation

$$\frac{\partial v_N}{\partial t} = -\frac{\partial w_N}{\partial x}$$

est vérifiée pour les modes $k = 0, ..., (N-2)$. On peut donc l'écrire :

$$\frac{\partial v_N}{\partial t} + \frac{\partial w_N}{\partial x} = \tau_1 L_N(x) + \tau_2 L_{N-1}(x) \tag{3.13}$$

En effet, $\frac{\partial v_N}{\partial t}$ et $\frac{\partial w_N}{\partial x}$ sont des polynômes de degré resp. N et $N-1$ en x. On prend le produit scalaire de (3.13) avec $L_k(x)$ pour $0 \leqslant k \leqslant N$:

$$\left(\frac{\partial v_N}{\partial t}, L_k\right) + \left(\frac{\partial w_N}{\partial x}, L_k\right) = (\tau_1 L_N, L_k) + (\tau_2 L_{N-1}, L_k)$$

$$\Rightarrow \quad \frac{d\hat{v}_k}{dt} + \hat{w}_k^{(1)} = \tau_1\delta_{k,N} + \tau_2\delta_{k,N-1}$$

La relation de dérivation dans l'espace spectral (1.38) :

$$\hat{w}_k^{(1)} = (2k+1) \sum_{\substack{p=k+1 \\ p+k \text{ impair}}}^{N} \hat{w}_p$$

entraine

$$\hat{w}_N^{(1)} = 0 \quad \text{et} \quad \hat{w}_{N-1}^{(1)} = (2N-1)\hat{w}_N$$

Pour $k = N$, on a $\frac{d\hat{v}_N}{dt} = \tau_1$

Pour $k = (N-1)$, on a $\frac{d\hat{v}_{N-1}}{dt} + \hat{w}_{N-1}^{(1)} = \tau_2$; i.e. $\frac{d\hat{v}_{N-1}}{dt} + (2N-1)\hat{w}_N = \tau_2$.

Il reste à déterminer $\hat{w}_N$; de l'équation

$$\frac{\partial w_N}{\partial t} = -\frac{\partial v_N}{\partial x}$$

on extrait $\frac{d\hat{w}_N}{dt} = -\hat{v}_N^{(1)} = 0$, soit $\hat{w}_N(t) = \hat{w}_N(0) \ \ \forall t > 0$.
On peut alors choisir $\hat{w}_N = 0$, i.e. $\tau_2 = \frac{d\hat{v}_{N-1}}{dt}$.
On a ainsi obtenu $\tau_1 = \frac{d\hat{v}_N}{dt}$ et $\tau_2 = \frac{d\hat{v}_{N-1}}{dt}$.

On applique la méthode de l'énergie :

$$\begin{aligned}
\frac{1}{2}\ \frac{d}{dt}\int_{-1}^{1} v_N^2 + w_N^2 \,\mathrm{d}x &= \left(\frac{\partial v_N}{\partial t}, v_N\right) + \left(\frac{\partial w_N}{\partial t}, w_N\right) \\
&= -\left(\frac{\partial w_N}{\partial x}, v_N\right) - \left(\frac{\partial v_N}{\partial x}, w_N\right) + (\tau_1 L_N, v_N) + (\tau_2 L_{N-1}, v_N) \\
&= \frac{2}{2N+1}\,\hat{v}_N\,\frac{d\hat{v}_N}{dt} + \frac{2}{2N-1}\,\hat{v}_{N-1}\,\frac{d\hat{v}_{N-1}}{dt} \qquad \text{grâce aux conditions limites}
\end{aligned}$$

Comme

$$\frac{1}{2}\frac{d}{dt}\int_{-1}^{1} v_N^2 + w_N^2 \,\mathrm{d}x = \frac{1}{2}\frac{d}{dt}\left\{\sum_{n=0}^{N}\frac{2}{2n+1}\,\hat{v}_n^2 + \sum_{m=0}^{N}\frac{2}{2m+1}\,\hat{w}_m^2\right\}$$

Alors

$$\begin{aligned}
\frac{d}{dt}\left\{\sum_{n=0}^{N-2}\frac{2}{2n+1}\,\hat{v}_n^2 + \sum_{m=0}^{N}\frac{2}{2m+1}\,\hat{w}_m^2\right\} &= 0 \\
\Rightarrow \quad \frac{d}{dt}\left\{|\,P_{N-2}v_N\,|^2 + |\,w_N\,|^2\right\} &= 0 \\
\Rightarrow \quad |\,P_{N-2}v_N\,|^2 + |\,w_N\,|^2 &= \text{cste} \\
\text{i.e. } \ \|\,P_{N-2}\mathbf{V}\,\| &= \text{cste}
\end{aligned}$$

Le problème monodimensionnel semi-discrétisé en espace est bien posé. Nous appliquons maintenant cette idée pour le problème 2D qui nous intéresse.

Nous introduisons dès à présent des notations qui nous seront fort utiles pour la description des méthodes multi-niveaux.

$$\mathbf{U} = \begin{pmatrix} E_x \\ E_y \\ H_z \end{pmatrix} \ \text{ sera notée } \ \mathbf{U} = \begin{pmatrix} U_1 \\ U_2 \\ U_3 \end{pmatrix} \ \text{ plus maniable.}$$

On définit les espaces polynômiaux suivants pour N et M deux entiers strictement positifs.
$\mathcal{L}_N^1 = \{L_0(x), ..., L_N(x)\}$: espace vectoriel des polynômes de degré inférieur ou égal à N en x.
$\mathcal{L}_M^2 = \{L_0(y), ..., L_M(y)\}$: espace vectoriel des polynômes de degré inférieur ou égal à M en y.
où $L_p(\sigma)$ désigne le $p^{\text{ème}}$ polynôme de Legendre en la variable σ.

$$\begin{array}{lll} \text{Soit} \quad \mathcal{L}_{N,M} & = & \mathcal{L}_N^1 \otimes \mathcal{L}_M^2 \\ \text{i.e.} \quad \mathcal{L}_{N,M} & = & \{\Psi_{p,q}(x,y); 0 \leqslant p \leqslant N \text{ et } 0 \leqslant q \leqslant M\} \\ \text{avec} \quad \Psi_{p,q}(x,y) & = & L_p(x)L_q(y) \end{array}$$

Pour alléger les notations, nous ne précisons plus les indices N et M :

$$\mathbf{U}_{N,M} = \begin{pmatrix} U_{1,N,M} \\ U_{2,N,M} \\ U_{3,N,M} \end{pmatrix} = \begin{pmatrix} U_1 \\ U_2 \\ U_3 \end{pmatrix}$$

La solution $\mathbf{U}(x,y,t)$ de l'approximation Tau-Legendre ([26, 13]) de ce problème est pour tout $t \geqslant 0$ un vecteur, dont chacune des trois composantes est un élément de $\mathcal{L}_{N,M}$, vérifiant les conditions aux limites $U_1(x, y = \pm 1, t) = U_2(x = \pm 1, y, t) = 0$ et satisfaisant le système d'équations suivant

$$\left\{ \begin{array}{lll} \left(\left(\dfrac{\partial U_1}{\partial t} - \dfrac{\partial U_3}{\partial y}, f_1\right)\right) = 0 & \forall f_1 = \Phi_p\Phi_q \text{ avec } 0 \leqslant p \leqslant N-1 & \text{et } 0 \leqslant q \leqslant M-2 \\ \left(\left(\dfrac{\partial U_2}{\partial t} + \dfrac{\partial U_3}{\partial x}, f_2\right)\right) = 0 & \forall f_2 = \Phi_p\Phi_q \text{ avec } 0 \leqslant p \leqslant N-2 & \text{et } 0 \leqslant q \leqslant M-1 \\ \left(\left(\dfrac{\partial U_3}{\partial t} - \dfrac{\partial U_1}{\partial y} + \dfrac{\partial U_2}{\partial x}, f_3\right)\right) = 0 & \forall f_3 = \Phi_p\Phi_q \text{ avec } 0 \leqslant p \leqslant N-1 & \text{et } 0 \leqslant q \leqslant M-1 \end{array} \right.$$

où $\Phi_p \equiv \dfrac{2}{2p+1} L_p(x) \quad ; \quad \Phi_q \equiv \dfrac{2}{2q+1} L_p(y)$

$$\text{et on a} \left\{ \begin{array}{l} U_1(x,y,t) = \displaystyle\sum_{k=0}^{N-1} \sum_{j=0}^{M} \hat{u}_1(k,j,t) L_k(x) L_j(y) \\ U_2(x,y,t) = \displaystyle\sum_{k=0}^{N} \sum_{j=0}^{M-1} \hat{u}_2(k,j,t) L_k(x) L_j(y) \\ U_3(x,y,t) = \displaystyle\sum_{k=0}^{N-1} \sum_{j=0}^{M-1} \hat{u}_3(k,j,t) L_k(x) L_j(y) \end{array} \right.$$

Les conditions aux limites $U_1(x, y = \pm 1, t) = U_2(x = \pm 1, y, t) = 0$ sont imposées à l'aide des derniers modes.
$U_1(x, y = \pm 1, t) = 0$ s'écrit

$$\sum_{k=0}^{N-1} \sum_{j=0}^{M} \hat{u}_1(k,j,t) L_k(x) (\pm 1)^j = 0 \ ; \ \forall x \in [-1, 1]$$

soit

$$\sum_{j=0}^{M} \hat{u}_1(k,j,t) (\pm 1)^j = 0 \ ; \ \forall k \in [\![0,\ N-1]\!]$$

On en déduit facilement

$$\sum_{\substack{j=0 \\ j \text{ pair}}}^{M} \hat{u}_1(k,j,t) = 0 \quad \text{et} \quad \sum_{\substack{j=1 \\ j \text{ impair}}}^{M-1} \hat{u}_1(k,j,t) = 0 \qquad \forall k \in [\![0,\ N-1]\!]$$

soit

$$\left\{\begin{array}{lcll} \hat{u}_1(k,M,t) & = & -\displaystyle\sum_{\substack{j=0 \\ j \text{ pair}}}^{M-2} \hat{u}_1(k,j,t) & \\ & & & \forall k \in [\![0,\ N-1]\!] \\ \hat{u}_1(k,M-1,t) & = & -\displaystyle\sum_{\substack{j=1 \\ j \text{ impair}}}^{M-3} \hat{u}_1(k,j,t) & \end{array}\right. \tag{3.14}$$

Un raisonnement analogue pour U_2 mène à

$$\sum_{\substack{k=0 \\ k \text{ pair}}}^{N} \hat{u}_2(k,j,t) = 0 \quad \text{et} \quad \sum_{\substack{k=1 \\ k \text{ impair}}}^{N-1} \hat{u}_2(k,j,t) = 0 \qquad \forall j \in [\![0,\ M-1]\!]$$

soit encore

$$\left\{\begin{array}{lcll} \hat{u}_2(N,j,t) & = & -\displaystyle\sum_{\substack{k=0 \\ k \text{ pair}}}^{N-2} \hat{u}_2(k,j,t) & \\ & & & \forall j \in [\![0,\ M-1]\!] \\ \hat{u}_2(N-1,j,t) & = & -\displaystyle\sum_{\substack{k=1 \\ k \text{ impair}}}^{N-3} \hat{u}_2(k,j,t) & \end{array}\right. \tag{3.15}$$

Montrons maintenant que le problème semi-discrétisé en espace est bien posé. Pour cela on écrit les inconnues U_1, U_2 et U_3 sous la forme:

$$\begin{array}{rcccl} U_1(x,y,t) & = & \displaystyle\sum_{k=0}^{N-1}\sum_{j=0}^{M} \hat{u}_1(k,j,t)L_k(x)L_j(y) & = & \displaystyle\sum_{j=0}^{M} A_j(x)L_j(y) \\ U_2(x,y,t) & = & \displaystyle\sum_{k=0}^{N}\sum_{j=0}^{M-1} \hat{u}_2(k,j,t)L_k(x)L_j(y) & = & \displaystyle\sum_{k=0}^{N} B_k(y)L_k(x) \\ U_3(x,y,t) & = & \displaystyle\sum_{k=0}^{N-1}\sum_{j=0}^{M-1} \hat{u}_3(k,j,t)L_k(x)L_j(y) & = & \displaystyle\sum_{j=0}^{M-1} C_j(x)L_j(y) = \sum_{k=0}^{N-1} D_k(y)L_k(x) \end{array}$$

Comme U_1 et U_3 sont solution du problème, alors l'équation

$$\frac{\partial U_1}{\partial t} = \frac{\partial U_3}{\partial y}$$

est vérifiée pour les modes $j = 0, ..., (M-2)$.
On peut donc l'écrire:

$$\frac{\partial U_1}{\partial t} - \frac{\partial U_3}{\partial y} = \tau_1(x)L_M(y) + \tau_2(x)L_{M-1}(y) \tag{3.16}$$

En effet, $\dfrac{\partial U_1}{\partial t}$ et $\dfrac{\partial U_3}{\partial y}$ sont des polynômes de degré resp. M et $M-1$ en y.

On prend le produit scalaire de (3.16) avec $L_j(y)$ pour $0 \leqslant j \leqslant M$:

$$\begin{aligned} &\left(\frac{\partial U_1}{\partial t}, L_j\right) - \left(\frac{\partial U_3}{\partial y}, L_j\right) = (\tau_1 L_M, L_j) + (\tau_2 L_{M-1}, L_j) \\ \Rightarrow\quad &\frac{d\hat{A}_j}{dt} - \hat{C}_j^{(1)} = \tau_1 \delta_{j,M} + \tau_2 \delta_{j,M-1} \end{aligned}$$

On a aussi

$$\hat{C}_M^{(1)} = 0 \quad \text{et} \quad \hat{C}_{M-1}^{(1)} = (2M-1)\hat{C}_M = 0 \ \text{ par définition de } \ C_M$$

Pour $j = M$, on a $\dfrac{\partial \hat{A}_M}{\partial t} = \tau_1(x)$

Pour $j = (M-1)$, on a $\dfrac{\partial \hat{A}_{M-1}}{\partial t} - \hat{C}_{M-1}^{(1)} = \tau_2(x)$; i.e. $\dfrac{\partial \hat{A}_{M-1}}{\partial t} = \tau_2(x)$.

Comme U_2 et U_3 sont solution du problème, alors l'équation

$$\frac{\partial U_2}{\partial t} = -\frac{\partial U_3}{\partial x}$$

est vérifiée pour les modes $k = 0, ..., (N-2)$. On peut donc l'écrire :

$$\frac{\partial U_2}{\partial t} + \frac{\partial U_3}{\partial x} = \tau_3(y) L_N(x) + \tau_4(y) L_{N-1}(x) \tag{3.17}$$

En effet, $\dfrac{\partial U_2}{\partial t}$ et $\dfrac{\partial U_3}{\partial x}$ sont des polynômes de degré resp. N et $N-1$ en x. On prend le produit scalaire de (3.17) avec $L_k(x)$ pour $0 \leqslant k \leqslant N$:

$$\begin{aligned} &\left(\frac{\partial U_2}{\partial t}, L_k\right) + \left(\frac{\partial U_3}{\partial x}, L_k\right) = (\tau_3 L_N, L_k) + (\tau_4 L_{N-1}, L_k) \\ \Rightarrow\quad &\frac{d\hat{B}_k}{dt} + \hat{D}_k^{(1)} = \tau_3 \delta_{k,N} + \tau_4 \delta_{k,N-1} \end{aligned}$$

On a aussi

$$\hat{D}_N^{(1)} = 0 \quad \text{et} \quad \hat{D}_{N-1}^{(1)} = (2N-1)\hat{D}_N = 0 \ \text{ par définition de } \ D_N$$

Pour $k = N$, on a $\dfrac{\partial \hat{B}_N}{\partial t} = \tau_3(y)$

Pour $k = (N-1)$, on a $\dfrac{\partial \hat{B}_{N-1}}{\partial t} + \hat{D}_{N-1}^{(1)} = \tau_4(y)$; i.e. $\dfrac{\partial \hat{B}_{N-1}}{\partial t} = \tau_4(y)$.

On applique la méthode de l'énergie :

$$\frac{1}{2}\frac{d}{dt}\int_\Omega U_1^2 + U_2^2 + U_3^2 \, \mathrm{d}\Omega = \left(\frac{\partial U_1}{\partial t}, U_1\right) + \left(\frac{\partial U_2}{\partial t}, U_2\right) + \left(\frac{\partial U_3}{\partial t}, U_3\right)$$

$$
\begin{aligned}
= & \left(\frac{\partial U_3}{\partial y}, U_1\right) + \int_x \left\{\tau_1 \int_y L_M(y) U_1 \,\mathrm{d}y\right\} \mathrm{d}x + \int_x \left\{\tau_2 \int_y L_{M-1}(y) U_1 \,\mathrm{d}y\right\} \mathrm{d}x + \left(\frac{\partial U_1}{\partial y}, U_3\right) \\
- & \left(\frac{\partial U_3}{\partial x}, U_2\right) + \int_y \left\{\tau_3 \int_x L_N(x) U_2 \,\mathrm{d}x\right\} \mathrm{d}y + \int_y \left\{\tau_4 \int_x L_{N-1}(x) U_2 \,\mathrm{d}x\right\} \mathrm{d}y - \left(\frac{\partial U_2}{\partial x}, U_3\right) \\
= & \int_x \frac{\partial A_M}{\partial t} \frac{2}{2M+1} A_M(x) \,\mathrm{d}x + \int_x \frac{\partial A_{M-1}}{\partial t} \frac{2}{2M-1} A_{M-1}(x) \,\mathrm{d}x \\
+ & \int_x \frac{\partial B_N}{\partial t} \frac{2}{2N+1} B_N(y) \,\mathrm{d}y + \int_x \frac{\partial B_{N-1}}{\partial t} \frac{2}{2N-1} B_{N-1}(y) \,\mathrm{d}y \\
= & \frac{2}{2M+1} \frac{1}{2} \frac{d}{dt} \int_x A_M^2 \,\mathrm{d}x + \frac{2}{2M-1} \frac{1}{2} \frac{d}{dt} \int_x A_{M-1}^2 \,\mathrm{d}x \\
+ & \frac{2}{2N+1} \frac{1}{2} \frac{d}{dt} \int_y B_N^2 \,\mathrm{d}y + \frac{2}{2N-1} \frac{1}{2} \frac{d}{dt} \int_y B_{N-1}^2 \,\mathrm{d}y
\end{aligned}
$$

On calcule ces quatre intégrales :

$$
\begin{aligned}
\int_x A_M^2 \,\mathrm{d}x &= \int_x \left\{\sum_{k=0}^{N-1} \hat{u}_1(k, M) L_k(x)\right\} \mathrm{d}x &&= \sum_{k=0}^{N-1} \frac{2}{2k+1} \hat{u}_1(k, M)^2 \\
\int_x A_{M-1}^2 \,\mathrm{d}x &= \int_x \left\{\sum_{k=0}^{N-1} \hat{u}_1(k, M-1) L_k(x)\right\} \mathrm{d}x &&= \sum_{k=0}^{N-1} \frac{2}{2k+1} \hat{u}_1(k, M-1)^2 \\
\int_y B_N^2 \,\mathrm{d}y &= \int_y \left\{\sum_{j=0}^{M-1} \hat{u}_2(N, j) L_j(y)\right\} \mathrm{d}y &&= \sum_{j=0}^{M-1} \frac{2}{2j+1} \hat{u}_2(N, j)^2 \\
\int_y B_{N-1}^2 \,\mathrm{d}y &= \int_y \left\{\sum_{j=0}^{M-1} \hat{u}_2(N-1, j) L_j(y)\right\} \mathrm{d}y &&= \sum_{j=0}^{M-1} \frac{2}{2j+1} \hat{u}_2(N-1, j)^2
\end{aligned}
$$

Finalement, on obtient après substitutions

$$
\begin{aligned}
\Rightarrow \; & \frac{1}{2} \frac{d}{dt} \left\{ \sum_{k=0}^{N-1} \sum_{j=0}^{M} \frac{2}{2k+1} \frac{2}{2j+1} \hat{u}_1(k, j)^2 + \sum_{k=0}^{N} \sum_{j=0}^{M-1} \frac{2}{2k+1} \frac{2}{2j+1} \hat{u}_2(k, j)^2 \right. \\
& \qquad \left. + \sum_{k=0}^{N-1} \sum_{j=0}^{M-1} \frac{2}{2k+1} \frac{2}{2j+1} \hat{u}_3(k, j)^2 \right\} \\
= & \frac{1}{2} \frac{d}{dt} \left\{ \sum_{k=0}^{N-1} \frac{2}{2k+1} \left[\frac{2}{2M+1} \hat{u}_1(k, M)^2 + \frac{2}{2M-1} \hat{u}_1(k, M-1)^2\right] \right. \\
& \qquad \left. + \sum_{j=0}^{M-1} \frac{2}{2j+1} \left[\frac{2}{2N+1} \hat{u}_2(N, j)^2 + \frac{2}{2N-1} \hat{u}_2(N-1, j)^2\right] \right\} \\
\Rightarrow \; & \frac{d}{dt} \left\{ \sum_{k=0}^{N-1} \sum_{j=0}^{M-2} \frac{2}{2k+1} \frac{2}{2j+1} \hat{u}_1(k, j)^2 + \sum_{k=0}^{N-2} \sum_{j=0}^{M-1} \frac{2}{2k+1} \frac{2}{2j+1} \hat{u}_2(k, j)^2 \right. \\
& \qquad \left. + \sum_{k=0}^{N-1} \sum_{j=0}^{M-1} \frac{2}{2k+1} \frac{2}{2j+1} \hat{u}_3(k, j)^2 \right\} = 0
\end{aligned}
$$

$$\Rightarrow \quad \frac{d}{dt}\left\{|\, P_{N-1,\,M-2}U_1(t)\,|^2 + |\, P_{N-2,\,M-1}U_2(t)\,|^2 + |\, P_{N-1,\,M-1}U_3(t)\,|^2\right\} = 0$$

$$\Rightarrow \quad |\, P_{N-1,\,M-2}U_1(t)\,|^2 + |\, P_{N-2,\,M-1}U_2(t)\,|^2 + |\, P_{N-1,\,M-1}U_3(t)\,|^2 = \text{cste}$$

$$\Rightarrow \quad \|\, \widetilde{\mathcal{P}}_{N,\,M}\mathbf{U}\,\| = \text{cste}$$

On vient donc de montrer que le problème semi-discrétisé en espace est bien posé.

Remarque *9*

Si l'on applique la méthode Tau-Legendre usuelle, on obtient alors la discrétisation spatiale

$$\begin{cases} U_1(x,y,t) = \displaystyle\sum_{k=0}^{N}\sum_{j=0}^{M} \hat{u}_1(k,j,t)L_k(x)L_j(y) \\ U_2(x,y,t) = \displaystyle\sum_{k=0}^{N}\sum_{j=0}^{M} \hat{u}_2(k,j,t)L_k(x)L_j(y) \\ U_3(x,y,t) = \displaystyle\sum_{k=0}^{N}\sum_{j=0}^{M} \hat{u}_3(k,j,t)L_k(x)L_j(y) \end{cases}$$

qui mène au résultat

$$\begin{aligned} &\frac{1}{2}\quad \frac{d}{dt}\left\{|\, P_{N,\,M-2}U_1(t)\,|^2 + |\, P_{N-2,\,M}U_2(t)\,|^2 + |\, P_{N,\,M}U_3(t)\,|^2\right\} \\ &= \quad -\sum_{k=0}^{N} \hat{u}_3(k,M,t)\,\hat{u}_1(k,M-1,t)\,\frac{4}{2k+1} \\ &\qquad +\sum_{j=0}^{M} \hat{u}_3(N,j,t)\,\hat{u}_2(N-1,j,t)\,\frac{4}{2j+1} \end{aligned}$$

Et nous ne pouvons conclure.

Remarque *10*

L'existence et l'unicité de la solution du problème semi-discrétisé en espace sont obtenues par le même raisonnement que pour le cas continu, en utilisant le semi-groupe $\mathcal{S}_{N,M}(t)$ engendré par l'opérateur $-\mathcal{A}_{N,M} = -\mathcal{A}$.

L'orthogonalité des polynômes de Legendre pour le produit scalaire usuel de $L^2(\Omega)$ nous donne

$$\begin{cases} \dfrac{d\hat{u}_1(k,j,t)}{dt} = \hat{u}_3^{(0,1)}(k,j,t) & ; \begin{cases} 0 \leqslant k \leqslant N-1 \\ 0 \leqslant j \leqslant M-2 \end{cases} \\[2ex] \dfrac{d\hat{u}_2(k,j,t)}{dt} = -\hat{u}_3^{(1,0)}(k,j,t) & ; \begin{cases} 0 \leqslant k \leqslant N-2 \\ 0 \leqslant j \leqslant M-1 \end{cases} \\[2ex] \dfrac{d\hat{u}_3(k,j,t)}{dt} = \hat{u}_1^{(0,1)}(k,j,t) - \hat{u}_2^{(1,0)}(k,j,t) & ; \begin{cases} 0 \leqslant k \leqslant N-1 \\ 0 \leqslant j \leqslant M-1 \end{cases} \end{cases}$$

$$\left.\begin{array}{ll} \displaystyle\sum_{\substack{j=0\\ j \text{ pair}}}^{M} & \hat{u}_1(k,j,t) = 0 \quad \forall k \in [\![0,\ N-1]\!] \\ \displaystyle\sum_{\substack{j=1\\ j \text{ impair}}}^{M-1} & \hat{u}_1(k,j,t) = 0 \quad \forall k \in [\![0,\ N-1]\!] \end{array}\right\} \begin{array}{l}\text{que l'on note plus} \\ \text{simplement } \ M_y^{bc}\widehat{U}_1 = 0\end{array}$$

$$\left.\begin{array}{ll} \displaystyle\sum_{\substack{k=0\\ k \text{ pair}}}^{N} & \hat{u}_2(k,j,t) = 0 \quad \forall j \in [\![0,\ M-1]\!] \\ \displaystyle\sum_{\substack{k=1\\ k \text{ impair}}}^{N-1} & \hat{u}_2(k,j,t) = 0 \quad \forall j \in [\![0,\ M-1]\!] \end{array}\right\} \begin{array}{l}\text{que l'on note plus} \\ \text{simplement } \ M_x^{bc}\widehat{U}_2 = 0\end{array}$$

où $\hat{u}^{(p,q)}$ désigne le spectre de Legendre de $\dfrac{\partial^{p+q} u}{\partial x^p \partial y^q}$.

Nous écrivons cela matriciellement sous la forme

$$\left\{\begin{array}{lcl} P_{N-1,\,M-2}\dot{\widehat{U}}_1 & = & P_{N-1,\,M-2}D_y\widehat{U}_3 \\ P_{N-2,\,M-1}\dot{\widehat{U}}_2 & = & -\ P_{N-2,\,M-1}D_x\widehat{U}_3 \\ P_{N-1,\,M-1}\dot{\widehat{U}}_3 & = & P_{N-1,\,M-1}D_y\widehat{U}_1 - P_{N-1,\,M-1}D_x\widehat{U}_2 \\ M_y^{bc}\widehat{U}_1 = 0 & & \\ M_x^{bc}\widehat{U}_2 = 0 & & \end{array}\right. \tag{3.18}$$

avec $\left\{\begin{array}{l} P_{p,\,q} \text{ l'opérateur de projection orthogonal dans l'espace } \mathcal{L}_{p,q} \\ \quad \text{pour le produit scalaire usuel} \\ \quad \text{(notation utilisée aussi bien dans l'espace physique que dans l'espace spectral)} ; \\ D_y \text{ la matrice de l'opérateur } \dfrac{\partial}{\partial y} \text{ dans l'espace } \mathcal{L}_{N,M}; \\ D_x \text{ la matrice de l'opérateur } \dfrac{\partial}{\partial x} \text{ dans l'espace } \mathcal{L}_{N,M}. \end{array}\right.$

On écrit cela sous forme vectorielle

$$\left\{\begin{array}{lcl} \widetilde{\mathcal{P}}_{N,\,M}\dot{\widehat{\mathbf{U}}} & = & \mathcal{M}_{\mathcal{A}}\widehat{\mathbf{U}} \\ \mathcal{M}_{x,y}^{\mathrm{bc}}\widehat{\mathbf{U}} & = & 0 \end{array}\right. \tag{3.19}$$

avec

$$\widehat{\mathbf{U}} = \begin{pmatrix} \widehat{U}_1 \\ \widehat{U}_2 \\ \widehat{U}_3 \end{pmatrix} \ ; \ \widetilde{\mathcal{P}}_{N,\,M} = \begin{pmatrix} P_{N-1,\,M-2} \\ P_{N-2,\,M-1} \\ P_{N-1,\,M-1} \end{pmatrix} \ ; \ \mathcal{M}_{\mathcal{A}} = \begin{pmatrix} 0 & 0 & P_{N-1,\,M-2}D_y \\ 0 & 0 & -P_{N-2,\,M-1}D_x \\ P_{N-1,\,M-1}D_y & -P_{N-1,\,M-1}D_x & 0 \end{pmatrix}$$

3.4.2 Invariant.

Nous allons maintenant montrer que le problème semi-discrétisé en espace possède un invariant, analogue discret, de l'invariant du problème continu.

Le système différentiel (3.19) ne porte par sur $\mathbf{U}$ mais sur $\widetilde{\mathcal{P}}_{N,M}\mathbf{U}$ en raison des conditions aux limites. Nous avons obtenu

$$\left(\left(\frac{\partial \widetilde{\mathcal{P}}_{N,M}\mathbf{U}}{\partial t}, \widetilde{\mathcal{P}}_{N,M}\mathbf{U}\right)\right) = \frac{1}{2}\frac{d}{dt} \parallel \widetilde{\mathcal{P}}_{N,M}\mathbf{U}^2 \parallel = 0$$

soit

$$| P_{N-1,M-2}U_1(t) |^2 + | P_{N-2,M-1}U_2(t) |^2 + | P_{N-1,M-1}U_3(t) |^2 = \text{cste}$$

3.4.3 Discrétisation temporelle.

Les solutions des équations que nous étudions sont des ondes planes, des fonctions très régulières. Nous discrétisons en temps ces équations à l'aide de schémas précis : schéma de Runge-Kutta explicite d'ordre 4, schéma de Runge-Kutta semi-implicite d'ordre 4 et schéma de Crank-Nicholson semi-implicite d'ordre 2.

Les schémas de type Runge Kutta ont fait l'objet de nombreux travaux. On peut consulter notamment les ouvrages de ([12, 20]).

3.4.3.1 Forme générale des schémas de Runge-Kutta.

Pour approcher la solution $y : [0,T] \to \mathbb{R}^p$ d'une équation différentielle

$$y' = f(t;y) \ , \ y(0) = y_0 \tag{3.20}$$

où $f : [0,T] \times \mathbb{R}^p \to \mathbb{R}^p$ est une fonction suffisamment régulière, l'idée des méthodes de Runge-Kutta est de passer de t_k à $t_k + \Delta t = t_{k+1}$ en approchant l'intégrale de

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(t;y(t))\,\mathrm{d}t \tag{3.21}$$

par une formule de quadrature

$$y(t_{k+1}) = y(t_k) + \Delta t \sum_{i=1}^{q} b_i f(t_k + c_i\Delta t; y(t_k + c_i\Delta t)\ ) \tag{3.22}$$

à l'aide des poids $(b_i)_{i=1,q}$ et des points de quadrature $(c_i)_{i=1,q}$.
Nous notons $t_{k,i}$ pour $t_k + c_i\Delta t$.

Supposons que nous disposons d'une approximation y_k de $y(t_k)$; pour utiliser (3.22) nous avons besoin des valeurs $y_{k,i}$ à substituer aux $y(t_{k,i})$. On les calcule par quadrature sur les mêmes noeuds

$$y_{k,i} = y_k + \Delta t \sum_{j=1}^{q} a_{i,j} f(t_{k,j}; y_{k,j}) \ ; \ 1 \leqslant i \leqslant q \tag{3.23}$$

En général, il s'agit d'un ensemble d'équations implicites, que nous résolvons pour utiliser (3.22) et ainsi obtenir la prochaine valeur y_{k+1} de y :

$$y_{k+1} = y_k + \Delta t \sum_{i=1}^{q} b_i f(t_{k,i}; y_{k,i}) \tag{3.24}$$

Les formules (3.23) et (3.24) réunies définissent une méthode de Runge-Kutta que nous désignons en mettant ses coefficients sous la forme d'un tableau :

$$\begin{array}{c|cccc} c_1 & a_{1,1} & \dots & \dots & a_{1,q} \\ & & & & \\ c_q & a_{q,1} & \dots & \dots & a_{q,q} \\ \hline & b_1 & \dots & \dots & b_q \end{array} \qquad \begin{array}{c|c} \mathrm{C} & \mathrm{A} \\ \hline & b^T \end{array} \quad \text{et} \quad \begin{array}{rcl} \mathrm{A} & = & (a_{i,j})_{i,j} \\ \mathrm{b} & = & (b_i)_i \\ \mathrm{C} & = & diag(c_i)_i \end{array} \quad \text{vérifiant} \left\{ \begin{array}{l} \displaystyle\sum_{i=0}^{q} b_i = 1 \\ \displaystyle\sum_{i=0}^{q} a_{i,q} = c_i \end{array} \right.$$

3.4.3.2 Consistance des schémas de Runge-Kutta.

A l'aide des notations précédentes et de $e = (e_i)_i = (1)_i$, nous énonçons un lemme présentant les conditions nécessaires et suffisantes pour qu'une méthode soit d'ordre q, $1 \leqslant q \leqslant 4$.

Lemme *3.1*

Une condition nécessaire et suffisante pour qu'une méthode de Runge-Kutta soit
d'ordre 1 est $b^T e = 1$;
d'ordre 2 est ordre 1 ainsi que $b^T C e = b^T A e = \frac{1}{2}$;
d'ordre 3 est ordre 2 ainsi que $b^T C^2 e = \frac{1}{3}$; $b^T A C e = b^T A^2 e = \frac{1}{6}$;
d'ordre 4 est ordre 3 ainsi que $b^T C^3 e = \frac{1}{4}$; $b^T A C^2 e = \frac{1}{12}$; $b^T A^2 C e = b^T A^3 e = \frac{1}{24}$.

Ce démonstration de ce lemme se trouve par exemple dans ([17]).

3.4.3.3 Schéma (RK4).

Nous commençons par le schéma de Runge-Kutta explicite d'ordre 4, noté (RK4). Le tableau de ses coefficients est :

$$\begin{array}{c|cccc} 0 & 0 & & & \\ 1/2 & 1/2 & 0 & & \\ 1/2 & 0 & 1/2 & 0 & \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

On peut vérifier que ce schéma est d'ordre 4 en appliquant le lemme précédent.
Pour le problème linéaire qui nous intéresse

$$\left\{ \begin{array}{rcll} \widetilde{\mathcal{P}}_{N,M} \dot{\widehat{\mathbf{U}}} & = & \mathcal{M}_{\mathcal{A}} \widehat{\mathbf{U}} & \forall t > 0 \\ \mathcal{M}^{\mathrm{bc}}_{x,y} \widehat{\mathbf{U}} & = & 0 & \forall t \geqslant 0 \end{array} \right. \tag{3.25}$$

ce schéma s'écrit

$$\left\{ \begin{array}{l} \widehat{\mathbf{Y}}^{\mathbf{1}} = \widehat{\mathbf{U}}^{\mathbf{k}} \\ (\text{ donc } \mathcal{M}^{\mathrm{bc}}_{x,y} \widehat{\mathbf{Y}}^{\mathbf{1}} = 0) \end{array} \right. \qquad \left\{ \begin{array}{l} \widetilde{\mathcal{P}}_{N,M} \widehat{\mathbf{Y}}^{\mathbf{2}} = \widetilde{\mathcal{P}}_{N,M} \widehat{\mathbf{U}}^{\mathbf{k}} + \dfrac{\Delta t}{2} \mathcal{M}_{\mathcal{A}} \widehat{\mathbf{Y}}^{\mathbf{1}} \\ \mathcal{M}^{\mathrm{bc}}_{x,y} \widehat{\mathbf{Y}}^{\mathbf{2}} = 0 \end{array} \right.$$

$$\left\{ \begin{array}{l} \widetilde{\mathcal{P}}_{N,M} \widehat{\mathbf{Y}}^{\mathbf{3}} = \widetilde{\mathcal{P}}_{N,M} \widehat{\mathbf{U}}^{\mathbf{k}} + \dfrac{\Delta t}{2} \mathcal{M}_{\mathcal{A}} \widehat{\mathbf{Y}}^{\mathbf{2}} \\ \mathcal{M}^{\mathrm{bc}}_{x,y} \widehat{\mathbf{Y}}^{\mathbf{3}} = 0 \end{array} \right. \qquad \left\{ \begin{array}{l} \widetilde{\mathcal{P}}_{N,M} \widehat{\mathbf{Y}}^{\mathbf{4}} = \widetilde{\mathcal{P}}_{N,M} \widehat{\mathbf{U}}^{\mathbf{k}} + \Delta t \mathcal{M}_{\mathcal{A}} \widehat{\mathbf{Y}}^{\mathbf{3}} \\ \mathcal{M}^{\mathrm{bc}}_{x,y} \widehat{\mathbf{Y}}^{\mathbf{4}} = 0 \end{array} \right.$$

L'approximation $\widehat{\mathbf{U}}^{\mathbf{k+1}}$ de $\widehat{\mathbf{U}}((k+1)\Delta t)$ étant obtenue par

$$\left\{\begin{array}{l} \widetilde{\mathcal{P}}_{N,M}\widehat{\mathbf{U}}^{\mathbf{k+1}} = \widetilde{\mathcal{P}}_{N,M}\widehat{\mathbf{U}}^{\mathbf{k}} + \dfrac{\Delta t}{6}\left\{\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{Y}}^{\mathbf{1}} + 2\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{Y}}^{\mathbf{2}} + 2\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{Y}}^{\mathbf{3}} + \mathcal{M}_{\mathcal{A}}\widehat{\mathbf{Y}}^{\mathbf{4}}\right\} \\ \mathcal{M}^{\mathrm{bc}}_{x,y}\widehat{\mathbf{U}}^{\mathbf{k+1}} = 0 \end{array}\right. \tag{3.26}$$

3.4.3.4 Schéma (DIRK4).

Nous prenons aussi un schéma semi-implicite de Runge-Kutta d'ordre 4, noté (DIRK4). Le tableau de ses coefficients est :

$$\begin{array}{c|ccc} (1+\xi)/2 & (1+\xi)/2 & & \\ 1/2 & -\xi/2 & (1+\xi)/2 & \\ (1-\xi)/2 & (1+\xi) & -(1+2\xi) & (1+\xi)/2 \\ \hline & \dfrac{1}{6\xi^2} & 1-\dfrac{1}{3\xi^2} & \dfrac{1}{6\xi^2} \end{array}$$

avec ξ l'une des trois racines du polynôme $\xi^3 - \xi = 1/3$

$$\left(\xi_1 = \frac{2}{\sqrt{3}}\cos\frac{\pi}{18}\ ;\ \xi_2 = -\frac{2}{\sqrt{3}}\cos\frac{7\pi}{18}\ ;\ \xi_3 = -\frac{2}{\sqrt{3}}\cos\frac{5\pi}{18}\right).$$

Ici, pour des raisons de stabilité, on prendra $\xi = \xi_1 \approx 1,1371$.
Les coefficients $(a_{i,j})_{i,j}$ forment une matrice triangulaire inférieure dont les coefficients diagonaux sont identiques. La forme particulière de la diagonale donne le nom de *schéma diagonalement implicite* de Runge-Kutta (DIRK).
De plus, comme $a_{i,i}$ est identique pour $i = 1, 2, 3$, alors chacune des trois étapes nécessite la résolution d'un système linéaire avec la même matrice, ce qui est très commode.

On peut vérifier que ce schéma est d'ordre 4 en appliquant le lemme précédent.
Pour le problème linéaire qui nous intéresse

$$\left\{\begin{array}{rcll} \widetilde{\mathcal{P}}_{N,M}\dot{\widehat{\mathbf{U}}} & = & \mathcal{M}_{\mathcal{A}}\widehat{\mathbf{U}} & \forall t > 0 \\ \mathcal{M}^{\mathrm{bc}}_{x,y}\widehat{\mathbf{U}} & = & 0 & \forall t \geqslant 0 \end{array}\right. \tag{3.27}$$

il s'écrit

$$\left\{\begin{array}{l} \widetilde{\mathcal{P}}_{N,M}\widehat{\mathbf{Y}}^{\mathbf{1}} = \widetilde{\mathcal{P}}_{N,M}\widehat{\mathbf{U}}^{\mathbf{k}} + \dfrac{1+\xi}{2}\Delta t\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{Y}}^{\mathbf{1}} \\ \mathcal{M}^{\mathrm{bc}}_{x,y}\widehat{\mathbf{Y}}^{\mathbf{1}} = 0 \end{array}\right. \qquad \left\{\begin{array}{l} \widetilde{\mathcal{P}}_{N,M}\widehat{\mathbf{Y}}^{\mathbf{2}} = \widetilde{\mathcal{P}}_{N,M}\widehat{\mathbf{U}}^{\mathbf{k}} - \dfrac{\xi}{2}\Delta t\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{Y}}^{\mathbf{1}} + \dfrac{1+\xi}{2}\Delta t\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{Y}}^{\mathbf{2}} \\ \mathcal{M}^{\mathrm{bc}}_{x,y}\widehat{\mathbf{Y}}^{\mathbf{2}} = 0 \end{array}\right.$$

$$\left\{\begin{array}{l} \widetilde{\mathcal{P}}_{N,M}\widehat{\mathbf{Y}}^{\mathbf{3}} = \widetilde{\mathcal{P}}_{N,M}\widehat{\mathbf{U}}^{\mathbf{k}} + (1+\xi)\,\Delta t\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{Y}}^{\mathbf{1}} - (1+2\xi)\,\Delta t\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{Y}}^{\mathbf{2}} + \dfrac{1+\xi}{2}\Delta t\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{Y}}^{\mathbf{3}} \\ \mathcal{M}^{\mathrm{bc}}_{x,y}\widehat{\mathbf{Y}}^{\mathbf{3}} = 0 \end{array}\right.$$

L'approximation $\widehat{\mathbf{U}}^{\mathbf{k+1}}$ de $\widehat{\mathbf{U}}((n+1)\Delta t)$ étant obtenue par

$$\left\{\begin{array}{l} \widetilde{\mathcal{P}}_{N,M}\widehat{\mathbf{U}}^{\mathbf{k+1}} = \widetilde{\mathcal{P}}_{N,M}\widehat{\mathbf{U}}^{\mathbf{k}} + \Delta t\left\{\dfrac{1}{6\xi^2}\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{Y}}^{\mathbf{1}} + \left(1-\dfrac{1}{3\xi^2}\right)\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{Y}}^{\mathbf{2}} + \dfrac{1}{6\xi^2}\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{Y}}^{\mathbf{3}}\right\} \\ \mathcal{M}^{\mathrm{bc}}_{x,y}\widehat{\mathbf{U}}^{\mathbf{k+1}} = 0 \end{array}\right. \tag{3.28}$$

Pour les trois sous-pas précédents, nous devons résoudre un système linéaire de la forme

$$\left\{\begin{array}{l} \left(\widetilde{\mathcal{P}}_{N,M} - \alpha \Delta t \mathcal{M}_{\mathcal{A}}\right) \widehat{\mathbf{Y}} = \widehat{\mathbf{B}} \\ \mathcal{M}^{\mathrm{bc}}_{x,y} \widehat{\mathbf{Y}} = 0 \end{array}\right. \tag{3.29}$$

en posant $\alpha = \dfrac{1+\xi}{2}$, que l'on écrit

$$\overline{\mathcal{M}} \widehat{\mathbf{Y}} = \widehat{\mathbf{B}} \tag{3.30}$$

pour $\widehat{\mathbf{Y}} = \widehat{\mathbf{Y}}^{\mathbf{i}}$ et $\widehat{\mathbf{B}} = \widehat{\mathcal{B}}^{\mathbf{i}}$ et

$$\left\{\begin{array}{lcl} \widehat{\mathcal{B}}^{\mathbf{1}} & = & \widetilde{\mathcal{P}}_{N,M} \widehat{\mathbf{U}}^{\mathbf{k}} \\ \widehat{\mathcal{B}}^{\mathbf{2}} & = & \widetilde{\mathcal{P}}_{N,M} \widehat{\mathbf{U}}^{\mathbf{k}} - \dfrac{\xi}{2} \Delta t \mathcal{M}_{\mathcal{A}} \widehat{\mathbf{Y}}^{\mathbf{1}} \\ \widehat{\mathcal{B}}^{\mathbf{3}} & = & \widetilde{\mathcal{P}}_{N,M} \widehat{\mathbf{U}}^{\mathbf{k}} + (1+\xi) \Delta t \mathcal{M}_{\mathcal{A}} \widehat{\mathbf{Y}}^{\mathbf{1}} - (1+2\xi) \Delta t \mathcal{M}_{\mathcal{A}} \widehat{\mathbf{Y}}^{\mathbf{2}} \end{array}\right. \qquad 1 \leqslant i \leqslant 3$$

La matrice $\overline{\mathcal{M}}$ est une matrice réelle non symétrique de rang $(3\,NM + N + M)$ qui n'est pas à diagonale strictement dominante. Nous allons maintenant voir que nous pouvons nous restreindre à la résolution d'un système linéaire dont la matrice est toujours réelle non symétrique mais de rang $N\,M$ et surtout à diagonale strictement dominante *sous conditions sur* Δt.
On peut écrire le système linéaire (3.29) sous la forme

$$\begin{pmatrix} \left(P_{N-1,\,M-2} + M^{bc}_y\right) & 0 & -\alpha \Delta t P_{N-1,\,M-2} D_y \\ 0 & \left(P_{N-2,\,M-1} + M^{bc}_x\right) & \alpha \Delta t P_{N-2,\,M-1} D_x \\ -\alpha \Delta t P_{N-1,\,M-1} D_y & \alpha \Delta t P_{N-1,\,M-1} D_x & Id \end{pmatrix} \begin{pmatrix} \widehat{Y}_1 \\ \widehat{Y}_2 \\ \widehat{Y}_3 \end{pmatrix} = \begin{pmatrix} \widehat{B}_1 \\ \widehat{B}_2 \\ \widehat{B}_3 \end{pmatrix}$$

en posant $\widehat{\mathbf{Y}} = \left(\widehat{Y}_1, \widehat{Y}_2, \widehat{Y}_3\right)^T$ et $\widehat{\mathbf{B}} = \left(\widehat{B}_1, \widehat{B}_2, \widehat{B}_3\right)^T$.
Que l'on note plus simplement

$$\begin{pmatrix} I_1 & 0 & -\alpha \Delta t D^0_y \\ 0 & I_2 & \alpha \Delta t D^0_x \\ -\alpha \Delta t D_y & \alpha \Delta t D_x & Id \end{pmatrix} \begin{pmatrix} \widehat{Y}_1 \\ \widehat{Y}_2 \\ \widehat{Y}_3 \end{pmatrix} = \begin{pmatrix} \widehat{B}_1 \\ \widehat{B}_2 \\ \widehat{B}_3 \end{pmatrix} \tag{3.31}$$

Les matrices I_1 et I_2 ont une forme voisine de la matrice identité et leur inverse se calculent facilement.
De là on obtient :

$$\left\{\begin{array}{lcl} \widehat{Y}_1 & = & (I_1)^{-1} \left\{ \widehat{B}_1 + \alpha \Delta t D^0_y \widehat{Y}_3 \right\} \\ \widehat{Y}_2 & = & (I_2)^{-1} \left\{ \widehat{B}_2 - \alpha \Delta t D^0_x \widehat{Y}_3 \right\} \end{array}\right. \tag{3.32}$$

que l'on substitue dans la troisième équation pour obtenir

$$\left\{ Id - \alpha^2 \Delta t^2 \left[D_y (I_1)^{-1} D^0_y + D_x (I_2)^{-1} D^0_x \right] \right\} \widehat{Y}_3 = \widetilde{B}_3 \tag{3.33}$$

avec

$$\widetilde{B}_3 = \widehat{B}_3 + \alpha \Delta t D_y (I_1)^{-1} \widehat{B}_1 - \alpha \Delta t D_x (I_2)^{-1} \widehat{B}_2$$

soit encore

$$\widetilde{M_3}\,\widehat{Y_3} = \widetilde{B}_3 \tag{3.34}$$

La matrice $\widetilde{M_3}$ est bien conditionnée, voir le tableau (3.1), valeurs obtenues à l'aide de Matlab pour $\Delta t = 10^{-3}$:

TAB. 3.1 – *Conditionnement de $\widetilde{M_3}$ pour (DIRK4).*

(N,M)	Rang($\widetilde{M_3}$)	Cond($\widetilde{M_3}$)
(16,16)	256	1,0210
(32,32)	1032	1,3177
(48,48)	2304	2,5778
(64,64)	4096	5,9378

Nous utilisons l'algorithme du *BI-CG Stab* dont nous rappelons le principe pour le système linéaire type $Ax = b$:

Initialisations

$$\begin{array}{lcl} r_0 &=& b - Ax_0 \\ r_0^* &=& r_0 \\ p_0 &=& r_0 \end{array}$$

Pour $j > 0$, jusqu'à convergence

$$\begin{array}{lcl} \alpha_j &=& \dfrac{((r_0, r_j))}{((Ap_j, r_0^*))} \\ s_j &=& r_j - \alpha_j A p_j \\ \omega_j &=& \dfrac{((As_j, s_j))}{((As_j, As_j))} \\ x_{j+1} &=& x_j + \alpha_j p_j + \omega_j s_j \\ r_{j+1} &=& s_j - \omega_j A s_j \\ \beta_j &=& \dfrac{((r_{j+1}, r_0^*))}{((r_j, r_0^*))}\,\dfrac{\alpha_j}{\omega_j} \\ p_{j+1} &=& r_{j+1} + \beta_j\,(p_j - \omega_j A p_j) \end{array}$$

Nous prendrons comme critère d'arrêt $\epsilon = 10^{-15}$ et le vecteur x_0 sera généralement $\widehat{U}_3^k$, l'approximation de la troisième composante de la solution à l'instant $k\Delta t$.
Une fois $\widehat{Y_3}$ obtenu par cet algorithme itératif, on construit très facilement $\widehat{Y_1}$ et $\widehat{Y_2}$.

3.4.3.5 Forme conservative des schémas de Runge-Kutta.

Le problème qui nous intéresse possède un invariant $\| \mathbf{U}(t) \| = \| \mathbf{U}(0) \|$. Il est intéressant d'utiliser des schémas numériques qui préservent cet invariant.
Les schémas de Runge-Kutta n'ont pas intrinsèquement cette propriété. Par contre on peut leur apporter une modification mineure qui rend ces schémas conservatifs. Nous suivons la démarche de ([20]).

Voyons cela de manière générale sur le problème à valeur initiale suivant :

$$\left\{\begin{array}{rcl} \dfrac{dy}{dt} & = & f(t;y) \quad \forall t>0 \\ y(0) & = & y_0 \end{array}\right. \tag{3.35}$$

qui représente une approximation semi-discrète, continue en temps d'un problème à valeur initiale donné, pour une équation aux dérivées partielles. Nous supposons qu'il existe une unique solution et que f définie de $[0,T]\times\mathbb{R}^p$ dans $\mathbb{R}^p$ possède toute la régularité nécessaire. Nous supposons également que

$$\frac{d\parallel y\parallel}{dt}=0$$

Prenons comme exemple le schéma (RK4).

$$\begin{array}{rcl} y_i & = & y_k+\Delta t\displaystyle\sum_{j=1}^{4}a_{i,j}f(t_k+c_j\Delta t;y_j)\ ;\ 1\leqslant i\leqslant 4 \\ y_{k+1} & = & y_k+\Delta t\displaystyle\sum_{i=1}^{4}b_if(t_k+c_i\Delta t;y_i) \end{array}$$

Nous désignons par $((.,.))$ le produit scalaire usuel de l'espace $L^2(\Omega)$ et par $\parallel.\parallel$ la norme associée. De plus, pour alléger les notations, on écrira $f_j=\Delta tf(t_k+c_j\Delta t;y_j)$. Alors

$$\parallel y_{k+1}\parallel^2=\parallel y_k\parallel^2+2\sum_{i=1}^{4}b_i\,((y_k,f_i))+\sum_{i,j=1}^{4}b_ib_j\,((f_i,f_j)) \tag{3.36}$$

En prenant le produit scalaire de y_i avec f_i, nous avons

$$((y_k,f_i))=((y_i,f_i))-\sum_{j=1}^{4}a_{i,j}\,((f_i,f_j))$$

que la substitution dans la relation précédente donne

$$\parallel y_{k+1}\parallel^2=\parallel y_k\parallel^2-2\sum_{i=1}^{4}b_i\,((y_i,f_i))-Q \tag{3.37}$$

avec

$$Q=\sum_{i,j=1}^{4}m_{i,j}\,((f_i,f_j))\ ;\ m_{i,j}=b_ia_{i,j}+b_ja_{j,i}-b_ib_j \tag{3.38}$$

L'idée est de rendre les poids b_j tels que $Q=0$. Le schéma préserve alors l'invariant.
On note $(\tilde{b}_i)_i$ les poids initiaux du schéma. Soit $b_i=\gamma\tilde{b}_i,\ 1\leqslant i\leqslant 4$, où γ est un paramètre utilisé pour satisfaire la condition $Q=0$.
En remplaçant les $(b_i)_i$ par leur expression dans (3.38), la condition $Q=0$ est vérifiée si

$$\left\{\begin{array}{rcl} \gamma & = & 1-\dfrac{\delta}{\eta} \\ \delta & = & \eta-\tilde{b}_2\,((f_1,f_2))-\tilde{b}_3\,((f_2,f_3))-\tilde{b}_4\,((f_3,f_4)) \\ \eta & = & \parallel\displaystyle\sum_{i=1}^{4}\tilde{b}_if_i\parallel^2 \end{array}\right. \tag{3.39}$$

Pour le schéma (DIRK4), nous obtenons de même:

$$\left\{\begin{array}{lcl}\gamma & = & \dfrac{\delta}{\eta} \\ \delta & = & (1+\xi)\displaystyle\sum_{i=1}^{3}\tilde{b}_i \parallel f_i \parallel^2 - \xi\tilde{b}_2\,((f_1,f_2)) \\ & + & 2\tilde{b}_3\,[(1+\xi)\,((f_1,f_3)) - (1+2\xi)\,((f_2,f_3))] \\ \eta & = & \parallel \displaystyle\sum_{i=1}^{3}\tilde{b}_i f_i \parallel^2 \end{array}\right. \tag{3.40}$$

Nous appelons respectivement (CRK4) et (CDIRK4) les versions conservatives des schémas (RK4) et (DIRK4). Nous comparerons la conservation de l'invariant pour les quatre schémas précédents.
De plus les changements ne concernent pas la résolution des systèmes linéaires pour (CDIRK4).

Examinons le comportement asymptotique de γ lorsque $\Delta t \to 0$. Trivialement $\eta = \mathcal{O}(\Delta t^2)$ puisque $f_i = \Delta t f(t_k + c_i\Delta t, y_i)$ et $\displaystyle\sum_{i=1}^{4}\tilde{b}_i = 1$.
Des calculs fastidieux permettent de montrer que pour (RK4)

$$\delta = \left\{\parallel f_1 - f_2 - f_3 + f_4 \parallel^2 + 3 \parallel f_3 - f_2 \parallel^2 + 6\,((f_3 - f_2, f_1 - f_4))\right\}/36$$

De là, on peut prouver que $\delta = \mathcal{O}(\Delta t^5)$ de telle sorte que $\gamma = 1 + \mathcal{O}(\Delta t^3)$ lorsque $\Delta t \to 0$.
Nous obtenons ainsi un schéma de Runge-Kutta explicite à 4 pas qui est conservatif pour notre problème d'électromagnétisme.
Puisque les poids b_i sont des perturbations en $\mathcal{O}(\Delta t^3)$ des poids originaux b_i, le schéma est d'ordre 3 au lieu de 4 si l'on considère y_{k+1} comme l'approximation de $y(t)$ au temps $t = t_k + \Delta t$. Cependant, si y_{k+1} est vue comme l'approximation de $y(t)$ au temps $t = t_k + \gamma\Delta t$, le schéma est consistant à l'ordre 4. Cela peut être étudié en examinant le développement de l'erreur locale de y_{k+1} en puissance de $\gamma\Delta t$ plutôt que de Δt.
De manière générale, il est recommandé d'avoir γ proche de 1. Pour résumer, nous détaillons les schémas (CRK4) et (CDIRK4):

$$\begin{array}{c|cccc} 0 & 0 & & & \\ 1/2 & 1/2 & 0 & & \\ 1/2 & 0 & 1/2 & 0 & \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \dfrac{\gamma}{6} & \dfrac{\gamma}{3} & \dfrac{\gamma}{3} & \dfrac{\gamma}{6} \end{array} \qquad \begin{array}{c|ccc} (1+\xi)/2 & (1+\xi)/2 & & \\ 1/2 & -\xi/2 & (1+\xi)/2 & \\ (1-\xi)/2 & (1+\xi) & -(1+2\xi) & (1+\xi)/2 \\ \hline & \dfrac{\gamma}{6\xi^2} & \gamma\left[1-\dfrac{1}{3\xi^2}\right] & \dfrac{\gamma}{6\xi^2} \end{array}$$

(CRK4) (CDIRK4)

avec γ défini par (3.39) ou (3.40).

Ces schémas en temps sont inconditionnellement stables. En effet, de $y_k = \widetilde{\mathcal{P}}_{N,M}\mathbf{U}^k$, nous en déduisons $\| \widetilde{\mathcal{P}}_{N,M}\mathbf{U}^k \|^2 = \| \widetilde{\mathcal{P}}_{N,M}\mathbf{U}^0 \|^2$ qui avec les conditions aux limites entraine après quelques calculs :

$$\begin{aligned} \| \mathcal{P}_{N,M}\mathbf{U}^k \| &\leqslant C(N,M) \ \| \widetilde{\mathcal{P}}_{N,M}\mathbf{U}^k \| \\ &\leqslant C(N,M) \ \| \widetilde{\mathcal{P}}_{N,M}\mathbf{U}^0 \| \leqslant C(N,M) \ \| \mathcal{P}_{N,M}\mathbf{U}^0 \| \end{aligned}$$

où $C(N,M)$ est une constante ne dépendant que de N et M et pas de Δt.

3.4.3.6 Schéma (CN2).

Nous considérons aussi le schéma classique de Crank-Nicholson, d'ordre 2, conservatif et inconditionnellement stable, que l'on désigne par (CN2).
Pour le problème linéaire

$$\begin{cases} \widetilde{\mathcal{P}}_{N,M}\dot{\widehat{\mathbf{U}}} &= \mathcal{M}_{\mathcal{A}}\widehat{\mathbf{U}} \quad \forall t > 0 \\ \mathcal{M}^{\mathrm{bc}}_{x,y}\widehat{\mathbf{U}} &= 0 \qquad\quad \forall t \geqslant 0 \end{cases} \tag{3.41}$$

il s'écrit

$$\begin{cases} \widetilde{\mathcal{P}}_{N,M}\widehat{\mathbf{U}}^{\mathbf{k+1}} = \widetilde{\mathcal{P}}_{N,M}\widehat{\mathbf{U}}^{\mathbf{k}} + \frac{\Delta t}{2}\mathcal{M}_{\mathcal{A}}\left\{\widehat{\mathbf{U}}^{\mathbf{k+1}} + \widehat{\mathbf{U}}^{\mathbf{k}}\right\} \\ \mathcal{M}^{\mathrm{bc}}_{x,y}\widehat{\mathbf{U}}^{\mathbf{k+1}} = 0 \end{cases} \tag{3.42}$$

La résolution du système linéaire est similaire à celle du schéma (DIRK4). On doit résoudre

$$\begin{cases} \left(\widetilde{\mathcal{P}}_{N,M} - \frac{\Delta t}{2}\mathcal{M}_{\mathcal{A}}\right)\widehat{\mathbf{U}}^{\mathbf{k+1}} = \widehat{\mathbf{B}} \\ \mathcal{M}^{\mathrm{bc}}_{x,y}\widehat{\mathbf{U}}^{\mathbf{k+1}} = 0 \end{cases} \tag{3.43}$$

que l'on note sous la forme

$$\begin{pmatrix} I_1 & 0 & -\frac{\Delta t}{2}D^0_y \\ 0 & I_2 & \frac{\Delta t}{2}D^0_x \\ -\frac{\Delta t}{2}D_y & \frac{\Delta t}{2}D_x & Id \end{pmatrix} \begin{pmatrix} \widehat{U}^{k+1}_1 \\ \widehat{U}^{k+1}_2 \\ \widehat{U}^{k+1}_3 \end{pmatrix} = \begin{pmatrix} \widehat{B}_1 \\ \widehat{B}_2 \\ \widehat{B}_3 \end{pmatrix} \tag{3.44}$$

avec

$$\begin{pmatrix} \widehat{B}_1 \\ \widehat{B}_2 \\ \widehat{B}_3 \end{pmatrix} = \begin{pmatrix} I_1 & 0 & \frac{\Delta t}{2}D^0_y \\ 0 & I_2 & -\frac{\Delta t}{2}D^0_x \\ \frac{\Delta t}{2}D_y & -\frac{\Delta t}{2}D_x & Id \end{pmatrix} \begin{pmatrix} \widehat{U}^{k}_1 \\ \widehat{U}^{k}_2 \\ \widehat{U}^{k}_3 \end{pmatrix}$$

En substituant

$$\begin{cases} \widehat{U}^{k+1}_1 &= (I_1)^{-1}\left\{\widehat{B}_1 + \frac{\Delta t}{2}D^0_y\,\widehat{U}^{k+1}_3\right\} \\ \widehat{U}^{k+1}_2 &= (I_2)^{-1}\left\{\widehat{B}_2 - \frac{\Delta t}{2}D^0_x\,\widehat{U}^{k+1}_3\right\} \end{cases} \tag{3.45}$$

dans la troisième équation, on obtient

$$\left\{ Id - \frac{\Delta t^2}{4} \left[D_y \left(I_1\right)^{-1} D_y^0 + D_x \left(I_2\right)^{-1} D_x^0 \right] \right\} \widehat{U}_3^{k+1} = \widetilde{B}_3 \tag{3.46}$$

avec

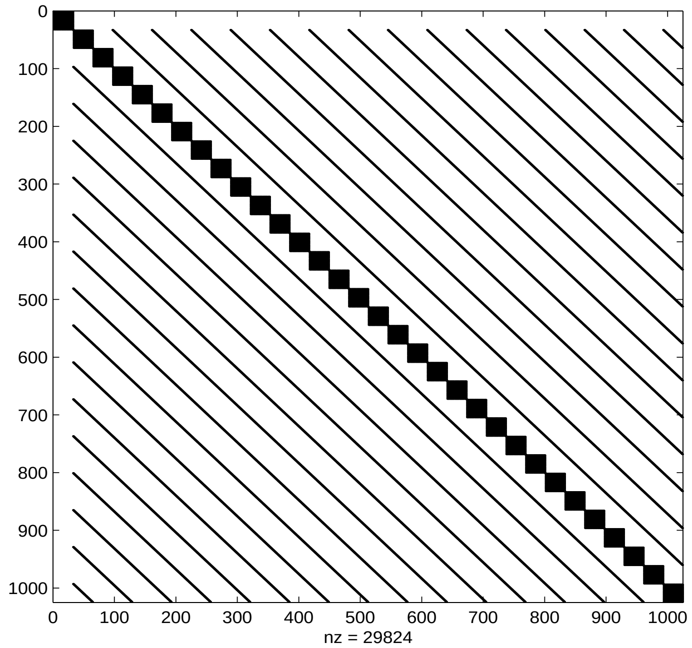
$$\widetilde{B}_3 = \widehat{B}_3 + \frac{\Delta t}{2} D_y \left(I_1\right)^{-1} \widehat{B}_1 - \frac{\Delta t}{2} D_x \left(I_2\right)^{-1} \widehat{B}_2$$

soit encore

$$\widetilde{M_3}\ \widehat{U}_3^{k+1} = \widetilde{B}_3 \tag{3.47}$$

La matrice $\widetilde{M_3}$ est un peu mieux conditionnée que celle du schéma (DIRK4). De même, elle est réelle, non symétrique et à diagonale strictement dominante sous condition sur Δt, ce que nous étudierons au paragraphe suivant. Nous obtenons le tableau (3.2), à l'aide de Matlab pour $\Delta t = 10^{-3}$.
Nous employons là encore l'algorithme du Bi-CG Stab avec le critère d'arrêt $\epsilon = 10^{-15}$.
Une fois $\widehat{U}_3^{k+1}$ obtenu par cet algorithme itératif, on construit très facilement $\widehat{U}_1^{k+1}$ et $\widehat{U}_2^{k+1}$.

FIG. 3.1 – *Squelette de la matrice $\widetilde{M_3}$ pour (CN2) avec $(N, M) = (32, 32)$.*



Nous montrons maintenant que le schéma (CN2) est conservatif et inconditionnellement stable. Nous partons du produit scalaire de l'équation (3.42) avec $\widetilde{\mathcal{P}}_{N,M}\left[\widehat{\mathbf{U}}^{\mathbf{k+1}} + \widehat{\mathbf{U}}^{\mathbf{k}}\right]$.
Cela s'écrit

$$\| \widetilde{\mathcal{P}}_{N,M} \widehat{\mathbf{U}}^{\mathbf{k+1}} \|^2 - \| \widetilde{\mathcal{P}}_{N,M} \widehat{\mathbf{U}}^{\mathbf{k}} \|^2 = \frac{\Delta t}{2} \left(\left(\widetilde{\mathcal{P}}_{N,M} \mathcal{M}_{\mathcal{A}} \left[\widehat{\mathbf{U}}^{\mathbf{k+1}} + \widehat{\mathbf{U}}^{\mathbf{k}} \right], \widetilde{\mathcal{P}}_{N,M} \left[\widehat{\mathbf{U}}^{\mathbf{k+1}} + \widehat{\mathbf{U}}^{\mathbf{k}} \right] \right) \right) = 0$$

en appliquant les résultats du paragraphe 3.1 au vecteur $\mathbf{U} = \left(\mathbf{U}^{k+1} + \mathbf{U}^k\right)$.
Ainsi le schéma en temps (CN2) est conservatif. comme pour les schémas de type Runge-Kutta, cela entraine une inconditionnelle stabilité :

$$\begin{aligned} \| \mathcal{P}_{N,M} \mathbf{U}^k \| &\leqslant C(N,M) \, \| \widetilde{\mathcal{P}}_{N,M} \mathbf{U}^k \| \\ &\leqslant C(N,M) \, \| \widetilde{\mathcal{P}}_{N,M} \mathbf{U}^0 \| \leqslant C(N,M) \, \| \mathcal{P}_{N,M} \mathbf{U}^0 \| \end{aligned}$$

3.4.3.7 Etude de la dominance diagonale des matrices $\widetilde{M}_3$.

Les schémas en temps (CN2), (DIRK4) et (CDIRK4) nécessitent la résolution de systèmes linéaires dont la matrice, $\widetilde{M}_3$, est réelle, non symétrique et à diagonale strictement dominante sous condition sur Δt. La figure (3.1) nous fournit le squelette de $\widetilde{M}_3$.

TAB. 3.2 – *Conditionnement de $\widetilde{M}_3$ pour (CN2).*

(N,M)	Rang($\widetilde{M}_3$)	Cond($\widetilde{M}_3$)
(16,16)	256	1,0046
(32,32)	1024	1,0696
(48,48)	2304	1,3455
(64,64)	4096	2,0811

Sa structure par bloc permet d'étudier plus facilement la dominance diagonale.
Notons A_p un bloc matriciel de rang p défini par :

$$A_p(i,j) = \begin{cases} \dfrac{\alpha^2}{2}(2i+1)j(j+1) & \left\{\begin{array}{ll} 4 \leqslant i \leqslant p-2, & i \text{ pair} \\ 2 \leqslant j \leqslant i-2, & j \text{ pair} \end{array}\right. \text{ou} \left\{\begin{array}{ll} 3 \leqslant i \leqslant p-1, & i \text{ impair} \\ 1 \leqslant j \leqslant i-2, & j \text{ impair} \end{array}\right. \\ \dfrac{\alpha^2}{2}(2i+1)i(i+1) \quad \text{avec} & 2 \leqslant i \leqslant p-2, i \text{ pair ou } 1 \leqslant i \leqslant p-1, i \text{ impair} \\ \dfrac{\alpha^2}{2}(2i+1)i(i+1) & \left\{\begin{array}{ll} 2 \leqslant i \leqslant p-4, & i \text{ pair} \\ i+2 \leqslant j \leqslant p-2, & j \text{ pair} \end{array}\right. \text{ou} \left\{\begin{array}{ll} 1 \leqslant i \leqslant p-3, & i \text{ impair} \\ i+2 \leqslant j \leqslant p-1, & j \text{ impair} \end{array}\right. \end{cases}$$

$$A_p(0,0) = 1$$

Chacun des bloc diagonaux de $\widetilde{M}_3$ est formé du bloc A_N (avec $p = N$).
Les co-diagonales de $\widetilde{M}_3$ sont constantes par bloc. La valeur du $j^{\text{ème}}$ bloc d'une co-diagonale de $\widetilde{M}_3$ est le $j^{\text{ème}}$ coefficient de la même co-diagonale de A_M (pour $p = M$). On note que la diagonale principale de $\widetilde{M}_3$ est donnée par les diagonales principales de A_M et A_N augmentée de 1, la contribution de la matrice identité.
Posons

$$\alpha = \begin{cases} \dfrac{1}{2} & \text{pour (CN2)} \\ \dfrac{1+\xi}{2} & \text{pour (CDIRK4)} \end{cases}$$

Alors pour chaque ligne L de $\widetilde{M}_3$ on peut écrire le terme diagonal sous la forme

$$1 + \alpha^2 \Delta t^2 r(i,j), \quad \left\{\begin{array}{ccccc} 0 & \leqslant & i & \leqslant & N-1 \\ 0 & \leqslant & j & \leqslant & M-1 \end{array}\right.$$

et la somme des termes non diagonaux

$$\alpha^2 \Delta t^2 s(i,j), \quad \left\{\begin{array}{ccccc} 0 & \leqslant & i & \leqslant & N-1 \\ 0 & \leqslant & j & \leqslant & M-1 \end{array}\right.$$

avec la notation, basée sur la division euclidienne de L en

$$L = j(M-1) + i \ , \left\{\begin{array}{ccccc} 0 & \leqslant & i & \leqslant & N-1 \\ 0 & \leqslant & j & \leqslant & M-1 \end{array}\right.$$

La matrice $\widetilde{M_3}$ est à diagonale strictement dominante si et seulement si

$$1 + \alpha^2 \Delta t^2 r(i,j) \; > \; \alpha^2 \Delta t^2 s(i,j) \, , \qquad \left\{ \begin{array}{lllll} 0 & \leqslant & i & \leqslant & N-1 \\ 0 & \leqslant & j & \leqslant & M-1 \end{array} \right. \tag{3.48}$$

Nous décomposons l'étude en trois cas distincts :

- (i) Si $r(i,j) = s(i,j)$ alors (3.48) est vérifiée.
- (ii) Si $r(i,j) > s(i,j)$ alors (3.48) est vérifiée.
- (iii) Si $r(i,j) < s(i,j)$ alors Δt doit satisfaire la relation

$$\Delta t < \frac{1}{\alpha \sqrt{s(i,j) - r(i,j)}}$$

A l'aide de Maple, pour les valeurs suivantes de (N, M), nous obtenons les majorants pour Δt donnés dans le tableau (3.3).

TAB. 3.3 – *Majorant de Δt pour la diagonale dominance de $\widetilde{M_3}$.*

(N,M)	(CN2)	(CDIRK4)	Rang($\widetilde{M_3}$)
(32,32)	$4,13.10^{-3}$	$1,83.10^{-3}$	1024
(64,64)	$9,27.10^{-4}$	$4,34.10^{-4}$	4096
(128,128)	$2,21.10^{-4}$	$1,03.10^{-4}$	16384

Nous obtenons ainsi une condition nécessaire et suffisante pour que la matrice $\widetilde{M_3}$ soit inversible. Nous remarquons que la restriction croît un peu plus vite que le rang du système à inverser.

3.4.4 Analyse des résultats.

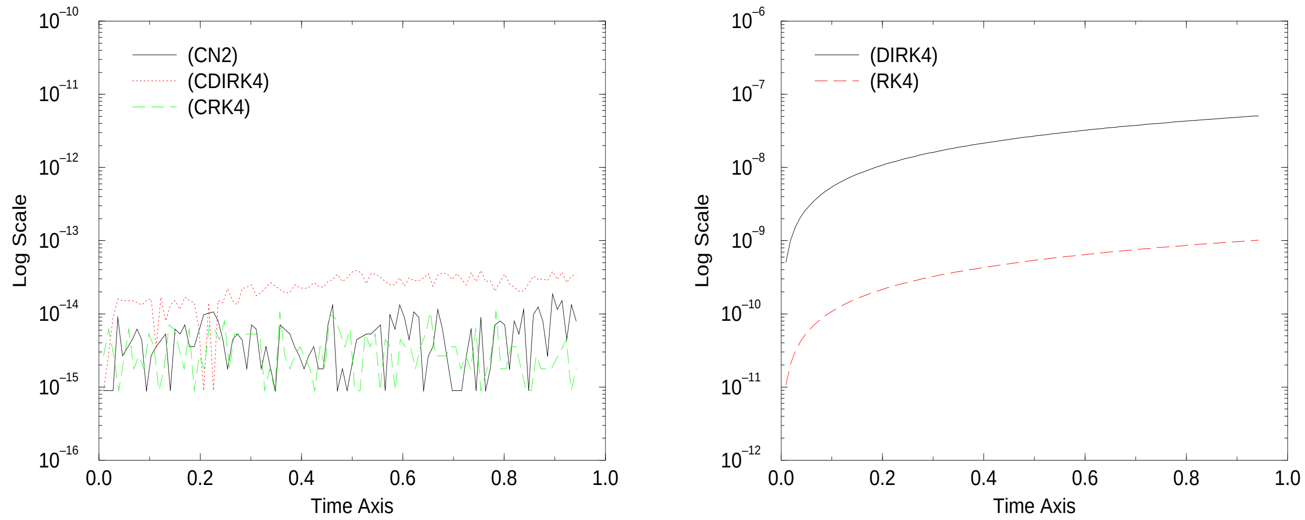
Pour valider les cinq méthodes précédentes (cinq schémas en temps avec la même discrétisation spatiale), nous considérons le petit problème suivant, qui servira par la suite de premier test de comparaison.

Nous prenons le vecteur d'onde $\mathbf{k} = (k_1, k_2) = (3,3)$. Nous fixons les scalaires λ_{3i}, $1 \leqslant i \leqslant 4$ par : $\lambda_3 = 1$, $\lambda_6 = 2$, $\lambda_9 = -1$ et $\lambda_{12} = 1$. Nous intégrons en temps les équations de Maxwell sur les intervalles $[0, 2T]$ et $[0, 10T]$ avec T définie de la manière suivante :
il faut considérer la propagation d'une onde sur une durée au moins égale à deux fois sa période. Nous prenons donc ici deux fois et dix fois la période de l'onde. Il reste à déterminer T. Pour les pulsations ω_i, $1 \leqslant i \leqslant 4$, il n'est pas possible de déterminer T_0 de telle sorte que T_0 soit un multiple entier des périodes T_i associées resp. à ω_i pour $1 \leqslant i \leqslant 4$. Ainsi nous prenons la plus grande des quatre périodes, qui correspond à la plus petite pulsation $\omega_1 = \pi \, |\mathbf{k}|$; i.e. $T = T_1 = \dfrac{2}{3\sqrt{2}} \approx 0,47$.

Nous pouvons alors déterminer les erreurs des schémas de discrétisation, la conservation de l'invariant théorique, l'imposition des conditions aux limites ainsi que les temps calcul pour les méthodes considérées. D'autre part, nous ferons une petite étude statistique des paramètres de l'algorithme du bi-CG stabilisé (nombre d'itérations et résidu).

Les figures (3.2) représentent l'évolution en temps de la différence entre l'invariant théorique et l'invariant calculé. Les conclusions sont immédiates : les versions conservatives des schémas de Runge-Kutta (CRK4) et (CDIRK4) préservent nettement mieux l'invariant que les versions classiques (RK4) et (DIRK4).

FIG. 3.2 – *Evolution en temps de la différence entre les invariants théorique et calculé pour (CN2), (CRK4) et (CDIRK4) (à gauche) ainsi que pour (RK4) et (DIRK4) (à droite).*



Les tableaux (3.4), (3.5) et (3.6) présentent respectivement pour les 3 composantes de

TAB. 3.4 – *Normes et erreurs pour* U_1.

$(t = 2T)$	$\lvert U_1 \rvert_2$	$\lvert U_1 \rvert_\infty$	$\lvert U_1 - U_1^{\text{ex}} \rvert_2$	$\lvert U_1 - U_1^{\text{ex}} \rvert_\infty$	B.C.
(CN2)	0.1567E+01	0.2120E+01	0.2427E-03	0.3794E-03	0.6854E-15
(RK4)	0.1567E+01	0.2120E+01	0.5977E-08	0.9192E-08	0.1364E-14
(DIRK4)	0.1567E+01	0.2120E+01	0.1020E-06	0.1555E-06	0.1349E-14
(CRK4)	0.1567E+01	0.2120E+01	0.6058E-08	0.9310E-08	0.1375E-14
(CDIRK4)	0.1567E+01	0.2120E+01	0.9787E-07	0.1495E-06	0.1439E-14
$(t = 10T)$					
(CN2)	0.1577E+01	0.1890E+01	0.1199E-02	0.1929E-02	0.1931E-14
(RK4)	0.1578E+01	0.1892E+01	0.3047E-07	0.4790E-07	0.7127E-15
(DIRK4)	0.1578E+01	0.1892E+01	0.4588E-06	0.7261E-06	0.6753E-15
(CRK4)	0.1578E+01	0.1892E+01	0.3006E-07	0.4725E-07	0.5832E-15
(CDIRK4)	0.1578E+01	0.1892E+01	0.4790E-06	0.7591E-06	0.1740E-14

$\mathbf{U} = (U_1, U_2, U_3)$ (i.e. E_x, E_y, H_z) les valeurs des normes L^2, L^∞ des composantes et des erreurs absolues sur celles-ci ainsi que l'imposition des conditions aux limites (colonne BC), s'il y a lieu, aux instants $2T$ et $10T$, pour les différents schémas.

TAB. 3.5 – *Normes et erreurs pour U_2.*

$(t = 2T)$	$\vert U_2\vert_2$	$\vert U_2\vert_\infty$	$\vert U_2 - U_2^{\text{ex}}\vert_2$	$\vert U_2 - U_2^{\text{ex}}\vert_\infty$	B.C.
(CN2)	0.1595E+01	0.2226E+01	0.2345E-03	0.3724E-03	0.2259E-14
(RK4)	0.1595E+01	0.2226E+01	0.5809E-08	0.9037E-08	0.2942E-14
(DIRK4)	0.1595E+01	0.2226E+01	0.9910E-07	0.1523E-06	0.4806E-14
(CRK4)	0.1595E+01	0.2226E+01	0.5885E-08	0.9148E-08	0.2802E-14
(CDIRK4)	0.1595E+01	0.2226E+01	0.9519E-06	0.1468E-06	0.5541E-14
$(t = 10T)$					
(CN2)	0.1583E+01	0.2005E+01	0.1190E-02	0.1922E-02	0.2826E-14
(RK4)	0.1584E+01	0.2005E+01	0.3025E-07	0.4770E-07	0.2357E-14
(DIRK4)	0.1584E+01	0.2005E+01	0.4576E-06	0.7251E-06	0.3532E-14
(CRK4)	0.1584E+01	0.2005E+01	0.2989E-07	0.4709E-07	0.5053E-14
(CDIRK4)	0.1584E+01	0.2005E+01	0.4758E-06	0.7562E-06	0.5792E-14

TAB. 3.6 – *Normes et erreurs pour U_3.*

$(t = 2T)$	$\vert U_3\vert_2$	$\vert U_3\vert_\infty$	$\vert U_3 - U_3^{\text{ex}}\vert_2$	$\vert U_3 - U_3^{\text{ex}}\vert_\infty$
(CN2)	0.1414E+01	0.2350E+01	0.5284E-03	0.7094E-03
(RK4)	0.1414E+01	0.2350E+01	0.1342E-07	0.1726E-07
(DIRK4)	0.1414E+01	0.2350E+01	0.2101E-06	0.2763E-06
(CRK4)	0.1414E+01	0.2350E+01	0.1335E-07	0.1727E-07
(CDIRK4)	0.1414E+01	0.2350E+01	0.2136E-06	0.2756E-06
$(t = 10T)$				
(CN2)	0.1416E+01	0.2415E+01	0.2641E-02	0.3192E-02
(RK4)	0.1414E+01	0.2412E+01	0.6628E-07	0.7808E-07
(DIRK4)	0.1414E+01	0.2412E+01	0.1091E-05	0.1279E-05
(CRK4)	0.1414E+01	0.2412E+01	0.6666E-07	0.7821E-07
(CDIRK4)	0.1414E+01	0.2412E+01	0.1073E-05	0.1260E-05

Les tableaux (3.7) et (3.8) illustrent la très bonne aptitude de l'algorithme du bi-gradient conjugué stabilisé à inverser le système

$$\widetilde{M}_3\, \widehat{U}_3^{k+1} = \widetilde{B}_3$$

puisqu'en deux itérations seulement on obtient un résidu nettement en deça de 10^{-16}. Les temps calcul des différents schémas sont présentés dans le tableau (3.9). Les versions conservatives des schémas de Runge-Kutta présentent un surcoût CPU de l'ordre de 10% en plus pour une amélioration d'un facteur 10^5 ou 10^7 de la conservation de l'invariant.

TAB. 3.7 – *Statistiques sur le nombre d'itérations du Bi-CG Stab.*

	Nombre d'itérations					Nombre d'itérations			
$(t = 2T)$	Min	Max	Moy.	Ec-type	$(t = 10T)$	Min	Max	Moy.	Ec-type
(CN2)	2	2	2	0	(CN2)	2	2	2	0
(DIRK4)	2	2	2	0	(DIRK4)	2	2	2	0
(CDIRK4)	2	2	2	0	(CDIRK4)	2	2	2	0

TAB. 3.8 – *Statistiques sur le résidu du Bi-CG Stab.*

	Résidu du système linéaire					Résidu du système linéaire			
$(t = 2T)$	Min	Max	Moy.	Ec-type	$(t = 10T)$	Min	Max	Moy.	Ec-type
(CN2)	4E-21	1E-20	7E-21	2E-21	(CN2)	4E-21	2E-20	1E-20	5E-21
(DIRK4)	4E-19	2E-18	1E-18	6E-19	(DIRK4)	4E-19	2E-18	9E-19	5E-19
(CDIRK4)	5E-19	2E-18	1E-18	5E-19	(CDIRK4)	4E-19	2E-18	1E-18	5E-19

Cette différence provient de la mise en œuvre des schémas. En effet, pour la version conservative du schéma explicite, (CRK4), nous calculons et stockons f_1, f_2, f_3, f_4 alors que pour la version "classique" non conservative nous employons un algorithme qui nécessite moins de stockage et de calculs ([50]). Par conséquent nous n'utiliserons plus que les versions conservatives (CRK4) et (CDIRK4) de ces schémas.

TAB. 3.9 – *Temps calcul des différentes discrétisations sur Sparc 10.*

	(CN2)	(RK4)	(CRK4)	(DIRK4)	(CDIRK4)
$(t = 2T)$	35 s	25 s	26 s	82 s	86 s
$(t = 10T)$	161 s	106 s	118 s	371 s	402 s

3.5 Méthode multi-niveaux.

Nous souhaitons écrire chaque inconnue sous la forme d'une somme de plusieurs termes, chacun étant traité spécifiquement.
Dans ce chapitre, nous nous intéressons à la modélisation de phénomènes électromagnétiques. Un découpage naturel des inconnues se fait en basses fréquences / hautes fréquences. Une difficulté importante réside dans le traitement des conditions aux limites.
Plaçons nous dans un cas monodimensionnel. On écrit une inconnue U sous la forme $U = V + W$ et nous devons imposer $U(\pm 1) = 0$.
i.e. $U(\pm 1) = V(\pm 1) + W(\pm 1) = 0$.
Il faudra choisir (au moins) deux modes pour avoir $U(\pm 1) = 0$.
Différentes possibilités s'offrent à nous :
• une première possibilité consiste à imposer $U(\pm 1) = V(\pm 1) + W(\pm 1) = 0$ en prenant $V(\pm 1) = W(\pm 1) = 0$.
Cela revient à employer 4 degrés de liberté pour 2 conditions aux limites.
D'autre part, nous nous restreignons car

$$U(\pm 1) = V(\pm 1) + W(\pm 1) = 0 \Leftrightarrow V(\pm 1) = -W(\pm 1)$$

ne se limite pas au cas particulier

$$V(\pm 1) = W(\pm 1) = 0$$

• Une autre possibilité est :

$$U(\pm 1) = V(\pm 1) + W(\pm 1) = 0 \Leftrightarrow W(\pm 1) = -V(\pm 1)$$

Avec l'aide du Professeur Labrosse ([35]), nous choisissons d'imposer cette égalité à l'aide des deux derniers degrés de liberté de W, V restant inchangé.
Dans ce cas, nous introduisons un troisième terme T, extrait de W, que nous appellerons "*modes Tau*", défini par

$$T(\pm 1) = -\{V(\pm 1) + W(\pm 1)\}$$

Dans le cadre de notre problème d'électromagnétisme bidimensionnel, nous écrivons

$$\mathbf{U} = \mathbf{V} + \mathbf{W} + \mathbf{T} \text{ avec } \mathbf{V} = \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix} \; ; \; \mathbf{W} = \begin{pmatrix} W_1 \\ W_2 \\ W_3 \end{pmatrix} \; ; \; \mathbf{T} = \begin{pmatrix} T_1 \\ T_2 \\ 0 \end{pmatrix} \tag{3.49}$$

En effet, la troisième inconnue, $U_3 = H_z$ n'a pas de conditions aux limites.
Construisons maintenant les espaces fonctionnels et les opérateurs de projection pour une telle décomposition.

3.5.1 Discrétisation spatiale.

Le problème d'évolution étudié est

$$\left\{ \begin{array}{lcl} \dfrac{\partial U_1}{\partial t} & = & \dfrac{\partial U_3}{\partial y} \\ \dfrac{\partial U_2}{\partial t} & = - & \dfrac{\partial U_3}{\partial x} \\ \dfrac{\partial U_3}{\partial t} & = & \dfrac{\partial U_1}{\partial y} - \dfrac{\partial U_2}{\partial x} \end{array} \right. \qquad \begin{array}{l} \forall x, y \in (-1, +1) \\ \forall t > 0 \end{array} \tag{3.50}$$

muni des conditions aux limites

$$\begin{cases} U_1(x, y = \pm 1, t) &= 0 \\ U_2(x = \pm 1, y, t) &= 0 \end{cases} \tag{3.51}$$

Nous rappelons que la notation $P_{p,\,q}$ désigne l'opérateur de projection orthogonale dans l'espace $\mathcal{L}_{p,q}$ pour le produit scalaire usuel de l'espace $L^2(\Omega)$, avec

$$\mathcal{L}_{p,q} = \mathcal{L}^1_p \otimes \mathcal{L}^2_q = \{\Psi_{k,j}(x,y); 0 \leqslant k \leqslant p \text{ et } 0 \leqslant j \leqslant q\} \text{ avec } \Psi_{k,j}(x,y) = L_k(x)L_j(y)$$

Soient les entiers strictement positifs N, M, N_1, M_1 tels que

$$0 < N_1 < N \quad \text{et} \quad 0 < M_1 < M$$

On définit l'opérateur de projection orthogonale $\mathcal{P}_{N_1,\,M_1}$, sur les basses fréquences pour le produit scalaire usuel de $L^2(\Omega)$ par

$$\mathbf{V} = \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix} = \mathcal{P}_{N_1,\,M_1}\mathbf{U} = \begin{pmatrix} P_{N_1,\,M_1}U_1 \\ P_{N_1,\,M_1}U_2 \\ P_{N_1,\,M_1}U_3 \end{pmatrix} \tag{3.52}$$

de même pour l'opérateur de projection sur les hautes fréquences$\mathcal{Q}_{N_1,\,M_1}$

$$\mathbf{W} = \begin{pmatrix} W_1 \\ W_2 \\ W_3 \end{pmatrix} = \mathcal{Q}_{N_1,\,M_1}\mathbf{U} = \begin{pmatrix} Q_{N-1,\,M-2}U_1 \\ Q_{N-2,\,M-1}U_2 \\ Q_{N-1,\,M-1}U_3 \end{pmatrix} = \begin{pmatrix} (P_{N-1,\,M-2} - P_{N_1,\,M_1})\,U_1 \\ (P_{N-2,\,M-1} - P_{N_1,\,M_1})\,U_2 \\ (P_{N-1,\,M-1} - P_{N_1,\,M_1})\,U_3 \end{pmatrix} \tag{3.53}$$

et enfin pour l'opérateur $\mathcal{R}_{N_1,\,M_1}$ intervenant dans l'imposition des conditions aux limites

$$\mathbf{T} = \begin{pmatrix} T_1 \\ T_2 \\ 0 \end{pmatrix} = \mathcal{R}_{N_1,\,M_1}\mathbf{U} = \begin{pmatrix} R_{N-1,\,M-2}U_1 \\ R_{N-2,\,M-1}U_2 \\ 0 \end{pmatrix} = \begin{pmatrix} (P_{N-1,\,M} - P_{N-1,\,M-2})\,U_1 \\ (P_{N,\,M-1} - P_{N-2,\,M-1})\,U_2 \\ 0 \end{pmatrix} \tag{3.54}$$

Ainsi la solution

$$\begin{cases} U_1(x,y,t) = \displaystyle\sum_{k=0}^{N-1}\sum_{j=0}^{M} \hat{u}_1(k,j,t)\,L_k(x)\,L_j(y) \\ U_2(x,y,t) = \displaystyle\sum_{k=0}^{N}\sum_{j=0}^{M-1} \hat{u}_2(k,j,t)\,L_k(x)\,L_j(y) \\ U_3(x,y,t) = \displaystyle\sum_{k=0}^{N-1}\sum_{j=0}^{M-1} \hat{u}_3(k,j,t)\,L_k(x)\,L_j(y) \end{cases}$$

s'écrit sous la forme

$$\begin{cases} U_1 = V_1 + W_1 + T_1 \\ U_2 = V_2 + W_2 + T_2 \\ U_3 = V_3 + W_3 \end{cases}$$

avec

$$\begin{cases} V_1(x,y,t) = \displaystyle\sum_{k=0}^{N_1}\sum_{j=0}^{M_1} \hat{u}_1(k,j,t)\,L_k(x)\,L_j(y) \\ V_2(x,y,t) = \displaystyle\sum_{k=0}^{N_1}\sum_{j=0}^{M_1} \hat{u}_2(k,j,t)\,L_k(x)\,L_j(y) \\ V_3(x,y,t) = \displaystyle\sum_{k=0}^{N_1}\sum_{j=0}^{M_1} \hat{u}_3(k,j,t)\,L_k(x)\,L_j(y) \end{cases}$$

$$\left\{\begin{array}{l} W_1(x,y,t) = \displaystyle\sum_{(k,j)\in\mathbb{I}_{N-1,M-2}\setminus\mathbb{I}_{N_1,M_1}} \hat{u}_1(k,j,t)\, L_k(x)\, L_j(y) \\ W_2(x,y,t) = \displaystyle\sum_{(k,j)\in\mathbb{I}_{N-2,M-1}\setminus\mathbb{I}_{N_1,M_1}} \hat{u}_2(k,j,t)\, L_k(x)\, L_j(y) \\ W_3(x,y,t) = \displaystyle\sum_{(k,j)\in\mathbb{I}_{N-1,M-1}\setminus\mathbb{I}_{N_1,M_1}} \hat{u}_3(k,j,t)\, L_k(x)\, L_j(y) \end{array}\right.$$

$$\left\{\begin{array}{l} T_1(x,y,t) = \displaystyle\sum_{(k,j)\in\mathbb{I}_{N-1,M}\setminus\mathbb{I}_{N-1,M-2}} \hat{u}_1(k,j,t)\, L_k(x)\, L_j(y) \\ T_2(x,y,t) = \displaystyle\sum_{(k,j)\in\mathbb{I}_{N,M-1}\setminus\mathbb{I}_{N-2,M-1}} \hat{u}_2(k,j,t)\, L_k(x)\, L_j(y) \end{array}\right.$$

et la notation

$$\mathbb{I}_{p,q} = [\![0,\ p]\!] \otimes [\![0,\ q]\!] \tag{3.55}$$

A l'aide des opérateurs $\mathcal{P}_{N_1,\,M_1}$, $\mathcal{Q}_{N_1,\,M_1}$ et $\mathcal{R}_{N_1,\,M_1}$, nous prenons les projections pour le produit scalaire usuel du système semi-discrétisé

$$\left\{\begin{array}{lcl} \widetilde{\mathcal{P}}_{N,\,M}\dot{\widehat{\mathbf{U}}} & = & \mathcal{M}_{\mathcal{A}}\widehat{\mathbf{U}} \\ \mathcal{M}^{\mathrm{bc}}_{x,y}\widehat{\mathbf{U}} & = & 0 \end{array}\right. \tag{3.56}$$

Nous obtenons ainsi

$$\left\{\begin{array}{lcl} \mathcal{P}_{N_1,\,M_1}\dot{\widehat{\mathbf{U}}} & = & \mathcal{P}_{N_1,\,M_1}\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{U}} \\ \mathcal{Q}_{N_1,\,M_1}\dot{\widehat{\mathbf{U}}} & = & \mathcal{Q}_{N_1,\,M_1}\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{U}} \\ \mathcal{M}^{\mathrm{bc}}_{x,y}\widehat{\mathbf{U}} = 0 & & \end{array}\right. \tag{3.57}$$

soit encore

$$\left\{\begin{array}{lcl} \dot{\widehat{\mathbf{V}}} & = & \mathcal{M}_{\mathcal{A}}\widehat{\mathbf{V}} + \ \mathcal{P}_{N_1,\,M_1}\mathcal{M}_{\mathcal{A}}\left(\widehat{\mathbf{W}} + \widehat{\mathbf{T}}\right) \\ \dot{\widehat{\mathbf{W}}} & = & \qquad\qquad \mathcal{Q}_{N_1,\,M_1}\mathcal{M}_{\mathcal{A}}\left(\widehat{\mathbf{W}} + \widehat{\mathbf{T}}\right) \\ \mathcal{M}^{\mathrm{bc}}_{x,y}\left(\widehat{\mathbf{V}} + \widehat{\mathbf{W}} + \widehat{\mathbf{T}}\right) = 0 & & \end{array}\right. \tag{3.58}$$

3.5.2 Discrétisation temporelle.

Nous avions conclu précédemment que les versions non conservatives des schémas de Runge-Kutta étaient qualitativement moins intéressantes. Nous discrétiserons donc en temps les équations projetées précédentes seulement à l'aide des schémas conservatifs suivants : (CRK4), (CDIRK4), (CN2).

3.5.2.1 Schéma (CRK4).

Pour le schéma de Runge-Kutta explicite (CRK4), il n'y a pas de difficulté particulière : nous n'avons pas à faire face à des problèmes de dépendance. En effet, les quantités $\widehat{V}_i$, $\widehat{W}_i$ s'expriment en fonction des valeurs obtenues aux sous-pas précédents et les $\widehat{T}_i$ n'ont besoin que des $\widehat{V}_i$ et $\widehat{W}_i$. Les 4 sous-pas de ce schéma en temps ($1 \leqslant i \leqslant 4$) peuvent se mettre sous la forme générique

$$\left\{\begin{array}{lll} \mathcal{P}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{i}} & = & \mathcal{P}_{N_1,M_1}\widehat{\mathbf{U}}^{\mathbf{k}} \quad +c_i\Delta t\mathcal{M}_{\mathcal{A}}\left(\mathcal{P}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{i-1}}\right) \\ & & \qquad\qquad +c_i\Delta t\mathcal{P}_{N_1,M_1}\mathcal{M}_{\mathcal{A}}\left(\mathcal{Q}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{i-1}}+\mathcal{R}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{i-1}}\right) \\ \mathcal{Q}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{i}} & = & \mathcal{Q}_{N_1,M_1}\widehat{\mathbf{U}}^{\mathbf{k}} \quad +c_i\Delta t\mathcal{Q}_{N_1,M_1}\mathcal{M}_{\mathcal{A}}\left(\mathcal{Q}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{i-1}}+\mathcal{R}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{i-1}}\right) \\ \multicolumn{3}{l}{\mathcal{M}^{\mathrm{bc}}_{x,y}\mathcal{R}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{i}} = -\mathcal{M}^{\mathrm{bc}}_{x,y}\left(\mathcal{P}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{i}}+\mathcal{Q}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{i}}\right)} \end{array}\right. \tag{3.59}$$

en prenant la convention $\widehat{\mathbf{Y}}^{\mathbf{0}} = 0$.
Il en est de même pour la construction de $\widehat{\mathbf{U}}^{\mathbf{k+1}}$

$$\left\{\begin{array}{l} \mathcal{P}_{N_1,M_1}\widehat{\mathbf{U}}^{\mathbf{k+1}} = \mathcal{P}_{N_1,M_1}\widehat{\mathbf{U}}^{\mathbf{k}} \\ \qquad + \dfrac{\Delta t}{6}\Big\{ \mathcal{M}_{\mathcal{A}}\left(\mathcal{P}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{1}}\right) + 2\mathcal{M}_{\mathcal{A}}\left(\mathcal{P}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{2}}\right) \\ \qquad\qquad + 2\mathcal{M}_{\mathcal{A}}\left(\mathcal{P}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{3}}\right) + \mathcal{M}_{\mathcal{A}}\left(\mathcal{P}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{4}}\right) \Big\} \\ \mathcal{Q}_{N_1,M_1}\widehat{\mathbf{U}}^{\mathbf{k+1}} = \mathcal{Q}_{N_1,M_1}\widehat{\mathbf{U}}^{\mathbf{k}} \\ \qquad + \dfrac{\Delta t}{6}\Big\{ \mathcal{Q}_{N_1,M_1}\mathcal{M}_{\mathcal{A}}\left(\mathcal{Q}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{1}}\right) + 2\mathcal{Q}_{N_1,M_1}\mathcal{M}_{\mathcal{A}}\left(\mathcal{Q}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{2}}\right) \\ \qquad\qquad + 2\mathcal{Q}_{N_1,M_1}\mathcal{M}_{\mathcal{A}}\left(\mathcal{Q}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{3}}\right) + \mathcal{Q}_{N_1,M_1}\mathcal{M}_{\mathcal{A}}\left(\mathcal{Q}_{N_1,M_1}\widehat{\mathbf{Y}}^{\mathbf{4}}\right) \Big\} \\ \mathcal{M}^{\mathrm{bc}}_{x,y}\mathcal{R}_{N_1,M_1}\widehat{\mathbf{U}}^{\mathbf{k+1}} = -\mathcal{M}^{\mathrm{bc}}_{x,y}\left(\mathcal{P}_{N_1,M_1}\widehat{\mathbf{U}}^{\mathbf{k+1}}+\mathcal{Q}_{N_1,M_1}\widehat{\mathbf{U}}^{\mathbf{k+1}}\right) \end{array}\right. \tag{3.60}$$

Ce n'est plus le cas pour les deux schémas semi-implicites (CN2) et (CDIRK4). En effet, pour éviter une dépendance en $\widehat{\mathbf{W}}$ dans l'équation

$$\dot{\widehat{\mathbf{V}}} = \mathcal{M}_{\mathcal{A}}\widehat{\mathbf{V}} + \mathcal{P}_{N_1,M_1}\mathcal{M}_{\mathcal{A}}\left(\widehat{\mathbf{W}}+\widehat{\mathbf{T}}\right)$$

on commencera par résoudre l'équation

$$\dot{\widehat{\mathbf{W}}} = \mathcal{Q}_{N_1,M_1}\mathcal{M}_{\mathcal{A}}\left(\widehat{\mathbf{W}}+\widehat{\mathbf{T}}\right)$$

Malheureusement, ces deux équations présentent une dépendance en $\widehat{\mathbf{T}}$ par

$$\mathcal{M}^{\mathrm{bc}}_{x,y}\left(\widehat{\mathbf{V}}+\widehat{\mathbf{W}}+\widehat{\mathbf{T}}\right) = 0$$

que nous ne pouvons éviter.
Le lien entre les termes $\widehat{\mathbf{V}}$, $\widehat{\mathbf{W}}$ et $\widehat{\mathbf{T}}$, les conditions aux limites, sont très forts et essayer de les séparer brutalement conduit à des méthodes instables. On peut, par exemple, envisager de casser la dépendance à l'aide du processus suivant :
supposons que l'on dispose de la solution $\widehat{\mathbf{U}}^{\mathbf{k}} = \widehat{\mathbf{V}}^{\mathbf{k}} + \widehat{\mathbf{W}}^{\mathbf{k}} + \widehat{\mathbf{T}}^{\mathbf{k}}$ à l'instant t_k.
On discrétise en temps à l'aide du schéma (CN2). On extrapole la valeur de $\widehat{\mathbf{T}}^{\mathbf{k+1}}$, notée $\widetilde{\mathbf{T}}^{\mathbf{k+1}}$, à l'aide d'un développement limité à l'ordre 3 pour rester consistant. On peut alors

déterminer sans difficulté $\widehat{\mathbf{W}}^{\mathbf{k+1}}$ puis $\widehat{\mathbf{V}}^{\mathbf{k+1}}$. Mais ce schéma est instable et explose très vite numériquement.
Une autre idée consiste à voir $\widehat{\mathbf{W}}^{\mathbf{k+1}}$ et $\widehat{\mathbf{T}}^{\mathbf{k+1}}$ comme le terme des hautes fréquences (on "n'extrait" pas les modes Tau pour les conditions limites). La dépendance entre $\widehat{\mathbf{V}}^{\mathbf{k+1}}$ et $\widehat{\mathbf{W}}^{\mathbf{k+1}}$ étant dans l'équation en $\widehat{\mathbf{W}}^{\mathbf{k+1}}$:

$$\left\{\begin{array}{lcl} \dot{\widehat{\mathbf{V}}} & = & \mathcal{M}_{\mathcal{A}}\widehat{\mathbf{V}} + \mathcal{P}_{N_1,M_1}\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{W}} \\ \dot{\widehat{\mathbf{W}}} & = & \mathcal{Q}_{N_1,M_1}\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{W}} \\ \multicolumn{3}{l}{\mathcal{M}^{\mathrm{bc}}_{x,y}\left(\widehat{\mathbf{V}} + \widehat{\mathbf{W}}\right) = 0} \end{array}\right.$$

Un algorithme de point fixe est appliqué sur le système formé des équations pour $\widehat{\mathbf{V}}^{\mathbf{k+1}}$ et $\widehat{\mathbf{W}}^{\mathbf{k+1}}$. Après convergence numérique, on met à jour les modes servant à imposer les conditions aux limites à l'aide de la valeur de $\widehat{\mathbf{V}}^{\mathbf{k+1}}$ obtenue à la dernière itération du point fixe.
Cette méthode est un peu meilleure que la précédente, mais conduit au même phénomène d'explosion. Le problème vient du fait que, là encore, les conditions aux limites ne sont pas assez fortement imposées.
On considère alors le couplage des "modes Tau", $\widehat{\mathbf{T}}^{\mathbf{k+1}}$ avec les basses fréquences $\widehat{\mathbf{V}}^{\mathbf{k+1}}$. On a alors un système de deux équations, l'une en $\widehat{\mathbf{W}}^{\mathbf{k+1}}$ (avec une extrapolation pour les modes Tau), l'autre en $\widehat{\mathbf{V}}^{\mathbf{k+1}}$ et $\widehat{\mathbf{T}}^{\mathbf{k+1}}$. Cette alternative est meilleure que les deux précédentes car les conditions aux limites sont imposées plus fortement.
Pour résoudre ce problème, nous utilisons une méthode de point fixe sur $\widehat{\mathbf{T}}$. Nous verrons cela en détail pour chacun des deux schémas : on commence par présenter l'algorithme ensuite on étudie sa convergence et enfin on s'intéresse à la résolution des équations pour d'une part $\widehat{\mathbf{W}}^{\mathbf{k+1}}$ et $\widehat{\mathbf{V}}^{\mathbf{k+1}}$, $\widehat{\mathbf{T}}^{\mathbf{k+1}}$ d'autre part.

Nous devons intégrer en temps le problème semi-discrétisé en espace

$$\left\{\begin{array}{lcl} \dot{\widehat{\mathbf{V}}} & = & \mathcal{M}_{\mathcal{A}}\widehat{\mathbf{V}} + \mathcal{P}_{N_1,M_1}\mathcal{M}_{\mathcal{A}}\left(\widehat{\mathbf{W}} + \widehat{\mathbf{T}}\right) \\ \dot{\widehat{\mathbf{W}}} & = & \mathcal{Q}_{N_1,M_1}\mathcal{M}_{\mathcal{A}}\left(\widehat{\mathbf{W}} + \widehat{\mathbf{T}}\right) \\ \multicolumn{3}{l}{\mathcal{M}^{\mathrm{bc}}_{x,y}\left(\widehat{\mathbf{V}} + \widehat{\mathbf{W}} + \widehat{\mathbf{T}}\right) = 0} \end{array}\right. \tag{3.61}$$

3.5.2.2 Schéma (CN2).

La discrétisation du système (3.61) s'écrit :

$$\left\{\begin{array}{l} \begin{array}{lcl} \widehat{\mathbf{W}}^{\mathbf{k+1}} & = & \widehat{\mathbf{W}}^{\mathbf{k}} \\ & + & \dfrac{\Delta t}{2}\mathcal{Q}_{N_1,M_1}\mathcal{M}_{\mathcal{A}}\left\{\left(\widehat{\mathbf{W}}^{\mathbf{k+1}} + \widehat{\mathbf{T}}^{\mathbf{k+1}}\right) + \left(\widehat{\mathbf{W}}^{\mathbf{k}} + \widehat{\mathbf{T}}^{\mathbf{k}}\right)\right\} \end{array} \\ \\ \left\{\begin{array}{l} \begin{array}{lcl} \widehat{\mathbf{V}}^{\mathbf{k+1}} & = & \widehat{\mathbf{V}}^{\mathbf{k}} + \dfrac{\Delta t}{2}\mathcal{M}_{\mathcal{A}}\left\{\widehat{\mathbf{V}}^{\mathbf{k+1}} + \widehat{\mathbf{V}}^{\mathbf{k}}\right\} \\ & + & \dfrac{\Delta t}{2}\mathcal{P}_{N_1,M_1}\mathcal{M}_{\mathcal{A}}\left\{\left(\widehat{\mathbf{W}}^{\mathbf{k+1}} + \widehat{\mathbf{T}}^{\mathbf{k+1}}\right) + \left(\widehat{\mathbf{W}}^{\mathbf{k}} + \widehat{\mathbf{T}}^{\mathbf{k}}\right)\right\} \end{array} \\ \\ \mathcal{M}^{\mathrm{bc}}_{x,y}\left(\widehat{\mathbf{V}}^{\mathbf{k+1}} + \widehat{\mathbf{T}}^{\mathbf{k+1}}\right) = -\mathcal{M}^{\mathrm{bc}}_{x,y}\widehat{\mathbf{W}}^{\mathbf{k+1}} \end{array}\right. \end{array}\right. \tag{3.62}$$

Supposons $\widehat{\mathbf{T}}^{\mathbf{k+1,0}}$ obtenu par extrapolation.
L'algorithme du point fixe appliqué à ce système s'écrit pour $\nu \geqslant 0$, l'indice de ses itérations :

$$\left\{\begin{array}{l} \begin{array}{rcl} \widehat{\mathbf{W}}^{\mathbf{k+1},\nu+1} & = & \widehat{\mathbf{W}}^{\mathbf{k}} \\ & + & \dfrac{\Delta t}{2}\,\mathcal{Q}_{N_1,\,M_1}\mathcal{M}_{\mathcal{A}}\left\{\left(\widehat{\mathbf{W}}^{\mathbf{k+1},\nu+1}+\widehat{\mathbf{T}}^{\mathbf{k+1},\nu}\right)+\left(\widehat{\mathbf{W}}^{\mathbf{k}}+\widehat{\mathbf{T}}^{\mathbf{k}}\right)\right\} \end{array} \\ \left\{\begin{array}{rcl} \widehat{\mathbf{V}}^{\mathbf{k+1},\nu+1} & = & \widehat{\mathbf{V}}^{\mathbf{k}} + \dfrac{\Delta t}{2}\mathcal{M}_{\mathcal{A}}\left\{\widehat{\mathbf{V}}^{\mathbf{k+1},\nu+1}+\widehat{\mathbf{V}}^{\mathbf{k}}\right\} \\ & + & \dfrac{\Delta t}{2}\,\mathcal{P}_{N_1,\,M_1}\mathcal{M}_{\mathcal{A}}\left\{\left(\widehat{\mathbf{W}}^{\mathbf{k+1},\nu+1}+\widehat{\mathbf{T}}^{\mathbf{k+1},\nu+1}\right)+\left(\widehat{\mathbf{W}}^{\mathbf{k}}+\widehat{\mathbf{T}}^{\mathbf{k}}\right)\right\} \\ \multicolumn{3}{l}{\mathcal{M}^{\mathrm{bc}}_{x,y}\left(\widehat{\mathbf{V}}^{\mathbf{k+1},\nu+1}+\widehat{\mathbf{T}}^{\mathbf{k+1},\nu+1}\right)=-\mathcal{M}^{\mathrm{bc}}_{x,y}\widehat{\mathbf{W}}^{\mathbf{k+1},\nu+1}} \end{array}\right. \end{array}\right. \tag{3.63}$$

Remarque *11*

Nous obtenons la valeur de $\widehat{\mathbf{T}}^{\mathbf{k+1,0}}$ par un développement à l'ordre 3 de la manière suivante :

$$U_i|_{(k+1)\Delta t} = U_i|_{k\Delta t} + \Delta t\frac{\partial U_i}{\partial t}|_{k\Delta t} + \frac{\Delta t^2}{2}\frac{\partial^2 U_i}{\partial t^2}|_{k\Delta t} + \frac{\Delta t^3}{6}\frac{\partial^3 U_i}{\partial t^3}|_{k\Delta t} + \mathcal{O}(\Delta t^4)$$

Or si $\mathbf{U}$ est solution des équations de Maxwell alors ses composantes U_i sont chacune solution de l'équation des Ondes

$$\frac{\partial^2 f}{\partial t^2} - \left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right) f = 0$$

On obtient ainsi :

$$\begin{array}{rcl} U_1|_{(k+1)\Delta t} & = & U_1|_{k\Delta t} + \Delta t\dfrac{\partial U_1}{\partial t}|_{k\Delta t} + \dfrac{\Delta t^2}{2}\dfrac{\partial^2 U_1}{\partial t^2}|_{k\Delta t} + \dfrac{\Delta t^3}{6}\dfrac{\partial^3 U_1}{\partial t^3}|_{k\Delta t} + \mathcal{O}(\Delta t^4) \\ & = & U_1|_{k\Delta t} + \Delta t\dfrac{\partial U_3}{\partial y}|_{k\Delta t} + \dfrac{\Delta t^2}{2}\left(\dfrac{\partial^2}{\partial x^2}+\dfrac{\partial^2}{\partial y^2}\right)\left\{U_1|_{k\Delta t} + \dfrac{\Delta t}{3}\dfrac{\partial U_3}{\partial y}|_{k\Delta t}\right\} + \mathcal{O}(\Delta t^4) \end{array}$$

et

$$\begin{array}{rcl} U_2|_{(k+1)\Delta t} & = & U_2|_{k\Delta t} + \Delta t\dfrac{\partial U_2}{\partial t}|_{k\Delta t} + \dfrac{\Delta t^2}{2}\dfrac{\partial^2 U_2}{\partial t^2}|_{k\Delta t} + \dfrac{\Delta t^3}{6}\dfrac{\partial^3 U_2}{\partial t^3}|_{k\Delta t} + \mathcal{O}(\Delta t^4) \\ & = & U_2|_{k\Delta t} - \Delta t\dfrac{\partial U_3}{\partial x}|_{k\Delta t} + \dfrac{\Delta t^2}{2}\left(\dfrac{\partial^2}{\partial x^2}+\dfrac{\partial^2}{\partial y^2}\right)\left\{U_2|_{k\Delta t} - \dfrac{\Delta t}{3}\dfrac{\partial U_3}{\partial x}|_{k\Delta t}\right\} + \mathcal{O}(\Delta t^4) \end{array}$$

dont on projète la partie principale avec l'opérateur $\mathcal{R}_{N_1,\,M_1}$ pour obtenir $T_1^{k+1,0}$ et $T_2^{k+1,0}$.

L'algorithme s'arrête lorsque

$$|\,\widehat{\mathbf{U}}^{\mathbf{k+1},\nu+1} - \widehat{\mathbf{U}}^{\mathbf{k+1},\nu}\,|_2 < \varepsilon_{\text{relaxation}} \tag{3.64}$$

alors pour $\widehat{\mathbf{U}}^{\mathbf{k+1}}$, l'itéré suivant en temps, $\widehat{\mathbf{U}}^{\mathbf{k+1}} = \widehat{\mathbf{V}}^{\mathbf{k+1}} + \widehat{\mathbf{W}}^{\mathbf{k+1}} + \widehat{\mathbf{T}}^{\mathbf{k+1}}$, on prendra :

$$\left\{\begin{array}{lcl} \widehat{\mathbf{V}}^{\mathbf{k+1}} & = & \widehat{\mathbf{V}}^{\mathbf{k+1},\nu+1} \\ \widehat{\mathbf{W}}^{\mathbf{k+1}} & = & \widehat{\mathbf{W}}^{\mathbf{k+1},\nu+1} \\ \widehat{\mathbf{T}}^{\mathbf{k+1}} & = & \widehat{\mathbf{T}}^{\mathbf{k+1},\nu+1} \end{array}\right.$$

3.5.2.3 Schéma (CDIRK4).

Ce schéma nécessite la résolution de 3 systèmes linéaires, un par sous-pas de temps, chacun à l'aide d'un algorithme itératif.
Nous reprenons les notations de la décomposition pour $\widehat{\mathbf{Y}}^{\mathbf{i}}$, $(1 \leqslant i \leqslant 3)$, employées pour $\widehat{\mathbf{U}}^{\mathbf{k}}$, soit :

$$\widehat{\mathbf{Y}}^{\mathbf{i,k+1}} = \widehat{\mathbf{V}}^{\mathbf{k+1}} + \widehat{\mathbf{W}}^{\mathbf{k+1}} + \widehat{\mathbf{T}}^{\mathbf{k+1}}$$

Nous allons donc présenter ici l'algorithme du point fixe de manière générique avec la présence dans le second membre d'un terme $\widehat{\mathbf{B}}^{\mathbf{k}}$ correspondant aux contributions des sous-pas précédents (s'il y a lieu).
Soit $\widehat{\mathbf{T}}^{\mathbf{k+1,0}}$ donné, alors pour $\nu \geqslant 0$, l'indice des itérations, on a :

$$\left\{\begin{array}{l} \begin{array}{lcl} \widehat{\mathbf{W}}^{\mathbf{k+1},\nu+1} & = & \widehat{\mathbf{W}}^{\mathbf{k}} \\ & + & \dfrac{1+\xi}{2}\Delta t \mathcal{Q}_{N_1,\,M_1}\mathcal{M}_{\mathcal{A}}\left\{\left(\widehat{\mathbf{W}}^{\mathbf{k+1},\nu+1} + \widehat{\mathbf{T}}^{\mathbf{k+1},\nu}\right) + \left(\widehat{\mathbf{W}}^{\mathbf{k}} + \widehat{\mathbf{T}}^{\mathbf{k}}\right)\right\} \\ & + & \mathcal{Q}_{N_1,\,M_1}\widehat{\mathbf{B}}^{\mathbf{k}} \end{array} \\ \\ \left\{\begin{array}{l} \begin{array}{lcl} \widehat{\mathbf{V}}^{\mathbf{k+1},\nu+1} & = & \widehat{\mathbf{V}}^{\mathbf{k}} + \dfrac{1+\xi}{2}\Delta t \mathcal{M}_{\mathcal{A}}\left\{\widehat{\mathbf{V}}^{\mathbf{k+1},\nu+1} + \widehat{\mathbf{V}}^{\mathbf{k}}\right\} \\ & + & \dfrac{1+\xi}{2}\Delta t \mathcal{P}_{N_1,\,M_1}\mathcal{M}_{\mathcal{A}}\left\{\left(\widehat{\mathbf{W}}^{\mathbf{k+1},\nu+1} + \widehat{\mathbf{T}}^{\mathbf{k+1},\nu+1}\right) + \left(\widehat{\mathbf{W}}^{\mathbf{k}} + \widehat{\mathbf{T}}^{\mathbf{k}}\right)\right\} \\ & + & \mathcal{P}_{N_1,\,M_1}\widehat{\mathbf{B}}^{\mathbf{k}} \end{array} \\ \\ \mathcal{M}^{\mathrm{bc}}_{x,y}\left(\widehat{\mathbf{V}}^{\mathbf{k+1},\nu+1} + \widehat{\mathbf{T}}^{\mathbf{k+1},\nu+1}\right) = -\mathcal{M}^{\mathrm{bc}}_{x,y}\widehat{\mathbf{W}}^{\mathbf{k+1},\nu+1} \end{array}\right. \end{array}\right. \tag{3.65}$$

L'algorithme s'arrête lorsque

$$|\left(\mathbf{V}^{k+1,\nu+1} + \mathbf{W}^{k+1,\nu+1} + \mathbf{T}^{k+1,\nu+1}\right) - \left(\mathbf{V}^{k+1,\nu} + \mathbf{W}^{k+1,\nu} + \mathbf{T}^{k+1,\nu}\right)|_2 < \varepsilon_{\text{relaxation}}$$

alors

$$\left\{\begin{array}{lcl} \mathcal{P}_{N_1,\,M_1}\widehat{\mathbf{Y}}^{\mathbf{i,k+1}} & = & \widehat{\mathbf{V}}^{\mathbf{k+1},\nu+1} \\ \mathcal{Q}_{N_1,\,M_1}\widehat{\mathbf{Y}}^{\mathbf{i,k+1}} & = & \widehat{\mathbf{W}}^{\mathbf{k+1},\nu+1} \\ \mathcal{R}_{N_1,\,M_1}\widehat{\mathbf{Y}}^{\mathbf{i,k+1}} & = & \widehat{\mathbf{T}}^{\mathbf{k+1},\nu+1} \end{array}\right.$$

où $\widehat{\mathbf{Y}}^{\mathbf{i,k+1}}$ désigne la solution du $i^{\text{ème}}$ sous-pas du passage de l'étape k à l'étape $(k+1)$.

3.5.2.4 Etude de la convergence du point fixe.

L'algorithme du point fixe peut se mettre sous la forme

$$C\widehat{\mathbf{U}}^{\mathbf{k+1},\nu+1} = A\widehat{\mathbf{U}}^{\mathbf{k+1},\nu} + \widehat{\mathbf{B}} \tag{3.66}$$

avec $\widehat{\mathbf{U}}^{\mathbf{k+1,0}}$ donné,
$\widehat{\mathbf{B}}$ est un second membre constant, indépendant en particulier de ν,
C est une matrice inversible.

Notons $\widehat{\mathbf{U}}^{\mathbf{k+1}}$ la solution exacte de l'algorithme, elle vérifie donc

$$C\widehat{\mathbf{U}}^{\mathbf{k+1}} = A\widehat{\mathbf{U}}^{\mathbf{k+1}} + \widehat{\mathbf{B}} \tag{3.67}$$

A partir de là, déterminons des conditions nécessaires et suffisantes pour qu'il converge. On a

$$C\widehat{\mathbf{U}}^{\mathbf{k+1},\nu+1} = A\widehat{\mathbf{U}}^{\mathbf{k+1},\nu} + \widehat{\mathbf{B}}$$

et

$$C\widehat{\mathbf{U}}^{\mathbf{k+1}} = A\widehat{\mathbf{U}}^{\mathbf{k+1}} + \widehat{\mathbf{B}}$$

alors

$$\widehat{\mathbf{U}}^{\mathbf{k+1}} - \widehat{\mathbf{U}}^{\mathbf{k+1},\nu+1} = C^{-1}\left\{\left(A\widehat{\mathbf{U}}^{\mathbf{k+1}} + \widehat{\mathbf{B}}\right) - \left(A\widehat{\mathbf{U}}^{\mathbf{k+1},\nu} + \widehat{\mathbf{B}}\right)\right\}$$

d'où

$$\begin{aligned} |\,\widehat{\mathbf{U}}^{\mathbf{k+1}} - \widehat{\mathbf{U}}^{\mathbf{k+1},\nu+1}\,|_2 &\leqslant |\,C^{-1}A\,|_2^* \;\; |\,\widehat{\mathbf{U}}^{\mathbf{k+1}} - \widehat{\mathbf{U}}^{\mathbf{k+1},\nu}\,|_2 \\ &\leqslant \left(|\,C^{-1}A\,|_2^*\right)^\nu \;\; |\,\widehat{\mathbf{U}}^{\mathbf{k+1}} - \widehat{\mathbf{U}}^{\mathbf{k+1,0}}\,|_2 \end{aligned} \tag{3.68}$$

avec $|\,.\,|_2$ la norme L^2 vectorielle,
$|\,.\,|_2^*$ la norme L^2 matricielle induite.

Pour que l'algorithme soit convergent, il faut et il suffit que

$$|\,C^{-1}A\,|_2^* < 1$$

Pour cela, nous devons estimer

$$|\,C^{-1}A\,|_2^* = \left[\rho\left(C^{-1}A\right)^{\mathrm{T}}\left(C^{-1}A\right)\right]^{1/2} = \lambda_{\mathrm{MAX}}$$

Les termes de la matrice $(C^{-1}A)$ ne permettent pas de déterminer analytiquement λ_{MAX}; nous sommes obligés de faire une estimation numérique de cette valeur.
Soit $F = (C^{-1}A)^{\mathrm{T}}(C^{-1}A)$ alors $\lambda_{\mathrm{MAX}} = \rho(F)^{1/2}$.
Nous utilisons deux méthodes différentes pour obtenir une valeur approchée de λ_{MAX}.

Tout d'abord, pour localiser l'ensemble des valeurs propres de F, nous appliquons le théorème (3.7) de Gerschgörin-Hadamard ([36]).

Tab. 3.10 – *Rayon spectral de F pour (CN2).*

$N = M = 32$			$N = M = 64$			$N = M = 128$		
Δt	ρ^*	λ_{MAX}	Δt	ρ^*	λ_{MAX}	Δt	ρ^*	λ_{MAX}
10^{-3}	0,1962	0,1494	7.10^{-4}	1,1086	0,7805	2.10^{-4}	1,4483	0,9998
			6.10^{-4}	0,8003	0,5687	10^{-4}	0,3377	0,2417

Théorème *3.7*

Soit A une matrice carrée d'ordre N.
Les valeurs propres de A appartiennent l'union des N disques D_k du plan complexe, soit :

$$\lambda \in \bigcup_{k=1}^{N} D_k$$

où D_k, appelé le disque de Gerschgörin est défini par :

$$|\, z - a_{k,k} \,| \leqslant \Lambda_k = \sum_{\substack{j=1 \\ j\neq k}}^{N} |\, a_{k,j} \,|$$

Comme la matrice F est à coefficients réels, alors tous les disques sont centrés sur l'axe des abscisses. On obtient ρ^*, un majorant de $\rho\,(F)^{1/2}$, en prenant :

$$\rho^* = \max_k \{|\, a_{k,k} - \Lambda_k \,|, |\, a_{k,k} + \Lambda_k \,|\} = \max_k \{|\, a_{k,k} \,| + \Lambda_k\}^{1/2} \tag{3.69}$$

L'autre méthode consiste à calculer directement λ_{MAX} à l'aide de Matlab.
Les valeurs de ρ^* et de λ_{MAX} suivant les valeurs de Δt, N, M sont données dans les tableaux (3.10), (3.11). On impose $N_1 = 4$, $M_1 = 4$.

TAB. 3.11 – *Rayon spectral de F pour (CDIRK4).*

$N = M = 32$			$N = M = 64$			$N = M = 128$		
Δt	ρ^*	λ_{MAX}	Δt	ρ^*	λ_{MAX}	Δt	ρ^*	λ_{MAX}
10^{-3}	0,6429	0,4738	4.10^{-4}	1,4804	1,0590	9.10^{-5}	1,1632	0,8269
			3.10^{-4}	0,8402	0,6054	8.10^{-5}	0,9184	0,6545

Pour le cas $(N, M) = (32, 32)$, le pas de temps choisi $\Delta t = 10^{-3}$ fournit des résultats satisfaisants; c'est pour cette raison qu'il n'y a qu'une seule ligne remplie dans le tableau. La contrainte sur Δt est plus sévère que la condition trouvée pour obtenir la dominance diagonale de la matrice $\widetilde{M}_3$ dans le cas de la méthode classique.
Le choix des valeurs des paramètres de coupure N_1 et M_1 n'interviennent que peu ici pour la convergence de l'algorithme, comme on peut le voir en considérant les valeurs obtenues pour $N = M = 32$ (tableau 3.12). Ils ont une plus grande importance pour la résolution des équations en $\widehat{\mathbf{V}}^{\mathbf{k+1},\nu+1}$, $\widehat{\mathbf{W}}^{\mathbf{k+1},\nu+1}$ et $\widehat{\mathbf{T}}^{\mathbf{k+1},\nu+1}$, ce que nous verrons ultérieurement.

Remarque *12*

Dans la pratique nous prendrons comme critère d'arrêt $\varepsilon_{\text{relaxation}}$ de l'ordre de 10^{-15} ou 10^{-16}.

TAB. 3.12 – *Rayon spectral de F pour (CN2) et (CDIRK4).*

(CN2), $\Delta t = 10^{-3}$			(CDIRK4), $\Delta t = 10^{-3}$		
(N_1, M_1)	ρ^*	λ_{MAX}	(N_1, M_1)	ρ^*	λ_{MAX}
$(4,4)$	0,1962	0,1494	$(4,4)$	0,6429	0,4738
$(8,8)$	0,1954	0,1489	$(8,8)$	0,6404	0,4726
$(12,12)$	0,1943	0,1479	$(12,12)$	0,6366	0,4696
$(16,16)$	0,1926	0,1458	$(16,16)$	0,6293	0,4629

3.5.2.5 Résolution de l'équation pour $\widehat{\mathbf{W}}^{\mathbf{k+1},\nu+1}$.

Nous traitons cette résolution pour les deux schémas (CN2) et (CDIRK4) à la fois. Les deux seules différences sont la valeur du scalaire α et le second membre $\widehat{\mathbf{B}}^{\mathbf{k}}_{\text{W}} = \mathcal{Q}_{N_1, M_1}\widehat{\mathbf{B}}^{\mathbf{k}}$. L'équation à résoudre peut alors se mettre sous la forme :

$$\widehat{\mathbf{W}}^{\mathbf{k+1},\nu+1} = \alpha\Delta t\mathcal{Q}_{N_1, M_1}\mathcal{M}_{\mathcal{A}}\left(\widehat{\mathbf{W}}^{\mathbf{k+1},\nu+1} + \widehat{\mathbf{T}}^{\mathbf{k+1},\nu}\right) + \widehat{\mathbf{B}}^{\mathbf{k}}_{\text{W}} \tag{3.70}$$

où $\widehat{\mathbf{B}}^{\mathbf{k}}_{\text{W}}$ contient les contributions de l'étape k (ainsi que celle des sous-pas précédents pour (CDIRK4)).

$$\alpha = \begin{cases} \dfrac{1}{2} & \text{pour} \quad \text{(CN2)} \\ \dfrac{1+\xi}{2} & \text{pour} \quad \text{(CDIRK4)} \end{cases}$$

Ainsi nous avons :

$$\left(\mathbf{I}_{\text{W}} - \alpha\Delta t\mathcal{Q}_{N_1, M_1}\mathcal{M}_{\mathcal{A}}\right)\widehat{\mathbf{W}}^{\mathbf{k+1},\nu+1} = \mathcal{Q}_{N_1, M_1}\mathcal{M}_{\mathcal{A}}\widehat{\mathbf{T}}^{\mathbf{k+1},\nu} + \widehat{\mathbf{B}}^{\mathbf{k}}_{\text{W}} \tag{3.71}$$

On désigne par I_{W_i} $(1 \leqslant i \leqslant 3)$, la matrice identité pour chacun des 3 sous-espaces de $\mathcal{L}_{N,M}$ associé à la projection des inconnues dans les hautes fréquences. Ce qui s'écrit de manière plus détaillée :

$$\begin{aligned} &\begin{pmatrix} I_{\text{W}_1} & 0 & -\alpha\Delta t Q_{N,\,M-2} D_y \\ 0 & I_{\text{W}_2} & \alpha\Delta t Q_{N-2,\,M} D_x \\ -\alpha\Delta t Q_{N,\,M} D_y & \alpha\Delta t Q_{N,\,M} D_x & I_{\text{W}_3} \end{pmatrix} \begin{pmatrix} \widehat{W}_1^{k+1,\nu+1} \\ \widehat{W}_2^{k+1,\nu+1} \\ \widehat{W}_3^{k+1,\nu+1} \end{pmatrix} = \\ &\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\alpha\Delta t Q_{N,\,M} D_y & \alpha\Delta t Q_{N,\,M} D_x & 0 \end{pmatrix} \begin{pmatrix} \widehat{T}_1^{k+1,\nu} \\ \widehat{T}_2^{k+1,\nu} \\ 0 \end{pmatrix} + \begin{pmatrix} \widehat{B}^k_{\text{W}_1} \\ \widehat{B}^k_{\text{W}_2} \\ \widehat{B}^k_{\text{W}_3} \end{pmatrix} \end{aligned} \tag{3.72}$$

On désigne par I_{W_i} $(1 \leqslant i \leqslant 3)$, la matrice identité pour chacun des 3 sous-espaces de $\mathcal{L}_{N,M}$ associé à la projection des inconnues dans les hautes fréquences.
Pour résoudre ce système linéaire, nous allons procéder de manière similaire à la méthode classique, en 3 étapes :

(i) nous exprimons $\widehat{W}_1^{k+1,\nu+1}$ et $\widehat{W}_2^{k+1,\nu+1}$ en fonction de $\widehat{W}_3^{k+1,\nu+1}$;

(ii) nous résolvons alors l'équation en $\widehat{W}_3^{k+1,\nu+1}$;

(iii) nous obtenons enfin $\widehat{W}_1^{k+1,\nu+1}$ et $\widehat{W}_2^{k+1,\nu+1}$
à l'aide de leur expression trouvées en (i).

Ainsi

$$\left\{ \begin{array}{rcrl} \widehat{W}_1^{k+1,\nu+1} & = & \alpha\,\Delta t & Q_{N,\,M-2} D_y \widehat{W}_3^{k+1,\nu+1} + \widehat{B}_{\mathrm{W}_1}^k \\ \widehat{W}_2^{k+1,\nu+1} & = & -\alpha\,\Delta t & Q_{N-2,\,M} D_x \widehat{W}_3^{k+1,\nu+1} + \widehat{B}_{\mathrm{W}_2}^k \end{array} \right. \tag{3.73}$$

De là, on obtient

$$\left\{ I_{\mathrm{W}_3} - \alpha^2 \Delta t^2 \left[Q_{N,\,M} D_y Q_{N,\,M-2} D_y + Q_{N,\,M} D_x Q_{N-2,\,M} D_x \right] \right\} \widehat{W}_3^{k+1,\nu+1} = \widetilde{B}_{\mathrm{W}_3}^{k,\nu} \tag{3.74}$$

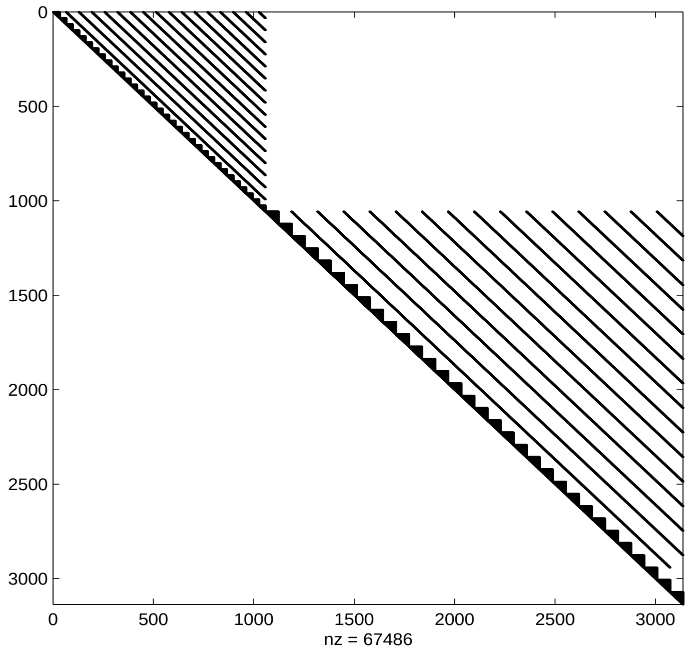
avec

$$\begin{array}{rl} \widetilde{B}_{\mathrm{W}_3}^{k,\nu} = \widehat{B}_{\mathrm{W}_3}^k & + \; \alpha\Delta t\, Q_{N,\,M} D_y \widehat{B}_{\mathrm{W}_1}^k \; - \; \alpha\Delta t\, Q_{N,\,M} D_x \widehat{B}_{\mathrm{W}_2}^k \\ & - \; \alpha\Delta t\, Q_{N,\,M} D_y \widehat{T}_1^{k+1,\nu} \; + \; \alpha\Delta t\, Q_{N,\,M} D_x \widehat{T}_2^{k+1,\nu} \end{array}$$

soit encore

$$\widetilde{M}_{\mathrm{W}_3}\ \widehat{W}_3^{k+1,\nu+1} = \widetilde{B}_{\mathrm{W}_3}^{k,\nu} \tag{3.75}$$

FIG. 3.3 – *Squelette de la matrice $\widetilde{M}_{W_3}$ pour (CN2) avec $(N, M) = (64, 64)$ et $(N_1, M_1) = (32, 32)$.*



La matrice $\widetilde{M}_{\mathrm{W}_3}$ est très intéressante puisqu'elle est triangulaire supérieure avec tous les coefficients diagonaux égaux à 1.
Elle s'inverse donc exactement très facilement. On peut aussi expliciter ses coefficients à l'aide des relations de dérivation dans l'espace spectral pour les polynômes de Legendre. Connaissant $\widehat{W}_3^{k+1,\nu+1}$, nous pouvons déterminer $\widehat{W}_1^{k+1,\nu+1}$ et $\widehat{W}_2^{k+1,\nu+1}$ à l'aide des relations (3.73).
D'autre part, il est clair que l'on a intérêt à rendre $\widehat{\mathbf{W}}$ prépondérant dans la décomposition $\widehat{\mathbf{U}} = \widehat{\mathbf{V}} + \widehat{\mathbf{W}} + \widehat{\mathbf{T}}$.
On verra que l'inversion exacte de la matrice issue du système linéaire relatif aux hautes fréquences n'est pas aussi facilement applicable au système linéaire des basses fréquences couplées aux "modes Tau", en raison des lignes pleines provenant de l'imposition des conditions aux limites.

3.5.2.6 Résolution de l'équation pour $\widehat{\mathbf{V}}^{\mathbf{k+1},\nu+\mathbf{1}}$ et $\widehat{\mathbf{T}}^{\mathbf{k+1},\nu+\mathbf{1}}$.

Le vecteur $\widehat{\mathbf{W}}^{\mathbf{k+1},\nu+\mathbf{1}}$ est maintenant supposé connu; il figurera donc au second membre. De même que pour l'équation en $\widehat{\mathbf{W}}^{\mathbf{k+1},\nu+\mathbf{1}}$, nous traitons en une fois les deux schémas (CN2) et (CDIRK4).

Ainsi nous avons le système linéaire suivant à résoudre, avec $\alpha = \dfrac{1}{2}$ ou $\alpha = \dfrac{1+\xi}{2}$:

$$\left\{\begin{array}{l} \widehat{\mathbf{V}}^{\mathbf{k+1},\nu+\mathbf{1}} = \alpha\Delta t \mathcal{M}_{\mathcal{A}} \widehat{\mathbf{V}}^{\mathbf{k+1},\nu+\mathbf{1}} + \alpha\Delta t\, \mathcal{P}_{N_1,\,M_1} \mathcal{M}_{\mathcal{A}} \widehat{\mathbf{T}}^{\mathbf{k+1},\nu+\mathbf{1}} + \widehat{\mathbf{B}}^{\mathbf{k}}_{\mathrm{V}} \\ \\ \mathcal{M}^{\mathrm{bc}}_{x,y} \left(\widehat{\mathbf{V}}^{\mathbf{k+1},\nu+\mathbf{1}} + \widehat{\mathbf{T}}^{\mathbf{k+1},\nu+\mathbf{1}}\right) = -\mathcal{M}^{\mathrm{bc}}_{x,y} \widehat{\mathbf{W}}^{\mathbf{k+1},\nu+\mathbf{1}} \end{array}\right. \tag{3.76}$$

où $\widehat{\mathbf{B}}^{\mathbf{k}}_{\mathrm{V}} = \mathcal{P}_{N_1,\,M_1} \widehat{\mathbf{B}}^{\mathbf{k}}$ correspond aux contributions de l'étape k, les termes provenant de $\widehat{\mathbf{W}}^{\mathbf{k+1},\nu+\mathbf{1}}$ et des sous-pas pour (CDIRK4) s'il y a lieu. Le système précédent s'écrit encore

$$\begin{pmatrix} I_{\mathrm{V}} & 0 & -\alpha\Delta t D_y & 0 & 0 \\ 0 & I_{\mathrm{V}} & \alpha\Delta t D_x & 0 & 0 \\ -\alpha\Delta t D_y & \alpha\Delta t D_x & I_{\mathrm{V}} & -\alpha\Delta t P_{N_1,\,M_1} D_y & \alpha\Delta t P_{N_1,\,M_1} D_x \\ M^{bc}_y & 0 & 0 & I_{\mathrm{T}_1} & 0 \\ 0 & M^{bc}_x & 0 & 0 & I_{\mathrm{T}_2} \end{pmatrix} \begin{pmatrix} \widehat{V}^{k+1,\nu+1}_1 \\ \widehat{V}^{k+1,\nu+1}_2 \\ \widehat{V}^{k+1,\nu+1}_3 \\ \widehat{T}^{k+1,\nu+1}_1 \\ \widehat{T}^{k+1,\nu+1}_2 \end{pmatrix}$$
$$= \begin{pmatrix} \widehat{B}^k_{\mathrm{V}_1} \\ \widehat{B}^k_{\mathrm{V}_2} \\ \widehat{B}^k_{\mathrm{V}_3} \\ \widehat{B}^{k+1,\nu+1}_{\mathrm{T}_1} \\ \widehat{B}^{k+1,\nu+1}_{\mathrm{T}_2} \end{pmatrix} = \begin{pmatrix} \widehat{B}_{\mathrm{V}^k_1} \\ \widehat{B}_{\mathrm{V}^k_2} \\ \widehat{B}_{\mathrm{V}^k_3} \\ -M^{bc}_y \widehat{W}^{k+1,\nu+1}_1 \\ -M^{bc}_x \widehat{W}^{k+1,\nu+1}_2 \end{pmatrix} \tag{3.77}$$

La matrice de ce système linéaire est réelle, non symétrique, de rang $3(N_1+1)(M_1+1) + 2[(N+1)+(M+1)]$, non à diagonale strictement dominante. Nous allons nous ramener à l'aide de substitutions successives à un système linéaire dont la matrice est réelle, non symétrique, de rang $(N_1+1)(M_1+1)$ à diagonale strictement dominante. On déduit du système linéaire (3.77) les quatre relations suivantes

$$\left\{\begin{array}{lcl} \widehat{T}^{k+1,\nu+1}_1 = \widehat{B}^{k+1,\nu+1}_{\mathrm{T}_1} & - & M^{bc}_y \widehat{V}^{k+1,\nu+1}_1 \\ \widehat{T}^{k+1,\nu+1}_2 = \widehat{B}^{k+1,\nu+1}_{\mathrm{T}_2} & - & M^{bc}_x \widehat{V}^{k+1,\nu+1}_2 \\ \widehat{V}^{k+1,\nu+1}_1 = \widehat{B}^k_{\mathrm{V}_1} & + & \alpha\Delta t D_y \widehat{V}^{k+1,\nu+1}_3 \\ \widehat{V}^{k+1,\nu+1}_2 = \widehat{B}^k_{\mathrm{V}_2} & - & \alpha\Delta t D_x \widehat{V}^{k+1,\nu+1}_3 \end{array}\right. \tag{3.78}$$

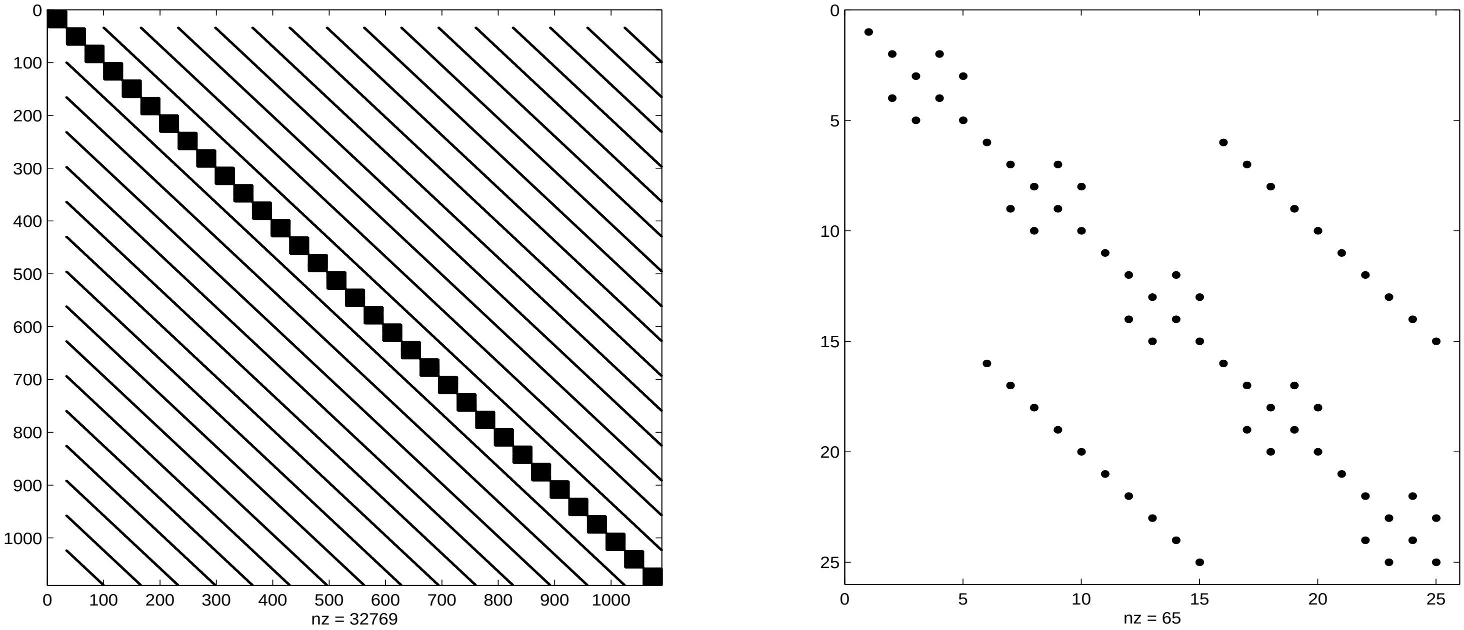
On substitue dans l'équation de $\widehat{V}^{k+1,\nu+1}_3$ les deux premières relations de (3.78), ensuite les deux suivantes pour n'obtenir ainsi que des termes en $\widehat{V}^{k+1,\nu+1}_3$. Ainsi celle-ci s'écrit

$$\left\{I_{\mathrm{V}} - \alpha^2\Delta t^2 \left[D^2_x + D^2_y - P_{N_1,\,M_1} D_y M^{bc}_y D_y - P_{N_1,\,M_1} D_x M^{bc}_x D_x\right]\right\} \widehat{V}^{k+1,\nu+1}_3 = \widetilde{B}^{k,\nu+1}_{\mathrm{V}_3} \tag{3.79}$$

avec

$$\begin{array}{rl} \widetilde{B}^{k,\nu+1}_{\mathrm{V}_3} = \widehat{B}^k_{\mathrm{V}_3} & + \ \alpha\Delta t \left[D_y - P_{N_1,\,M_1} D_y M^{bc}_y\right] \widehat{B}^k_{\mathrm{V}_1} \ + \ \alpha\Delta t\, P_{N_1,\,M_1} D_y \widehat{B}^{k+1,\nu+1}_{\mathrm{T}_1} \\ & - \ \alpha\Delta t \left[D_x - P_{N_1,\,M_1} D_x M^{bc}_x\right] \widehat{B}^k_{\mathrm{V}_2} \ - \ \alpha\Delta t\, P_{N_1,\,M_1} D_x \widehat{B}^{k+1,\nu+1}_{\mathrm{T}_2} \end{array}$$

FIG. 3.4 – *Squelette de la matrice* $\widetilde{M}_{\mathrm{V}_3}$ *pour (CN2) avec* $(N, M) = (64, 64)$ *et* $(N_1, M_1) = (32, 32)$ *(à gauche) ainsi que* $(N_1, M_1) = (4, 4)$ *(à droite).*



Soit encore

$$\widetilde{M}_{\mathrm{V}_3}\ \widehat{V}_3^{k+1,\nu+1} = \widetilde{B}_{\mathrm{V}_3}^{k,\nu+1} \tag{3.80}$$

Le tableau (3.13) présente les valeurs numériques de $\mathrm{cond}(\widetilde{M}_{\mathrm{V}_3})$ obtenues avec Matlab, pour $N = M = 64$, $\Delta t = 10^{-3}$ et différentes valeurs de N_1 et M_1. Nous n'avons aucune contrainte pour le choix des entiers N_1, M_1. Le splitting considéré est indépendant des opérateurs de dérivation partielle et n'a pour but que le traitement des conditions aux limites. Autant l'équation

$$\widetilde{M}_{\mathrm{W}_3}\ \widehat{W}_3^{k+1,\nu+1} = \widetilde{B}_{\mathrm{W}_3}^{k,\nu}$$

que l'équation

$$\widetilde{M}_{\mathrm{V}_3}\ \widehat{V}_3^{k+1,\nu+1} = \widetilde{B}_{\mathrm{V}_3}^{k,\nu+1}$$

peut être résolue à l'aide de l'algorithme du Bi-CG Stab présenté précédemment. Le conditionnement de ces matrices est très bon, voir le tableau (3.13). On peut alors escompter un faible nombre d'itérations comme pour la méthode classique. Cependant, le découpage de $\widehat{\mathbf{U}}$ en $\widehat{\mathbf{V}}$, $\widehat{\mathbf{W}}$ et $\widehat{\mathbf{T}}$ nous permet de rendre la matrice $\widetilde{M}_{\mathrm{V}_3}$ petite et donc de calculer facilement son inverse et cela de manière analytique. Pour des valeurs plus grandes de N_1 et M_1, cela reste bien entendu toujours possible bien que plus compliqué. De plus comme $\widetilde{M}_{\mathrm{W}_3}$ est triangulaire supérieure avec des 1 sur la diagonale un choix judicieux de N_1, M_1 nous fournit un système global plus économiquement inversible.
De là, nous prendrons $N_1 = M_1 = 4$ pour les schémas (CN2) et (CDIRK4). On peut alors calculer explicitement les coefficients de $\widetilde{M}_{\mathrm{V}_3}$. La figure (3.4) représente son squelette et on peut constater que le système linéaire peut se découpler en plusieurs systèmes de rang 2 et 4. Ce splitting est possible grâce aux propriétés de parité des dérivées des polynômes de Legendre.

TAB. 3.13 – *Conditionnement de la matrice $\widetilde{M}_{V_3}$ pour (CN2) et (CDIRK4).*

(CN2)			(CDIRK4)		
(N_1,M_1)	Rang($\widetilde{M}_{V_3}$)	Cond($\widetilde{M}_{V_3}$)	(N_1,M_1)	Rang($\widetilde{M}_{V_3}$)	Cond($\widetilde{M}_{V_3}$)
(4,4)	25	1,0001	(4,4)	25	1,0002
(8,8)	81	1,0005	(8,8)	81	1,0023
(12,12)	169	1,0020	(12,12)	169	1,0094
(16,16)	289	1,0058	(16,16)	289	1,0266
(32,32)	1089	1,0785	(32,32)	1089	1,3587

3.5.2.7 Accélération de la convergence.

Pour le système linéaire type $Ax = b$, on écrit le processus itératif sous la forme :

$$x^{(\nu+1)} = Fx^{(\nu)} + b$$

avec ν l'indice des itérations,
b le second membre de ce système, indépendant en particulier de ν.

On introduit un paramètre réel ω appelé coefficient de relaxation qui nous servira à améliorer la rapidité de la convergence.
Soit $x^{(\nu)}$ calculé et $\tilde{x}^{(\nu+1)}$ obtenu à partir de l'algorithme du point fixe précédent. On définit alors la combinaison linéaire

$$\begin{aligned} x^{(\nu+1)} &= \omega\, \tilde{x}^{(\nu+1)} + (1-\omega)x^{(\nu)} \\ &= \omega\left[Fx^{(\nu)} + b\right] + (1-\omega)x^{(\nu)} \end{aligned}$$

Les résultats suivants nous fournissent des conditions nécessaire et suffisante de convergence :

Proposition *1*

Une condition nécessaire pour que la méthode de relaxation converge est

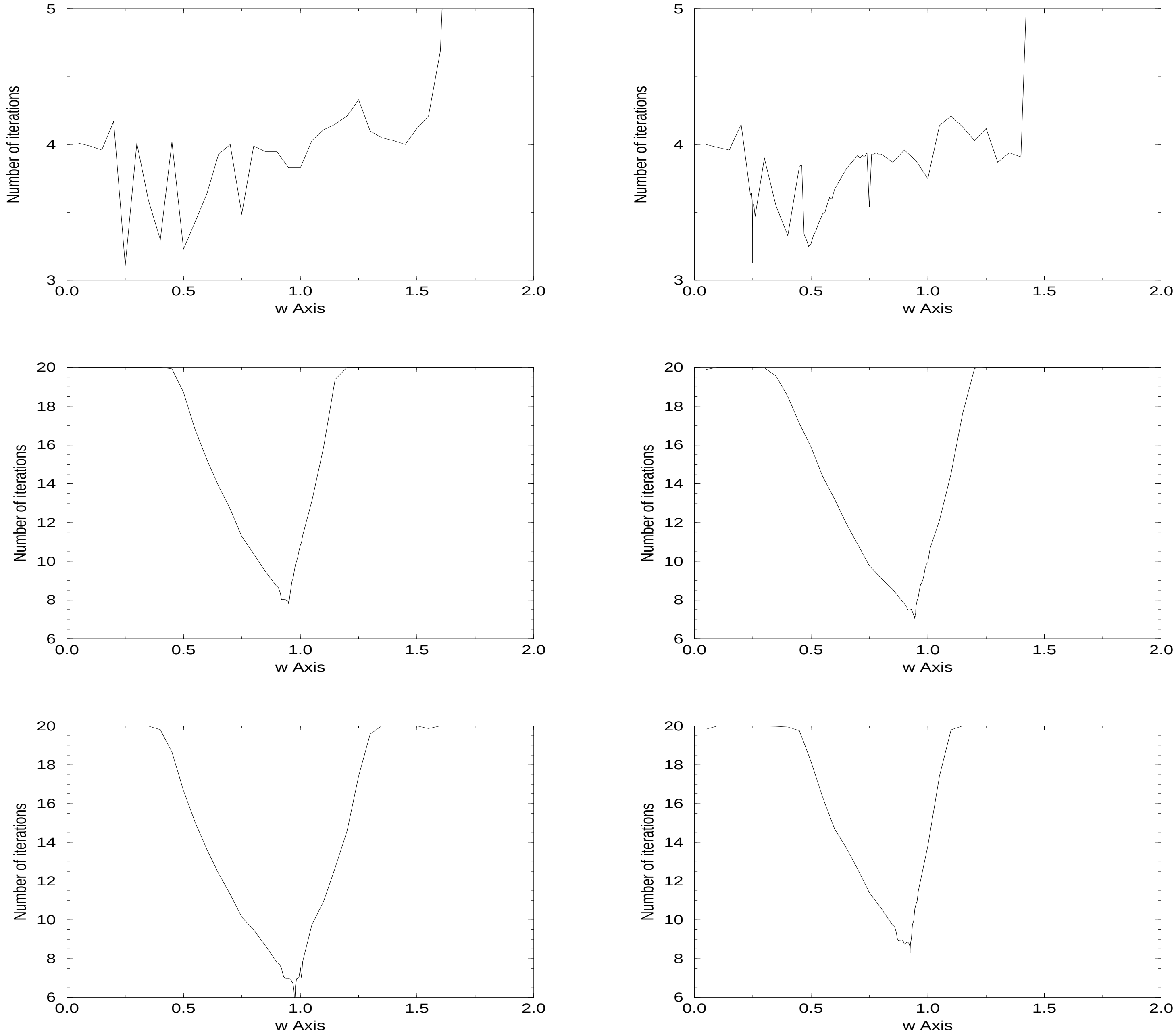
$$0 < \omega < 2$$

Proposition *2*

A étant une matrice à diagonale strictement dominante, si $0 < \omega \leqslant 1$ alors la méthode de relaxation est convergente.

Dans la littérature existante, Il y a peu de résultats théoriques concernant la détermination d'un paramètre optimal ω_b applicable à notre problème. Faire des tests est alors nécessaire pour déterminer une valeur ω_0 fournissant une accélération acceptable par rapport aux valeurs possibles de ω.
Ainsi, pour chacun des deux schémas (CN2) et (CDIRK4) et pour les trois cas tests considérés, nous effectuons la recherche de la manière suivante : nous fixons ω et faisons 100 itérations en temps. On calcule alors la moyenne des 100 valeurs du nombre d'itérations du point fixe. Les figures (3.5) représentent l'allure de la fonction $y = f(\omega)$ et on cherche à déterminer "un minimum acceptable". Le tableau (3.14) présente les valeurs obtenues pour ω_0 et le gain en terme moyen d'itérations par rapport au choix $\omega = 1$ qui correspond à l'algorithme de point fixe sans relaxation.

FIG. 3.5 – *Nombre moyen d'itérations en fonction de ω (CN2) à gauche et (CDIRK4) à droite.*



Pour chacun des trois tests, les courbes pour les deux schémas en temps sont assez semblables. Pour une raison qui reste encore à déterminer, les courbes pour le test 1 présentent une allure chaotique et la valeur ω_0 obtenue est surprenante.

TAB. 3.14 – *Comparaison des accélérations du point fixe.*

		ω_0	Nombre moyen d'itérations pour ω_0	Nombre moyen d'itérations pour $\omega = 1$	gain
Test 1	(CN2)	0,2500	3,11	3,83	19 %
	(CDIRK4)	0,2500	3,13	3,75	16,5 %
Test 2	(CN2)	0,9490	7,82	10,83	28 %
	(CDIRK4)	0,9445	7,05	9,94	29 %
Test 3	(CN2)	0,9765	6,00	7,51	20 %
	(CDIRK4)	0,9245	8,30	13,81	40 %

3.6 Analyse des résultats.

Nous avons effectué 3 tests, en augmentant les valeurs des composantes du vecteur d'onde, ce qui entraine une augmentation du nombre de modes nécessaires pour discrétiser la solution et de là une augmentation de la dimension des systèmes linéaires à inverser (du moins pour la méthode classique).
Pour chacun des tests, nous comparerons les erreurs globales, l'imposition des conditions aux limites, la conservation de l'invariant théorique ainsi que le temps calcul pour les différents schémas; à savoir, entre la version classique et la version à deux niveaux.

3.6.1 Test 1.

Nous prenons les mêmes valeurs du vecteur d'onde $\mathbf{k}$ que dans le cadre de la validation de la méthode classique, au paragraphe précédent, soit $\mathbf{k} = (k_1, k_2) = (3, 3)$. La "période" T est la même: $T = \frac{2}{3\sqrt{2}} \approx 0,47$ qui correspond à la plus longue des quatre périodes. Les scalaires λ_{3i} valent respectivement 1, 2, -1, 1 pour $1 \leqslant i \leqslant 4$.

3.6.1.1 Méthode explicite (CRK4).

Du point de vue de notre étude, la méthode la moins intéressante est la méthode explicite, basée sur la version conservative du schéma de Runge-Kutta d'ordre 4 (CRK4). En effet, le caractère explicite de ce schéma rend inopérant toute projection des équations initiales. Numériquement se rajoute le fait que les boucles des algorithmes sont fractionnées en plusieurs de longueur inférieure ce qui amoindrit les performances du code.

TAB. 3.15 – *Normes et erreurs pour le schéma (CRK4) classique.*

$(t = 2T)$	$\lvert U_i\rvert_2$	$\lvert U_i\rvert_\infty$	$\lvert U_i - U_i^{\text{ex}}\rvert_2$	$\lvert U_i - U_i^{\text{ex}}\rvert_\infty$	B.C.
U_1	0.1567E+01	0.2120E+01	0.6058E-08	0.9310E-08	0.1375E-14
U_2	0.1595E+01	0.2226E+01	0.5885E-08	0.9148E-08	0.2802E-14
U_3	0.1414E+01	0.2350E+01	0.1335E-07	0.1727E-07	
$(t = 10T)$					
U_1	0.1578E+01	0.1892E+01	0.3006E-07	0.4725E-07	0.5832E-15
U_2	0.1584E+01	0.2005E+01	0.2989E-07	0.4709E-07	0.5053E-14
U_3	0.1414E+01	0.2412E+01	0.6666E-07	0.7821E-07	

Cela se reflète dans les résultats, voir les tableaux (3.15), (3.16). Toutefois, il nous semble important ntéressant de considérer une intégration en temps explicite.
Le pas de temps, Δt, est déterminé de telle sorte que $\Delta t NM \approx \mathcal{O}(1)$, soit ici $\Delta t = \dfrac{T}{475} \approx 10^{-3}$. L'erreur commise sur l'invariant est de l'ordre de 3.10^{-15} aux instants $t = 2T$ et $t = 10T$.

TAB. 3.16 – *Normes et erreurs pour le schéma (CRK4) à 2 niveaux.*

$(t = 2T)$	$\lvert U_i \rvert_2$	$\lvert U_i \rvert_\infty$	$\lvert U_i - U_i^{\mathrm{ex}} \rvert_2$	$\lvert U_i - U_i^{\mathrm{ex}} \rvert_\infty$	B.C.
U_1	0.1567E+01	0.2120E+01	0.6058E-08	0.9310E-08	0.2450E-14
U_2	0.1595E+01	0.2226E+01	0.5885E-08	0.9148E-08	0.4753E-14
U_3	0.1414E+01	0.2350E+01	0.1335E-07	0.1727E-07	
$(t = 10T)$					
U_1	0.1578E+01	0.1892E+01	0.3006E-07	0.4725E-07	0.6128E-15
U_2	0.1584E+01	0.2005E+01	0.2989E-07	0.4709E-07	0.4077E-14
U_3	0.1414E+01	0.2412E+01	0.6666E-07	0.7821E-07	

Les temps calcul sont présentés dans le tableau (3.17).

TAB. 3.17 – *Temps calcul pour (CRK4) classique et (CRK4) à 2 niveaux.*

	(CRK4) classique	(CRK4) 2 niveaux
$(t = 2T)$	26 s	33 s
$(t = 10T)$	118 s	154 s

3.6.1.2 Méthode semi-implicite (CN2).

Pour ce schéma, nous appliquons les décompositions présentées au cours du paragraphe précédent. Les valeurs du critère d'arrêt de convergence du point fixe, $\varepsilon_{\text{relaxation}} \approx 10^{-15}$ ou 10^{-16} sont parfois difficiles à atteindre numériquement, pour cela nous avons formé un critère d'arrêt plus complexe basé sur les observations suivantes :
• Outre le test avec $\varepsilon_{\text{relaxation}}$, nous imposons un nombre maximal d'itérations pour éviter une non-convergence numérique du code : au bout de *maxiter* itérations, il y a sortie de l'algorithme du point-fixe.
• Mais cela mène parfois à faire des itérations qui n'améliorent pas pour autant la convergence et on complète le test avec le critère suivant : soit $\Delta_\nu = \lvert \widehat{\mathbf{U}}^{\mathbf{k+1},\nu+\mathbf{1}} - \widehat{\mathbf{U}}^{\mathbf{k+1},\nu} \rvert_2$;
si $\Delta_\nu - \Delta_{\nu-1} < \varepsilon_{\text{relaxation}}$ il y a sortie de la boucle du point fixe.
Ainsi le test d'arrêt de convergence est triple :

$$(\nu \geqslant maxiter)\ \mathbf{OU}\ \big(\Delta_\nu < \varepsilon_{\text{relaxation}}\big)\ \mathbf{OU}\ \big(|\Delta_\nu - \Delta_{\nu-1}| < \varepsilon_{\text{relaxation}}\big)$$

Nous prenons les valeurs suivantes pour *maxiter*: 2, 4, 6, 8, 10. Lorsque des figures illustrent un résultat, les numéros des courbes (de 1 à 5) en ordre croissant correspondent aux valeurs précédentes de *maxiter* données dans cet ordre. Dans le cas de figures avec des courbes numérotées de 1 à 6, 1 correspond à la méthode classique et de 2 à 6 aux valeurs de *maxiter* précédentes.

Une page de figures se scinde en deux groupes de 4 :

1	2		
3	4	celles du haut	présentent les résultats pour $t \in [0,\ 2T]$
5	6	celles du bas	présentent les résultats pour $t \in [0, 10T]$.
7	8		

Elles décrivent l'évolution en temps des quantités suivantes :

1. la norme L^∞ de U_3 sur $[0, 2T]$;
2. la différence entre les invariant théorique et calculé pour la méthode classique (1) et la méthode à deux niveaux (2 à 6) sur $[0, 2T]$;
3. le nombre d'itérations du point fixe de la méthode à deux niveaux sur $[0, 2T]$;
4. le résidu du point fixe sur $[0, 2T]$;
5. la norme L^2 de l'erreur absolue sur U_1 sur $[0, 10T]$;
6. la différence entre les invariant théorique et calculé pour la méthode classique (1) et la méthode à deux niveaux (2 à 6) sur $[0, 10T]$;
7. le nombre d'itérations du point fixe de la méthode à deux niveaux sur $[0, 10T]$;
8. le résidu du point fixe sur $[0, 10T]$.

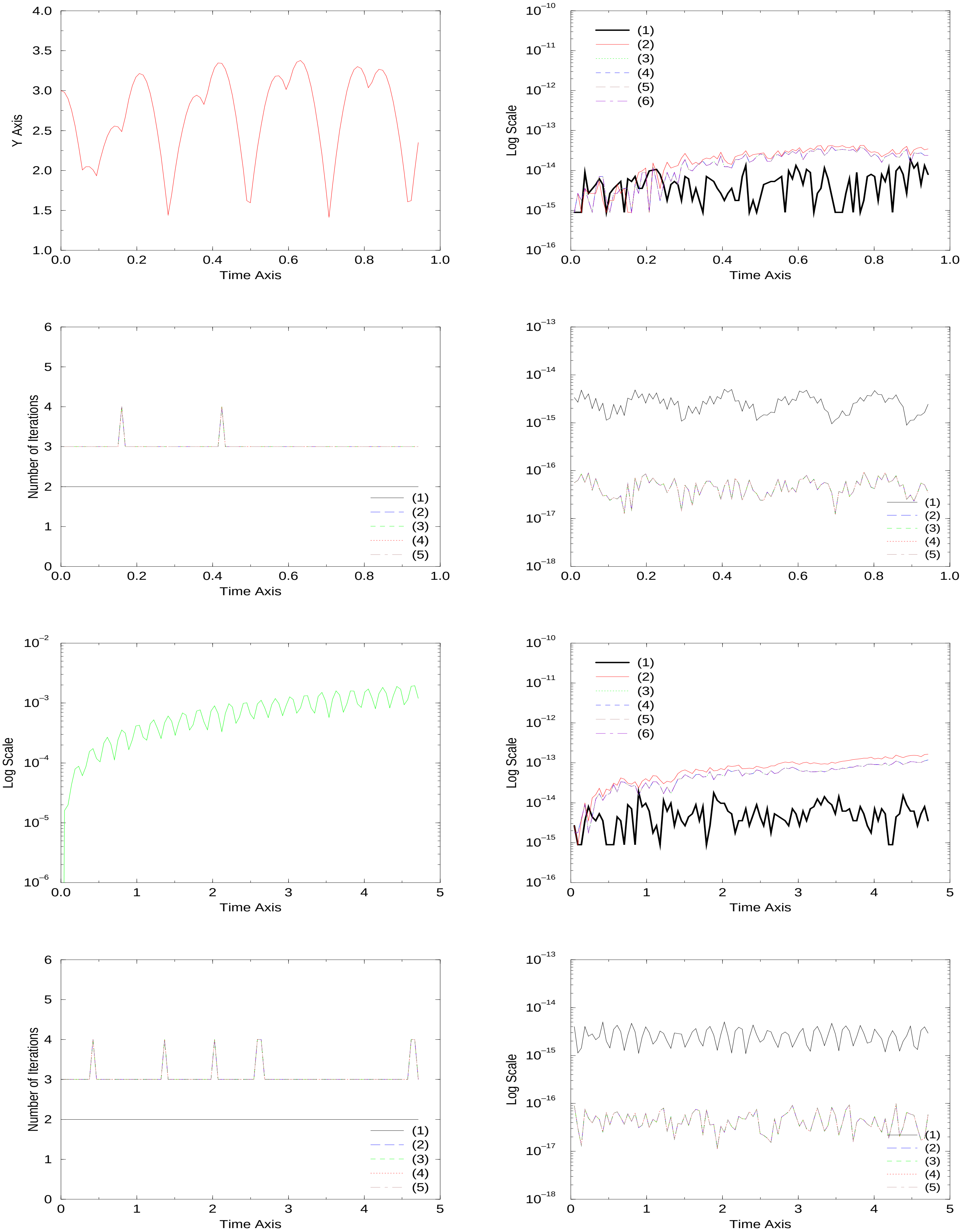
Les tracés de l'imposition des conditions aux limites ainsi que les spectres de la solution occupent une demie-page à part. Nous prenons $N_1 = M_1 = 4$ pour avoir facilement une inversion directe des matrices et le pas de temps est $\Delta t = 10^{-3}$, donné par le critère de convergence du point fixe et $\omega = 0,25$.

TAB. 3.18 – *Normes et erreurs pour (CN2) classique.*

$(t = 2T)$	$\lvert U_i \rvert_2$	$\lvert U_i \rvert_\infty$	$\lvert U_i - U_i^{\text{ex}} \rvert_2$	$\lvert U_i - U_i^{\text{ex}} \rvert_\infty$	B.C.
U_1	0.1567E+01	0.2120E+01	0.2427E-03	0.3794E-03	0.6854E-15
U_2	0.1595E+01	0.2226E+01	0.2345E-03	0.3724E-03	0.2259E-14
U_3	0.1414E+01	0.2350E+01	0.5284E-03	0.7094E-03	
$(t = 10T)$					
U_1	0.1577E+01	0.1890E+01	0.1199E-02	0.1929E-02	0.1931E-14
U_2	0.1583E+01	0.2005E+01	0.1190E-02	0.1922E-02	0.2826E-14
U_3	0.1416E+01	0.2415E+01	0.2641E-02	0.3192E-02	

Nous faisons plusieurs remarques.
Les conditions aux limites sont bien imposées, figure (3.8), avec des valeurs inférieures à 10^{-14} pour U_1 et U_2. Les spectres (pour U_2 à gauche, U_3 à droite) sont semblables pour toutes les simulations. Les normes et erreurs globales pour les deux méthodes sont très proches : les quatre premières décimales sont identiques, aussi bien sur $[0, 2T]$ que sur $[0, 10T]$, voir les tableaux (3.18), (3.19).

FIG. 3.6 – *Test 1 pour (CN2) :* $[0, 2T]$ *en haut et* $[0, 10T]$ *en bas.*

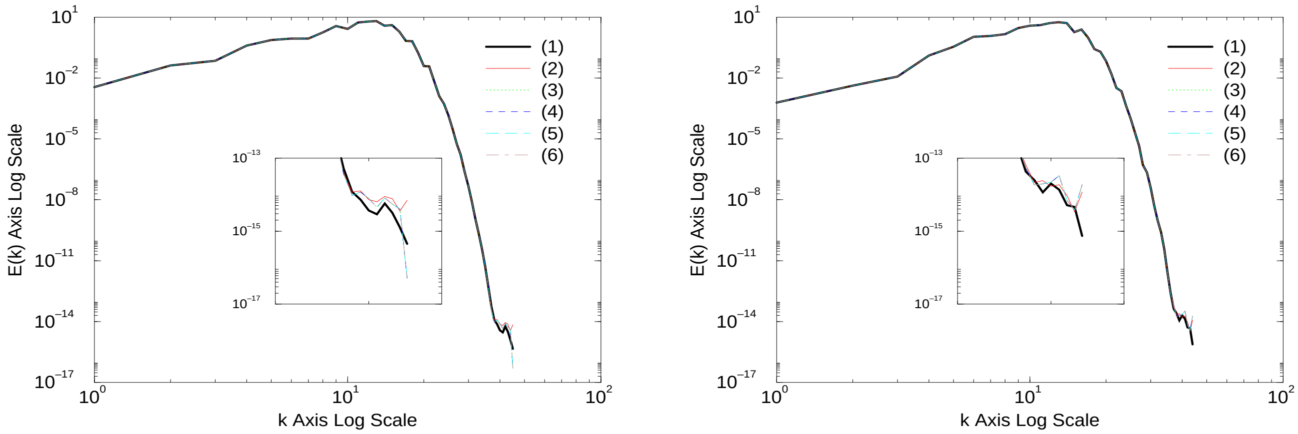


L'invariant est convenablement conservé avec une différence entre les invariants théorique et calculé majorée par 2.10^{-13}. On note que la conservation est meilleure pour la méthode classique.

TAB. 3.19 – *Normes et erreurs pour (CN2) maxiter = 2.*

$(t = 2T)$	$\vert U_i\vert_2$	$\vert U_i\vert_\infty$	$\vert U_i - U_i^{\text{ex}}\vert_2$	$\vert U_i - U_i^{\text{ex}}\vert_\infty$	B.C.
U_1	0.1567E+01	0.2120E+01	0.2427E-03	0.3794E-03	0.5665E-15
U_2	0.1595E+01	0.2226E+01	0.2345E-03	0.3724E-03	0.2319E-14
U_3	0.1414E+01	0.2350E+01	0.5284E-03	0.7094E-03	
$(t = 10T)$					
U_1	0.1577E+01	0.1890E+01	0.1199E-02	0.1929E-02	0.1053E-14
U_2	0.1583E+01	0.2005E+01	0.1190E-02	0.1922E-02	0.3230E-14
U_3	0.1416E+01	0.2415E+01	0.2641E-02	0.3192E-02	

FIG. 3.7 – *Spectre de Legendre pour (CN2) à $t = 2T$ et $t = 10T$.*



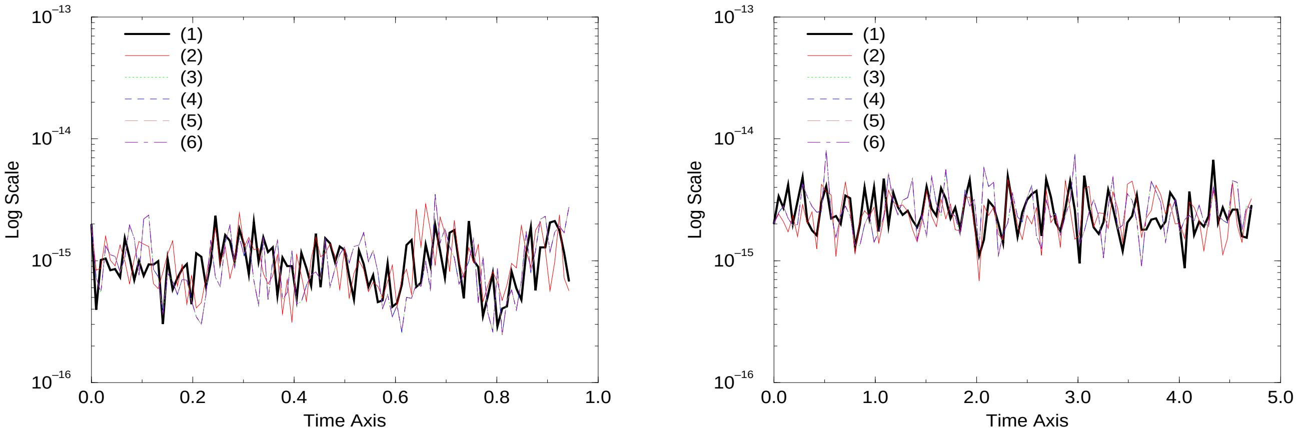
L'algorithme du point fixe nécessite généralement 3 itérations pour converger et 4 occasionnellement pour fournir un résidu inférieur à 10^{-16} : seul le cas avec *maxiter* = 2 se détache avec un résidu entre 10^{-14} et 10^{-15}.
A l'aide du tableau (3.20), nous constatons que les temps calcul pour la méthode à deux niveaux sont au mieux égaux à ceux de la méthode classique; autrement, il faut près de 50 % en plus pour une parfaite convergence du point fixe.

TAB. 3.20 – *Temps calcul pour (CN2) classique et (CN2) à 2 niveaux.*

	(CN2) cl	(CN2) 2	(CN2) 4	(CN2) 6	(CN2) 8	(CN2) 10
$(t = 2T)$	35 s	35 s	49 s	49 s	49 s	49 s
$(t = 10T)$	161 s	157 s	227 s	227 s	227 s	228 s

L'explication est simple :
bien que toutes les inconnues (les basses et les hautes fréquences de la troisième inconnue puis les autres par déduction) soient déterminées par l'inversion exacte d'une matrice, cela ne suffit pas à compenser les itérations du point fixe nécessaires à chaque pas de temps.

FIG. 3.8 – $U_1(x, y = \pm 1)$ *et* $U_2(x = \pm 1, y)$ *pour (CN2).*



A cela on peut ajouter le très conditionnement de la matrice $\widetilde{M_3}$, inversée par le bi-gradient conjugué stabilisé pour la méthode classique, tableau (3.21).

TAB. 3.21 – *Statistiques sur le nombre d'itérations du Bi-CG Stab pour (CN2).*

(CN2)	Nombre d'itérations				(CN2)	Nombre d'itérations			
$(t = 2T)$	Min	Max	Moy.	Ec-type	$(t = 10T)$	Min	Max	Moy.	Ec-type
cl	2	2	2	0	cl	2	2	2	0
(CN2)	Résidu du système linéaire				(CN2)	Résidu du système linéaire			
$(t = 2T)$	Min	Max	Moy.	Ec-type	$(t = 10T)$	Min	Max	Moy.	Ec-type
cl	4E-21	1E-20	7E-21	2E-21	cl	4E-21	2E-20	1E-20	5E-21

3.6.1.3 Méthode semi-implicite (CDIRK4).

De même que pour le schéma (CN2), nous utilisons la décomposition de la solution

$$\mathbf{U} = \mathbf{V} + \mathbf{W} + \mathbf{T}$$

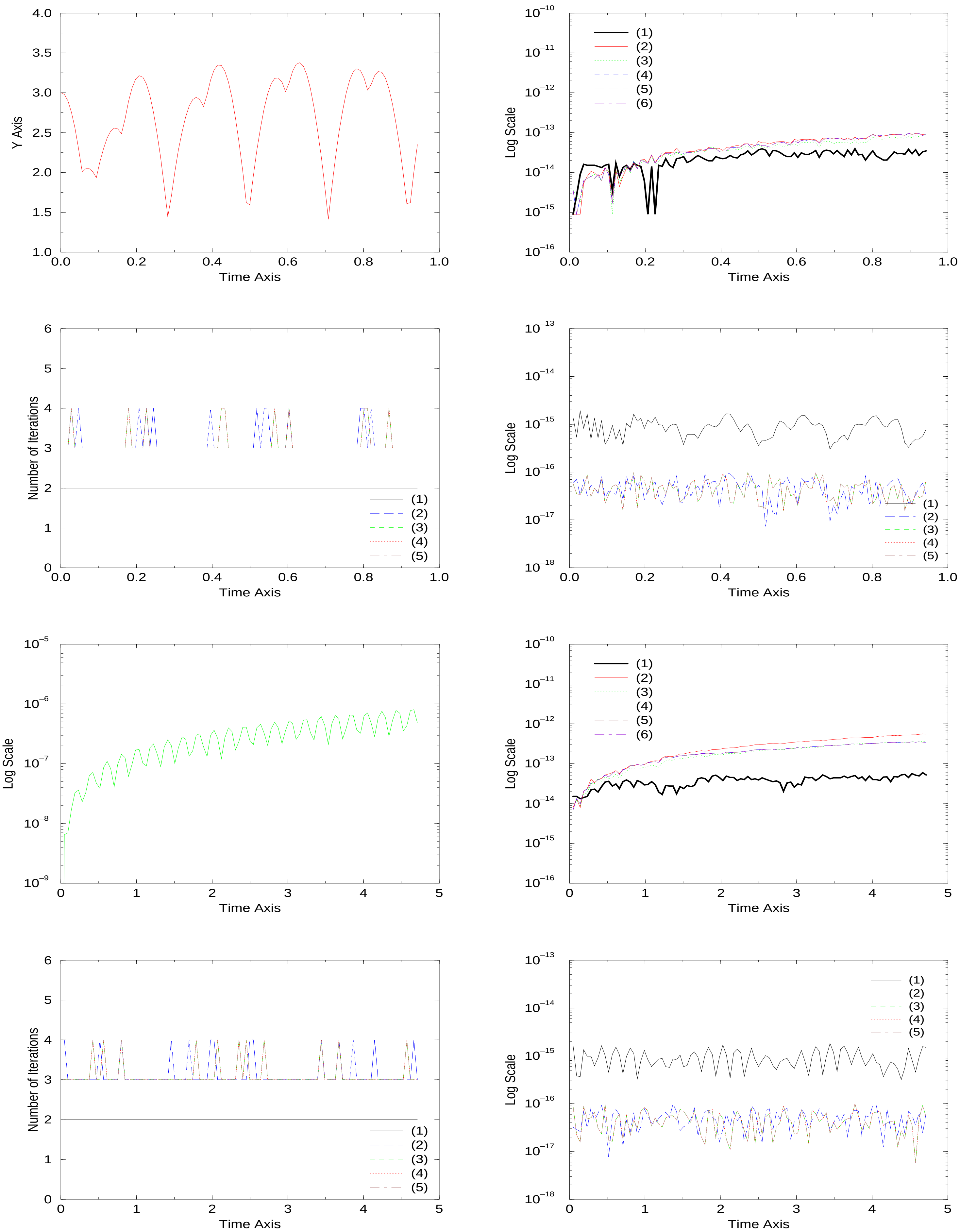
et nous résolvons alors à chacun des trois sous-pas de temps un système linéaire à l'aide d'un algorithme de point-fixe sous-relaxé ($\omega \approx 0,25$).

TAB. 3.22 – *Normes et erreurs pour (CDIRK4) classique.*

$(t = 2T)$	$\lvert U_i \rvert_2$	$\lvert U_i \rvert_\infty$	$\lvert U_i - U_i^{\mathrm{ex}} \rvert_2$	$\lvert U_i - U_i^{\mathrm{ex}} \rvert_\infty$	B.C.
U_1	0.1567E+01	0.2120E+01	0.9787E-07	0.1495E-06	0.1439E-14
U_2	0.1595E+01	0.2226E+01	0.9519E-07	0.1468E-06	0.5541E-14
U_3	0.1414E+01	0.2350E+01	0.2136E-06	0.2756E-06	
$(t = 10T)$					
U_1	0.1578E+01	0.1892E+01	0.4790E-06	0.7591E-06	0.1740E-14
U_2	0.1584E+01	0.2005E+01	0.4758E-06	0.7562E-06	0.5792E-14
U_3	0.1414E+01	0.2412E+01	0.1073E-05	0.1260E-05	

Le critère de convergence de l'algorithme ainsi que les valeurs du paramètre *maxiter* introduit pour le schéma (CN2) sont les mêmes.

FIG. 3.9 – *Test 1 pour (CDIRK4) :* $[0, 2T]$ *en haut et* $[0, 10T]$ *en bas.*



Nous prenons aussi $N_1 = M_1 = 4$ et le pas de temps $\Delta t = 10^{-3}$ pour satisfaire le critère de convergence du point fixe.
Les valeurs finales des normes L^2 et L^∞ des composantes U_1, U_2 et U_3 et de leurs erreurs absolues sont présentées dans les tableaux (3.22) et (3.23). Les valeurs finales sont du même ordre (quatre décimales identiques) et les valeurs obtenues pour les conditions aux limites sont proche de 10^{-15} soit la précision machine, figure (3.10).
L'algorithme du point fixe nécessite entre 3 et 4 itérations pour converger numériquement. Seul le cas $maxiter = 2$ est légèrement moins bon pour la conservation de l'invariant, ce que l'on retrouve aussi dans le résidu du point fixe, figure (3.9).

TAB. 3.23 – *Normes et erreurs pour (CDIRK4) maxiter = 2.*

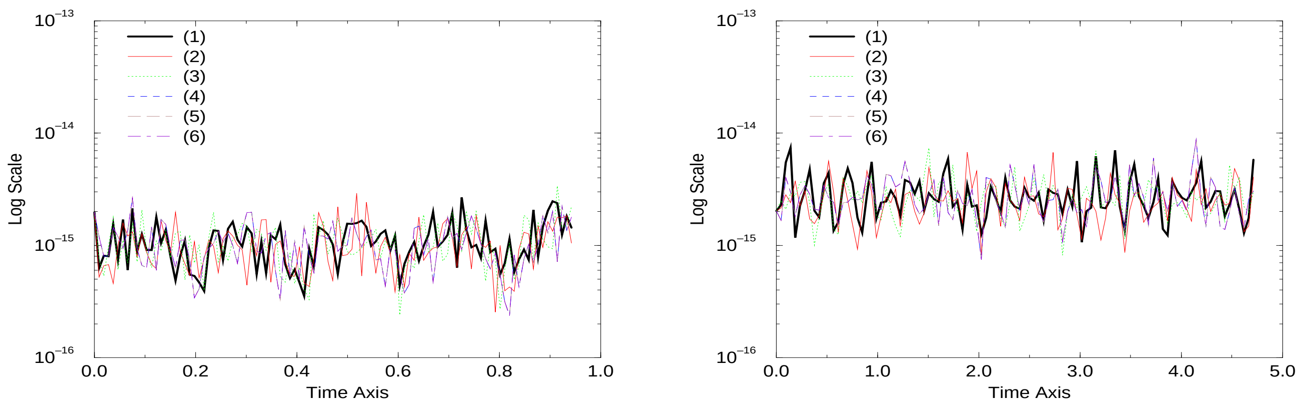
$(t = 2T)$	$\|U_i\|_2$	$\|U_i\|_\infty$	$\|U_i - U_i^{ex}\|_2$	$\|U_i - U_i^{ex}\|_\infty$	B.C.
U_1	0.1567E+01	0.2120E+01	0.9787E-07	0.1495E-06	0.1050E-14
U_2	0.1595E+01	0.2226E+01	0.9519E-07	0.1468E-06	0.2990E-14
U_3	0.1414E+01	0.2350E+01	0.2136E-06	0.2756E-06	
$(t = 10T)$					
U_1	0.1578E+01	0.1892E+01	0.4790E-06	0.7591E-06	0.1288E-14
U_2	0.1584E+01	0.2005E+01	0.4758E-06	0.7562E-06	0.4059E-14
U_3	0.1414E+01	0.2412E+01	0.1073E-05	0.1260E-05	

Les paramètres relatifs au Bi-CG Stab sont aussi corrects que pour le schéma (CN2).

TAB. 3.24 – *Statistiques sur le nombre d'itérations du Bi-CG Stab pour (CDIRK4).*

(CDIRK4)	Nombre d'itérations				(CDIRK4)	Nombre d'itérations			
$(t = 2T)$	Min	Max	Moy.	Ec-type	$(t = 10T)$	Min	Max	Moy.	Ec-type
cl	2	2	2	0	cl	2	2	2	0
(CDIRK4)	Résidu du système linéaire				(CDIRK4)	Résidu du système linéaire			
$(t = 2T)$	Min	Max	Moy.	Ec-type	$(t = 10T)$	Min	Max	Moy.	Ec-type
cl	5E-19	2E-18	1E-18	5E-19	cl	4E-19	2E-18	1E-18	5E-19

FIG. 3.10 – $U_1(x, y = \pm 1)$ *et* $U_2(x = \pm 1, y)$ *pour (CDIRK4).*



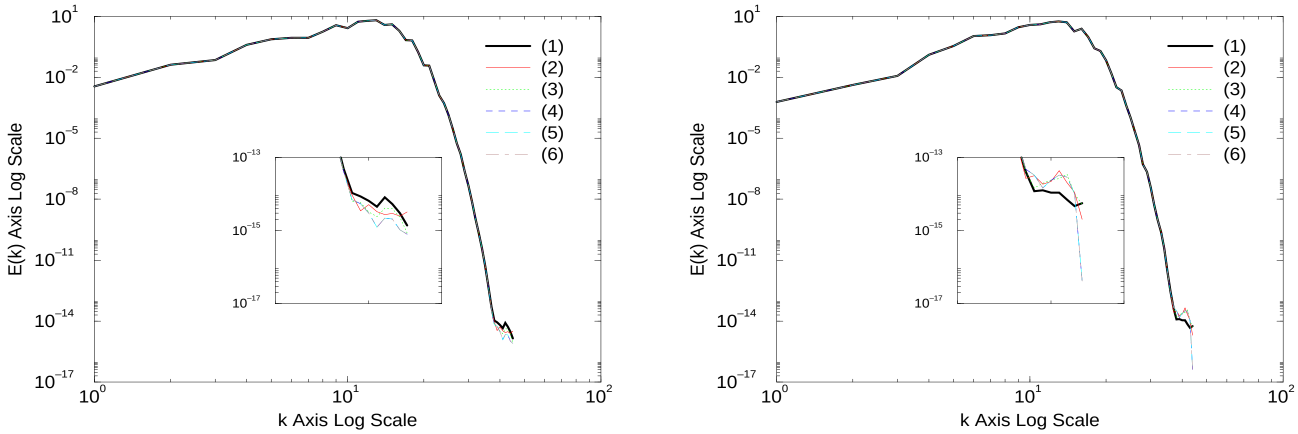
Pour le temps calcul des différentes simulations, tableau (3.25), la méthode à deux niveaux est au mieux un peu moins performante que la méthode classique. Cela peut s'expliquer par le fait que les itérations du point fixe ne compensent pas assez l'inversion en deux itérations du système linéaire issu de la méthode classique.

TAB. 3.25 – *Temps calcul pour (CDIRK4) classique et (CDIRK4) à 2 niveaux.*

(CDIRK4)	cl	2	4	6	8	10
$(t = 2T)$	86 s	86 s	148 s	155 s	155 s	155 s
$(t = 10T)$	402 s	414 s	734 s	781 s	781 s	781 s

Enfin nous traçons les spectres de la composante U_2 et U_3 resp. aux instants $t = 2T$ et $t = 10T$ avec un zoom de la queue du spectre. Les différentes courbes - (1) correspond à la méthode classique et (2), (3), (4), (5) et (6) - correspondent à la méthode à 2 niveaux pour différentes valeurs de *maxiter* - se superposent convenablement et seuls les derniers modes présentent des différences visibles. Il est difficile de préciser convenablement l'origine du phénomène. Nous sommes proches de la précision machine et de l'ordre de grandeur du critère d'arrêt du point fixe.

FIG. 3.11 – *Spectre de Legendre pour (CDIRK4) à $t = 2T$ et $t = 10T$.*



3.6.2 Test 2.

Nous augmentons les valeurs des composantes du vecteur d'onde, $\mathbf{k} = (k_1, k_2) = (12, 12)$. Les scalaires λ_{3i} gardent leur valeur : 1, 2, -1, 1.
On en déduit la période de référence T construite à l'aide de ω_1 : $T = T_1 = \frac{\sqrt{2}}{12} \approx 0,11$. Nous prenons alors $N = M = 64$ modes respectivement dans les direction x et y. Les paramètres N_1 et M_1 sont choisis égaux à 4.

3.6.2.1 Méthode explicite (CRK4) :

La relation $\Delta t NM \approx \mathcal{O}(1)$ nous fournit un pas de temps Δt de l'ordre de 2.10^{-4}.
Les valeurs finales à $t = 2T$ et $t = 10T$ des normes et erreurs sont très proches pour les deux versions de cette méthode explicite, tableaux (3.26), (3.27). Il y a là encore pas ou peu de différence entre la version classique du schéma (CRK4) et sa version à deux niveaux, comme l'illustrent les tableaux (3.26) et (3.27).

Nous notons la bonne conservation de l'invariant (2.10^{-14}) pour les deux versions.

TAB. 3.26 – *Normes et erreurs pour le schéma (CRK4) classique.*

$(t = 2T)$	$\|U_i\|_2$	$\|U_i\|_\infty$	$\|U_i - U_i^{\text{ex}}\|_2$	$\|U_i - U_i^{\text{ex}}\|_\infty$	B.C.
U_1	0.1232E+01	0.1919E+01	0.5457E-08	0.8304E-08	0.2159E-14
U_2	0.1218E+00	0.1888E+01	0.5492E-08	0.9666E-08	0.9476E-14
U_3	0.2000E+01	0.2959E+01	0.6247E-08	0.1036E-07	
$(t = 10T)$					
U_1	0.1217E+01	0.1902E+01	0.2727E-07	0.4021E-07	0.1621E-14
U_2	0.1232E+01	0.1932E+01	0.2707E-07	0.3962E-07	0.3520E-14
U_3	0.2000E+01	0.2947E+01	0.3102E-07	0.5036E-07	

TAB. 3.27 – *Normes et erreurs pour le schéma (CRK4) à 2 niveaux.*

$(t = 2T)$	$\|U_i\|_2$	$\|U_i\|_\infty$	$\|U_i - U_i^{\text{ex}}\|_2$	$\|U_i - U_i^{\text{ex}}\|_\infty$	B.C.
U_1	0.1232E+01	0.1919E+01	0.5457E-08	0.8304E-08	0.2045E-14
U_2	0.1218E+00	0.1888E+01	0.5492E-08	0.9666E-08	0.5364E-14
U_3	0.2000E+01	0.2959E+01	0.6247E-08	0.1036E-07	
$(t = 10T)$					
U_1	0.1217E+01	0.1902E+01	0.2727E-07	0.4021E-07	0.2333E-14
U_2	0.1232E+01	0.1932E+01	0.2707E-07	0.3962E-07	0.7222E-14
U_3	0.2000E+01	0.2947E+01	0.3102E-07	0.5036E-07	

Les temps calcul sont présentés dans le tableau (3.28).
La différence provient du morcellement des différents modules (produits matrice/vecteur, calculs du second membre) en parties plus petites, ce qui entraine un nombre d'appels à des subroutines plus important.

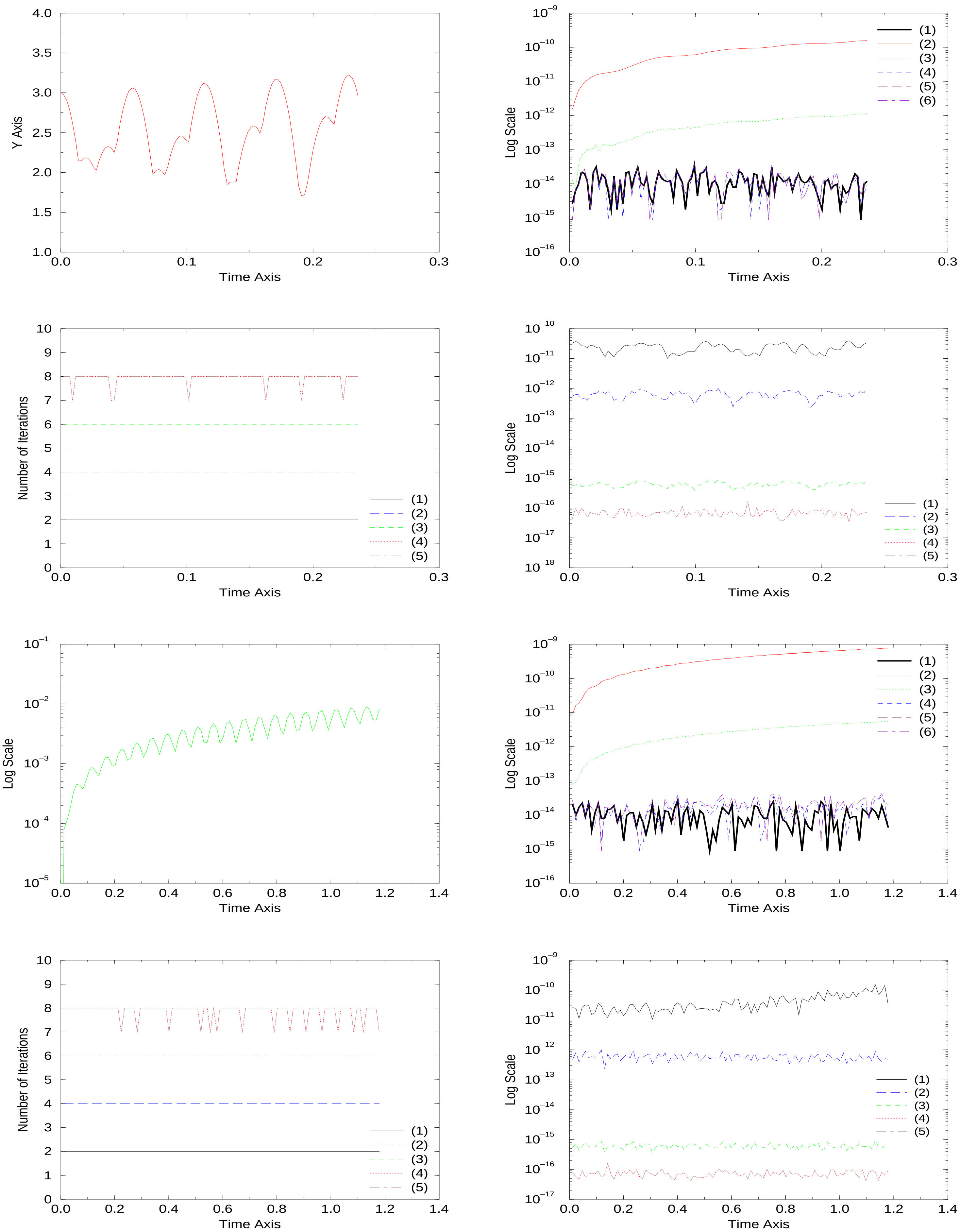
TAB. 3.28 – *Temps calcul pour (CRK4) classique et (CRK4) à 2 niveaux.*

	(CRK4) classique	(CRK4) 2 niveaux
$(t = 2T)$	135 s	192 s
$(t = 10T)$	627 s	912 s

3.6.2.2 Méthode semi-implicite (CN2).

Nous considérons le même test d'arrêt pour l'algorithme du point fixe sous-relaxé ($\omega = 0,949$) que précédemment. La dimension du problème à résoudre est quatre fois plus grande que pour le test 1. Le conditionnement de la matrice du système linéaire associé à la méthode classique est alors plus grand, i.e. moins bon. Le pas de temps a été déterminé de telle sorte que l'algorithme itératif soit convergent, soit $\Delta t \approx 6.10^{-4}$, d'après les résultats du tableau (3.10).
L'algorithme du point fixe nécessite entre 7 et 8 itérations pour satisfaire pleinement le critère de convergence. Un nombre maximal d'itérations inférieur à ceux-ci mène à une convergence médiocre (surtout pour $maxiter = 2$ ou 4) et la conservation de l'invariant en patît le plus. Cela est valable aussi bien sur $[0, 2T]$ que sur $[0, 10T]$, figure (3.12).

FIG. 3.12 – *Test 2 pour (CN2) :* $[0, 2T]$ *en haut et* $[0, 10T]$ *en bas.*



TAB. 3.29 – *Normes et erreurs pour (CN2) classique.*

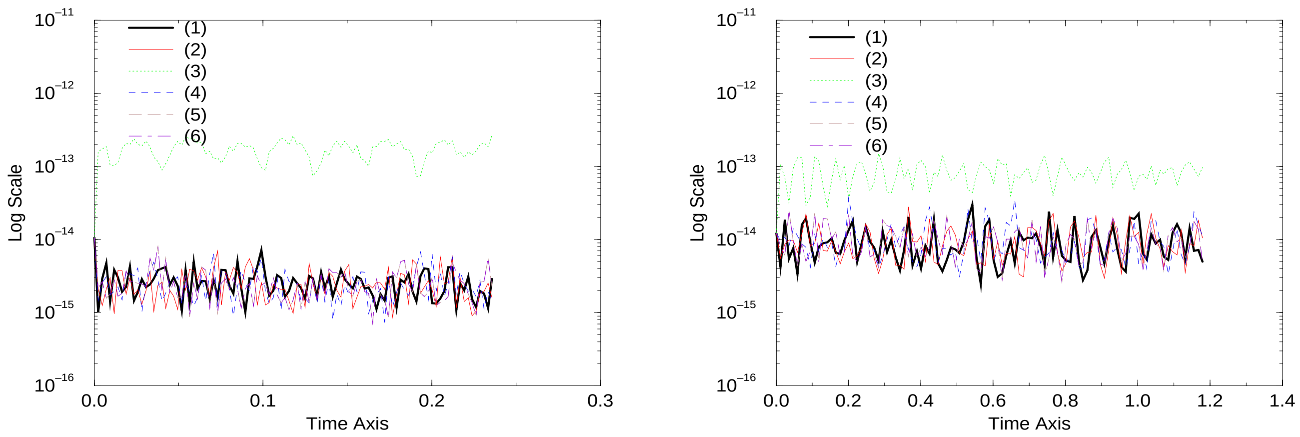
$(t = 2T)$	$\|U_i\|_2$	$\|U_i\|_\infty$	$\|U_i - U_i^{\mathrm{ex}}\|_2$	$\|U_i - U_i^{\mathrm{ex}}\|_\infty$	B.C.
U_1	0.1231E+01	0.1918E+01	0.1622E-02	0.2361E-02	0.2901E-14
U_2	0.1217E+01	0.1888E+01	0.1635E-02	0.2383E-02	0.8836E-14
U_3	0.2001E+01	0.2960E+01	0.1908E-02	0.2975E-02	
$(t = 10T)$					
U_1	0.1222E+01	0.1905E+01	0.8164E-02	0.1185E-01	0.4156E-14
U_2	0.1236E+01	0.1935E+01	0.8098E-02	0.1174E-01	0.5000E-14
U_3	0.1995E+01	0.2940E+01	0.9560E-02	0.1500E-02	

TAB. 3.30 – *Normes et erreurs pour (CN2) maxiter = 2.*

$(t = 2T)$	$\|U_i\|_2$	$\|U_i\|_\infty$	$\|U_i - U_i^{\mathrm{ex}}\|_2$	$\|U_i - U_i^{\mathrm{ex}}\|_\infty$	B.C.
U_1	0.1231E+01	0.1918E+01	0.1622E-02	0.2361E-02	0.1621E-14
U_2	0.1217E+01	0.1888E+01	0.1635E-02	0.2383E-02	0.7740E-14
U_3	0.2001E+01	0.2960E+01	0.1908E-02	0.2975E-02	
$(t = 10T)$					
U_1	0.1222E+01	0.1905E+01	0.8164E-02	0.1186E-01	0.5440E-14
U_2	0.1236E+01	0.1935E+01	0.8098E-02	0.1174E-01	0.1049E-13
U_3	0.1995E+01	0.2940E+01	0.9560E-02	0.1500E-02	

L'évolution des valeurs de la solution au bord montre la bonne imposition de celles-ci, figure (3.13). Pour une raison que nous n'arrivons pas à déterminer précisément le cas *maxiter = 4* est qualitativement moins bon que *maxiter = 2*. bien que les valeurs obtenues soient en 10^{-13} au lieu de 10^{-14} pour toutes les autres, elles demeurent acceptables pour des conditions aux limites de type Dirichlet homogène.

FIG. 3.13 – $U_1(x, y = \pm 1)$ *et* $U_2(x = \pm 1, y)$ *pour (CN2).*



Le Bi-CG Stab converge en trois itérations :

TAB. 3.31 – *Statistiques sur le nombre d'itérations du Bi-CG Stab pour (CN2).*

(CN2)	Nombre d'itérations				(CN2)	Nombre d'itérations			
$(t = 2T)$	Min	Max	Moy.	Ec-type	$(t = 10T)$	Min	Max	Moy.	Ec-type
cl	3	3	3	0	cl	3	3	3	0
(CN2)	Résidu du système linéaire				(CN2)	Résidu du système linéaire			
$(t = 2T)$	Min	Max	Moy.	Ec-type	$(t = 10T)$	Min	Max	Moy.	Ec-type
cl	4E-19	7E-18	2E-18	1E-18	cl	8E-19	7E-18	2E-18	9E-19

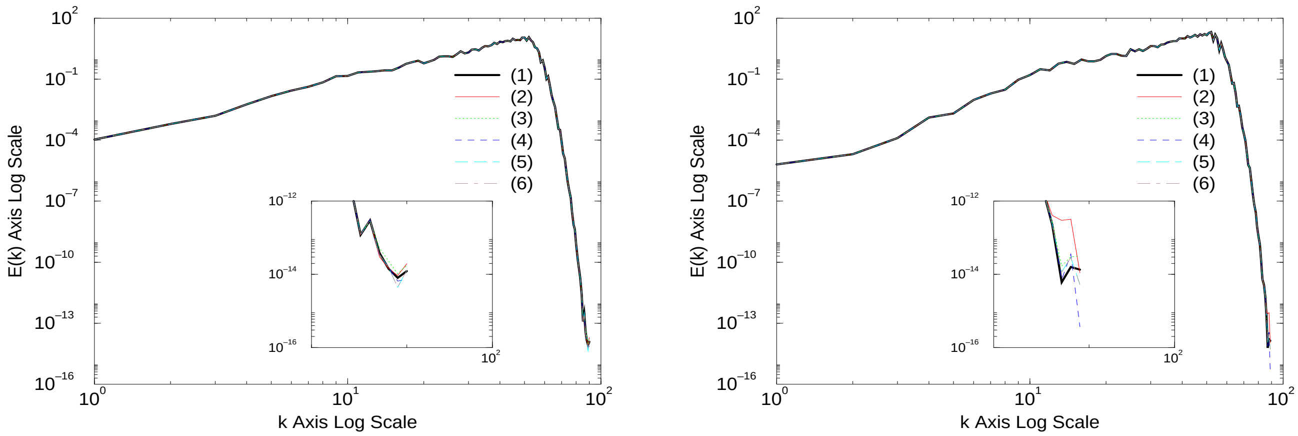
Les temps calcul font apparaitre un facteur 2 entre les deux versions mais les résultats pour $maxiter = 6$, encore acceptables, n'engendrent qu'un facteur 5/3.

TAB. 3.32 – *Temps calcul pour (CN2) classique et (CN2) à 2 niveaux.*

	cl	2	4	6	8	10
$(t = 2T)$	114 s	79 s	134 s	190 s	240 s	240 s
$(t = 10T)$	518 s	335 s	612 s	890 s	1143 s	1143 s

L'allure des spectres de la composante U_1 de la solution, figure (3.14), aux instants $t = 2T$ et $t = 10T$ est semblable pour les deux méthodes, seul $maxiter = 2$ se distingue des autres.

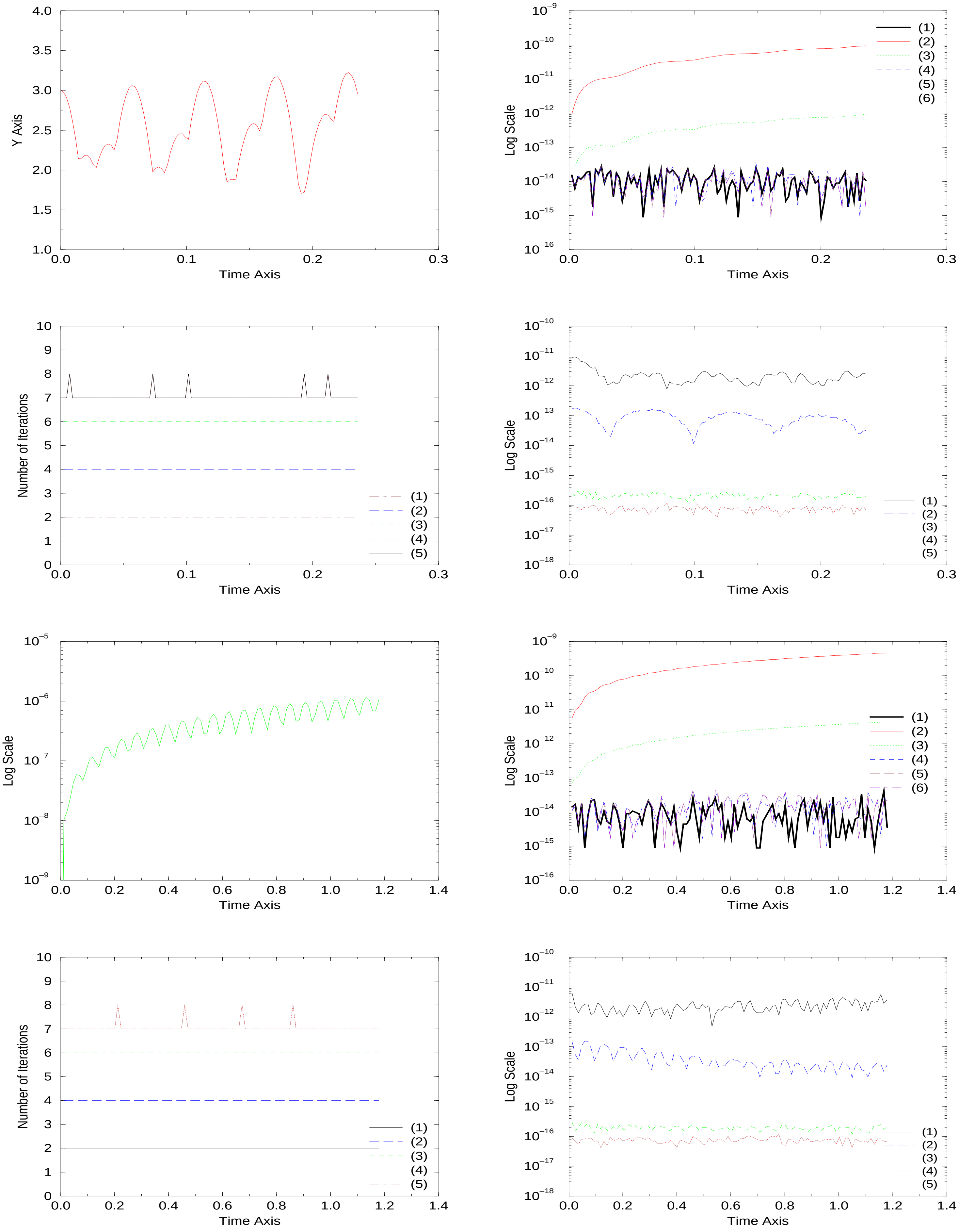
FIG. 3.14 – *Spectre de Legendre pour (CN2) à $t = 2T$ et $t = 10T$.*



3.6.2.3 Méthode semi-implicite (CDIRK4).

De même que pour (CN2), le pas de temps Δt est déterminé de telle sorte que l'algorithme itératif soit convergent. Nous obtenons numériquement (voir le tableau (3.11)) $\Delta t \approx 3.10^{-4}$. Le facteur de sous-relaxation est $\omega = 0,9445$.
Ce schéma d'ordre 4 donne naturellement des erreurs globales plus petites que le schéma précédent, d'ordre 2, ce qu'illustrent les tableaux (3.33) et (3.34).

FIG. 3.15 – *Test 2 pour (CDIRK4) :* $[0, 2T]$ *en haut et* $[0, 10T]$ *en bas.*



TAB. 3.33 – *Normes et erreurs pour (CDIRK4) classique.*

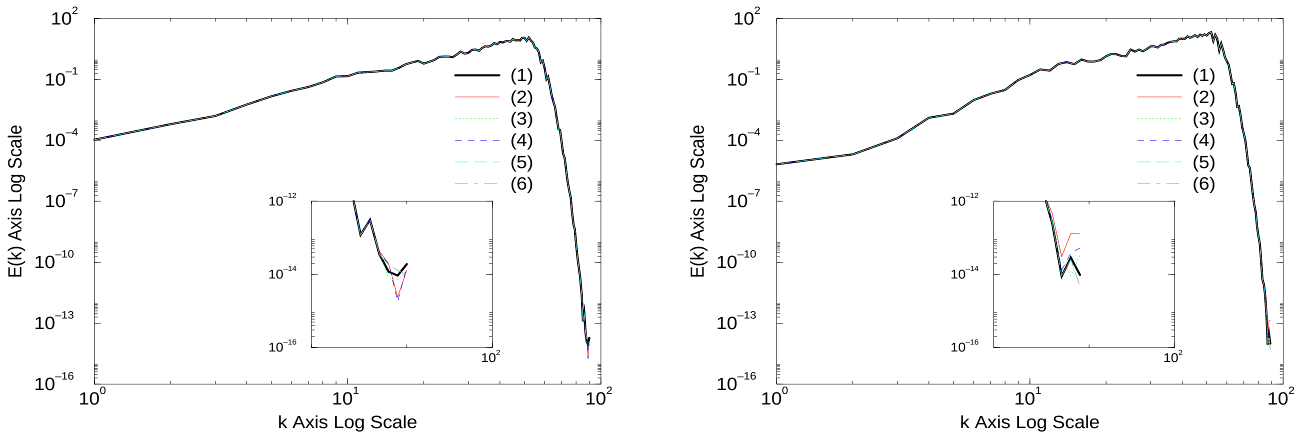
$(t = 2T)$	$\lvert U_i \rvert_2$	$\lvert U_i \rvert_\infty$	$\lvert U_i - U_i^{\text{ex}} \rvert_2$	$\lvert U_i - U_i^{\text{ex}} \rvert_\infty$	B.C.
U_1	0.1232E+01	0.1919E+01	0.2118E-06	0.3064E-06	0.2755E-14
U_2	0.1218E+01	0.1888E+01	0.2134E-06	0.3103E-06	0.6120E-14
U_3	0.2000E+01	0.2959E+01	0.2430E-06	0.3805E-06	
$(t = 10T)$					
U_1	0.1217E+01	0.1902E+01	0.1069E-05	0.1540E-05	0.4487E-14
U_2	0.1232E+01	0.1932E+01	0.1061E-05	0.1526E-05	0.8756E-14
U_3	0.2000E+01	0.2947E+01	0.1212E-05	0.1895E-05	

TAB. 3.34 – *Normes et erreurs pour (CDIRK4) maxiter = 2.*

$(t = 2T)$	$\lvert U_i \rvert_2$	$\lvert U_i \rvert_\infty$	$\lvert U_i - U_i^{\text{ex}} \rvert_2$	$\lvert U_i - U_i^{\text{ex}} \rvert_\infty$	B.C.
U_1	0.1232E+01	0.1919E+01	0.2118E-06	0.3063E-06	0.3295E-14
U_2	0.1218E+01	0.1888E+01	0.2134E-06	0.3103E-06	0.6000E-14
U_3	0.2000E+01	0.2959E+01	0.2430E-06	0.3806E-06	
$(t = 10T)$					
U_1	0.1217E+01	0.1902E+01	0.1069E-05	0.1540E-05	0.2506E-14
U_2	0.1232E+01	0.1932E+01	0.1061E-05	0.1526E-05	0.6794E-14
U_3	0.2000E+01	0.2947E+01	0.1212E-05	0.1896E-05	

Les conditions aux limites sont convenablement imposées comme le montre la figure (3.17). L'algorithme du point fixe nécessite entre 7 et 8 itérations pour converger numériquement. On remarque que parmi les valeurs inférieures de *maxiter*, la valeur 6 fournit des résultats qualitativement équivalents à *maxiter* = 8, 10 ou la version classique, figure (3.15). Il en est de même pour les spectres de U_2 et U_3 resp. à $t = 2T$ et $t = 10T$, (3.16).

FIG. 3.16 – *Spectre de Legendre pour (CDIRK4) à $t = 2T$ et $t = 10T$.*



Les statistiques du Bi-CG Stab, tableau (3.35), montrent la nécessité de faire 2 itérations pour la méthode classique, tout comme le schéma (CN2) précédemment.

TAB. 3.35 – *Statistiques sur le nombre d'itérations du Bi-CG Stab pour (CDIRK4).*

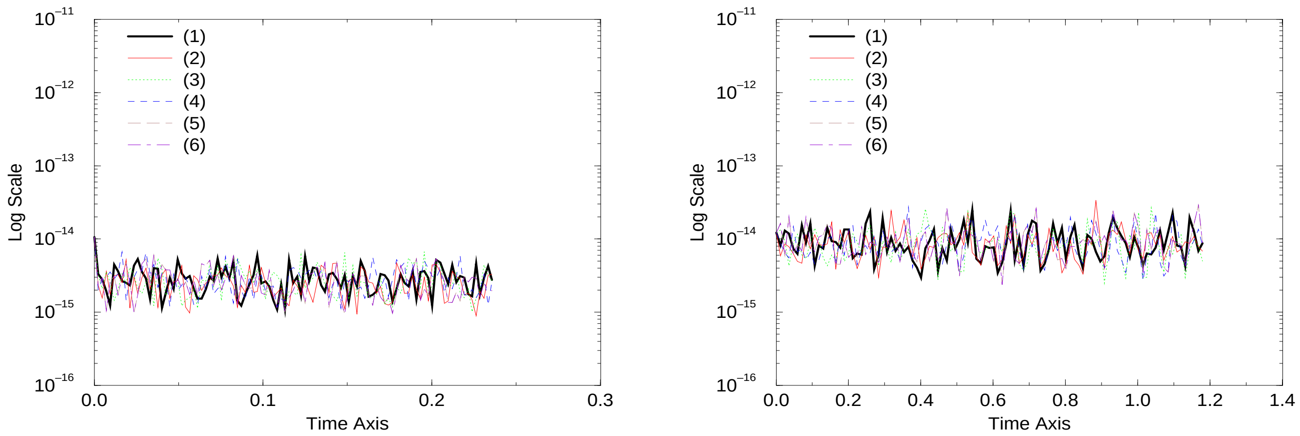
(CDIRK4)	Nombre d'itérations				(CDIRK4)	Nombre d'itérations			
$(t = 2T)$	Min	Max	Moy.	Ec-type	$(t = 10T)$	Min	Max	Moy.	Ec-type
cl	2	2	2	0	cl	2	2	2	0
(CDIRK4)	Résidu du système linéaire				(CDIRK4)	Résidu du système linéaire			
$(t = 2T)$	Min	Max	Moy.	Ec-type	$(t = 10T)$	Min	Max	Moy.	Ec-type
cl	4E-16	7E-16	7E-17	7E-17	cl	2E-17	7E-16	3E-17	2E-16

Les temps calcul entre les différentes simulations montrent qu'à même Δt, la méthode à deux niveaux demande entre 0,75 et 3 fois le temps calcul de la méthode classique.

TAB. 3.36 – *Temps calcul pour (CDIRK4) classique et (CDIRK4) à 2 niveaux.*

	cl	2	4	6	8	10
$(t = 2T)$	465 s	358 s	684 s	1010 s	1260 s	1260 s
$(t = 10T)$	2054 s	1728 s	3359 s	4990 s	6213 s	6133 s

FIG. 3.17 – $U_1(x, y = \pm 1)$ *et* $U_2(x = \pm 1, y)$ *pour (CDIRK4).*



TAB. 3.37 – *Temps calcul pour (CRK4) classique et (CRK4) à 2 niveaux.*

	(CRK4) classique	(CRK4) 2 niveaux
$(t = 2T)$	1262 s	1974 s
$(t = 10T)$	5929 s	9756 s

3.6.3 Test 3.

Pour ce troisième et dernier test nous fixons $\mathbf{k} = (k_1, k_2) = (30, 30)$ et nous conservons les valeurs des scalaires $\lambda_3 = 1, \lambda_6 = 2, \lambda_9 = -1, \lambda_{12} = 1$. Cela entraine $T = \frac{\sqrt{2}}{30} \approx 0,0471$. Nous prenons 128 modes dans chaque direction spatiale et le pas de temps Δt en conséquence, ainsi que $N_1 = M_1 = 4$ comme pour les deux tests précédents.

3.6.3.1 Méthode explicite (CRK4).

Le pas de temps est alors $\Delta t \approx 2.10^{-4}$.
Nous notons la bonne conservation de l'invariant (3.10^{-14}) et des conditions limites de Dirichlet (7.10^{-15} et 3.10^{-14}).

TAB. 3.38 – *Normes et erreurs pour le schéma (CRK4) classique.*

$(t = 2T)$	$\|U_i\|_2$	$\|U_i\|_\infty$	$\|U_i - U_i^{\text{ex}}\|_2$	$\|U_i - U_i^{\text{ex}}\|_\infty$	B.C.
U_1	0.1046E+01	0.1628E+01	0.1680E-08	0.6146E-08	0.5596E-14
U_2	0.1038E+01	0.1621E+01	0.1433E-08	0.3047E-08	0.1222E-13
U_3	0.2197E+01	0.3056E+01	0.2582E-08	0.1952E-07	
$(t = 10T)$					
U_1	0.1582E+01	0.2182E+01	0.3434E-08	0.8775E-08	0.7218E-14
U_2	0.1580E+01	0.2173E+01	0.3393E-08	0.8250E-08	0.4257E-14
U_3	0.1414E+01	0.2367E+01	0.6367E-08	0.1232E-07	

Comme précédemment, les valeurs obtenues pour les deux versions sont très proches, tableaux (3.38) et (3.39).

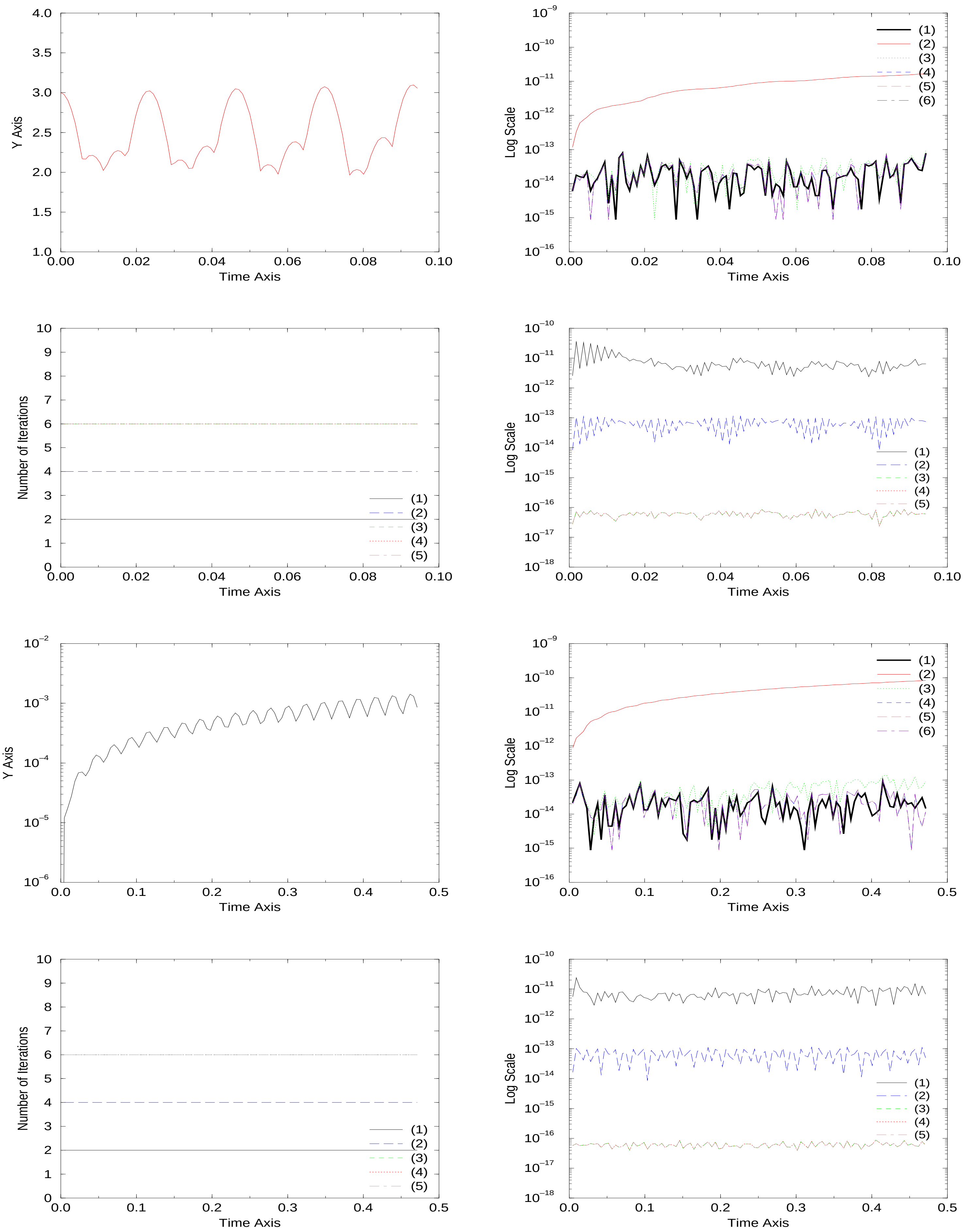
TAB. 3.39 – *Normes et erreurs pour le schéma (CRK4) à 2 niveaux.*

$(t = 2T)$	$\|U_i\|_2$	$\|U_i\|_\infty$	$\|U_i - U_i^{\text{ex}}\|_2$	$\|U_i - U_i^{\text{ex}}\|_\infty$	B.C.
U_1	0.1046E+01	0.1628E+01	0.1680E-08	0.6147E-08	0.8331E-14
U_2	0.1038E+01	0.1621E+01	0.1433E-08	0.3047E-08	0.8937E-14
U_3	0.2197E+01	0.3056E+01	0.2582E-08	0.1952E-08	
$(t = 10T)$					
U_1	0.1582E+01	0.2182E+01	0.3434E-08	0.8775E-08	0.1550E-13
U_2	0.1580E+01	0.2173E+01	0.3393E-08	0.8250E-08	0.4598E-13
U_3	0.1414E+01	0.2367E+01	0.6367E-08	0.1232E-07	

3.6.3.2 Méthode semi-implicite (CN2).

Nous prenons $\Delta t \approx 10^{-4} < 2.10^{-4}$ pour avoir la convergence théorique de l'algorithme itératif, voir le tableau (3.10). Les normes et erreurs globales de discrétisation sont quasi-identiques pour les deux méthodes considérées, tableaux (3.40) et (3.41).

FIG. 3.18 – *Test 3 pour (CN2) :* $[0, 2T]$ *en haut et* $[0, 10T]$ *en bas.*



TAB. 3.40 – *Normes et erreurs pour (CN2) classique.*

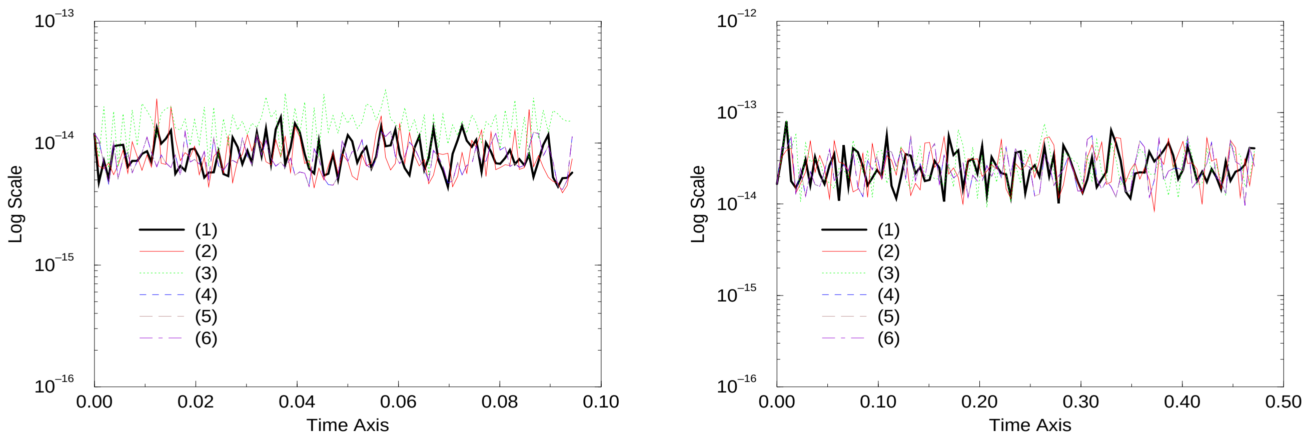
$(t=2T)$	$\|U_i\|_2$	$\|U_i\|_\infty$	$\|U_i-U_i^{ex}\|_2$	$\|U_i-U_i^{ex}\|_\infty$	B.C.
U_1	0.1046E+01	0.1628E+01	0.2682E-03	0.3693E-03	0.5745E-14
U_2	0.1038E+01	0.1621E+01	0.2691E-03	0.3703E-03	0.1935E-13
U_3	0.2197E+01	0.3056E+01	0.2476E-03	0.3884E-03	
$(t=10T)$					
U_1	0.1582E+01	0.2182E+01	0.8570E-03	0.1425E-02	0.1015E-13
U_2	0.1579E+01	0.2173E+01	0.8604E-03	0.1427E-02	0.4067E-13
U_3	0.1415E+01	0.2369E+01	0.1915E-02	0.2640E-02	

TAB. 3.41 – *Normes et erreurs pour (CN2) maxiter = 2.*

$(t=2T)$	$\|U_i\|_2$	$\|U_i\|_\infty$	$\|U_i-U_i^{ex}\|_2$	$\|U_i-U_i^{ex}\|_\infty$	B.C.
U_1	0.1046E+01	0.1628E+01	0.2682E-03	0.3693E-03	0.7414E-14
U_2	0.1038E+01	0.1631E+01	0.2691E-03	0.3703E-03	0.4546E-13
U_3	0.2197E+01	0.3056E+01	0.2476E-03	0.3884E-03	
$(t=10T)$					
U_1	0.1582E+01	0.2182E+01	0.8570E-03	0.1425E-02	0.1072E-13
U_2	0.1579E+01	0.2173E+01	0.8604E-03	0.1427E-02	0.3534E-13
U_3	0.1415E+01	0.2369E+01	0.1915E-02	0.2640E-02	

Les conditions aux limites sont bien imposées avec des valeurs majorées par 10^{-13}, figure (3.19).

FIG. 3.19 – $U_1(x, y=\pm 1)$ *et* $U_2(x=\pm 1, y)$ *pour (CN2).*



Pour la conservation de l'invariant, seul le cas *maxiter* $= 2$ se distingue des autres tracés, figure (3.18). La convergence numérique du point fixe est acquise généralement avec 6 itérations. Le facteur de sous-relaxation pour (CN2) est $\omega = 0,949$.

Les statistiques pour le Bi CG Stab sont les suivantes :

TAB. 3.42 – *Statistiques sur le nombre d'itérations du Bi-CG Stab pour (CN2).*

(CN2)	Nombre d'itérations				(CN2)	Nombre d'itérations			
($t = 2T$)	Min	Max	Moy.	Ec-type	($t = 10T$)	Min	Max	Moy.	Ec-type
cl	2	2	2	0	cl	2	2	2	0
(CN2)	Résidu du système linéaire				(CN2)	Résidu du système linéaire			
($t = 2T$)	Min	Max	Moy.	Ec-type	($t = 10T$)	Min	Max	Moy.	Ec-type
cl	2E-17	5E-17	4E-17	6E-18	cl	2E-17	5E-17	4E-17	6E-18

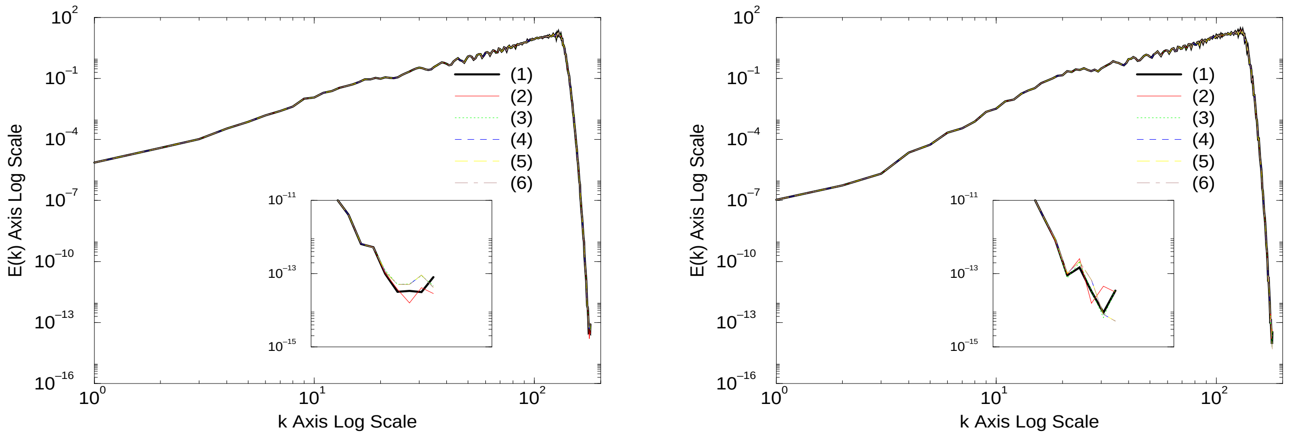
Les temps calcul de la méthode à deux niveaux représentent de 0,7 à 2 fois celui de la méthode classique.

TAB. 3.43 – *Temps calcul pour (CN2) classique et (CN2) à 2 niveaux.*

	cl	2	4	6	8	10
($t = 2T$)	1460 s	1067 s	1906 s	2741 s	2741 s	2741 s
($t = 10T$)	7000 s	4996 s	9189 s	13362 s	13213 s	13252 s

Les spectres des composantes U_2 et U_3, figure (3.20), ont la même allure, seules les valeurs des derniers modes permettent de les distinguer.

FIG. 3.20 – *Spectre de Legendre pour (CN2) à $t = 2T$ et $t = 10T$.*



3.6.3.3 Méthode semi-implicite (CDIRK4).

Le critère de convergence du point fixe sur-relaxé impose un pas de temps très petit pour ce schéma : $\Delta t \approx 8.10^{-5} < 9.10^{-5}$. Le test de convergence et les valeurs du paramètre *maxiter* sont les mêmes que précédemment.

Les résultats obtenus ici sont qualitativement similaires à ceux du schéma (CN2) avec une meilleure précision pour les normes globales, tableaux (3.44) et (3.45).

TAB. 3.44 – *Normes et erreurs pour (CDIRK4) classique.*

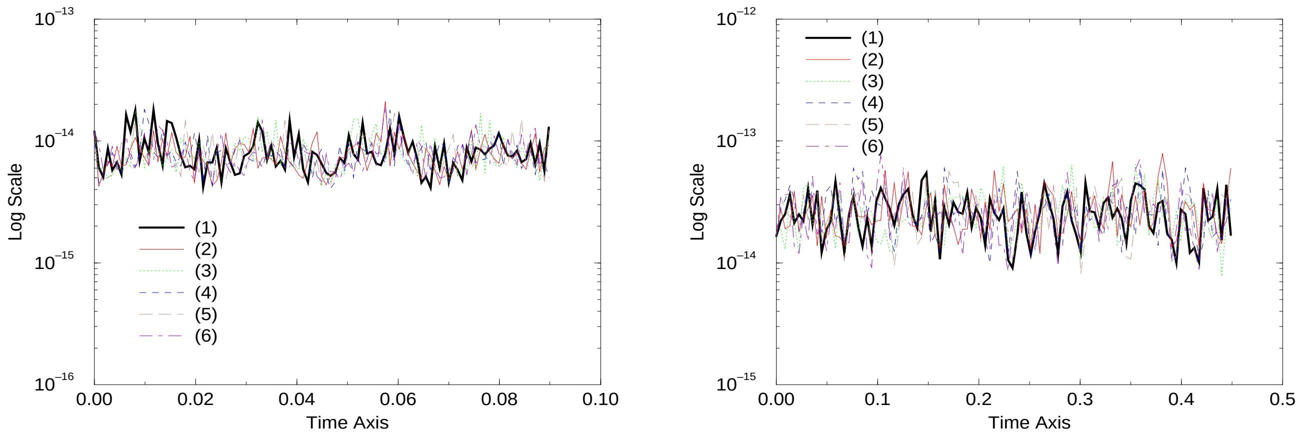
$(t = 2T)$	$\|U_i\|_2$	$\|U_i\|_\infty$	$\|U_i - U_i^{\mathrm{ex}}\|_2$	$\|U_i - U_i^{\mathrm{ex}}\|_\infty$	B.C.
U_1	0.1046E+01	0.1628E+01	0.7076E-07	0.9713E-07	0.5944E-14
U_2	0.1038E+01	0.1621E+01	0.7098E-07	0.9886E-07	0.1345E-13
U_3	0.2197E+01	0.3056E+01	0.6415E-07	0.1060E-06	
$(t = 10T)$					
U_1	0.1582E+01	0.2182E+01	0.2248E-06	0.3747E-06	0.8179E-14
U_2	0.1580E+01	0.2173E+01	0.2257E-06	0.3764E-06	0.2861E-13
U_3	0.1414E+01	0.2367E+01	0.5023E-06	0.7013E-06	

TAB. 3.45 – *Normes et erreurs pour (CDIRK4) maxiter = 2.*

$(t = 2T)$	$\|U_i\|_2$	$\|U_i\|_\infty$	$\|U_i - U_i^{\mathrm{ex}}\|_2$	$\|U_i - U_i^{\mathrm{ex}}\|_\infty$	B.C.
U_1	0.1046E+01	0.1628E+01	0.7076E-07	0.9727E-07	0.9905E-14
U_2	0.1038E+01	0.1621E+01	0.7098E-07	0.9867E-07	0.2064E-13
U_3	0.2197E+01	0.3056E+01	0.6415E-07	0.1128E-06	
$(t = 10T)$					
U_1	0.1582E+01	0.2182E+01	0.2316E-06	0.5483E-06	0.9047E-14
U_2	0.1580E+01	0.2163E+01	0.2262E-06	0.4032E-06	0.3187E-13
U_3	0.1414E+01	0.2367E+01	0.5074E-06	0.1879E-05	

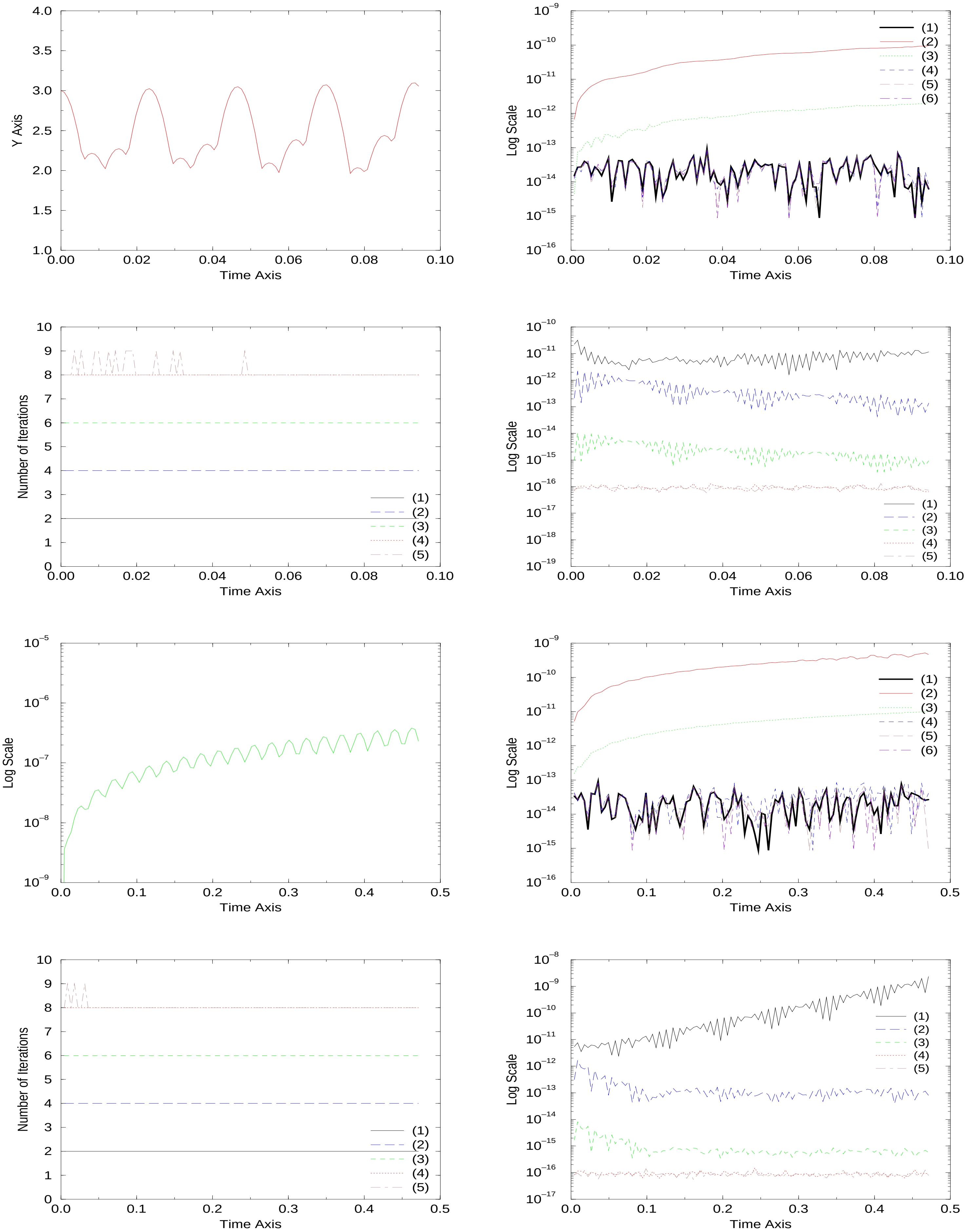
Les conditions aux limites sont imposées à 10^{-14} près aussi bien pour U_1 que pour U_2.

FIG. 3.21 – $U_1(x, y = \pm 1)$ *et* $U_2(x = \pm 1, y)$ *pour (CDIRK4).*



La conservation de l'invariant est moins bonne pour *maxiter = 2*, elle marque une nette augmentation, partant de 10^{-11} et allant jusqu'à 10^{-9} où elle semble se stabiliser, figure (3.22).

FIG. 3.22 – *Test 3 pour (CDIRK4) :* $[0, 2T]$ *en haut et* $[0, 10T]$ *en bas.*



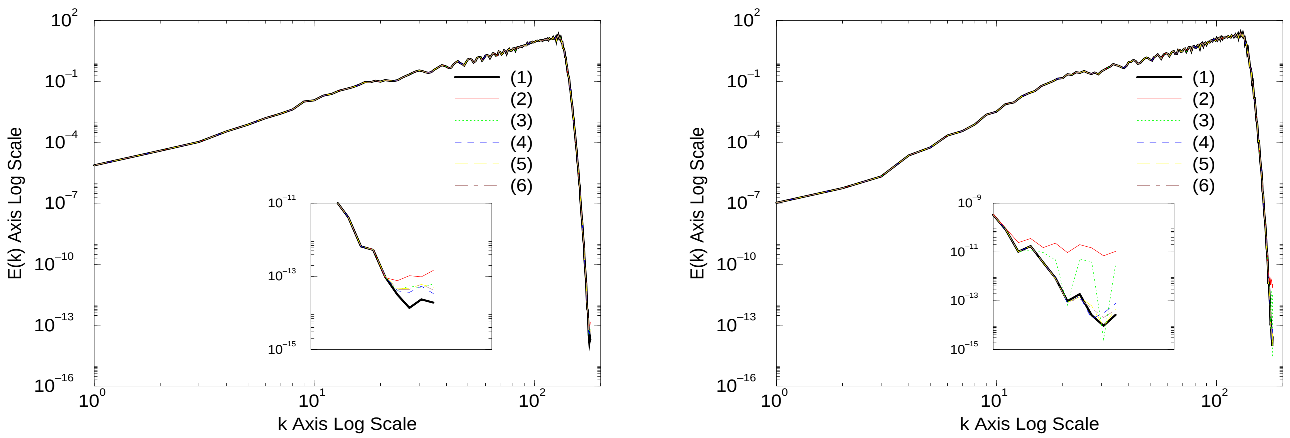
Le tracé du résidu du point fixe montre l'insuffisance de la résolution qui mène à un comportement instable. Cela se traduit par une remontée des spectres des fonctions, figure (3.23). Ce phénomène correspond parfaitement àl'insuffisance de résolution pour l'imposition des conditions aux limites.
L'algorithme du point fixe demande là aussi entre 8 et 9 itérations pour converger.
L'étude statistique du Bi-CG Stab montre la très bonne capacité de cet algorithme à résoudre notre système linéaire puisque 2 itérations suffisent pour obtenir un résidu inférieur à 10^{-15}. La même remarque est valable pour le schéma précédent, (CN2).

TAB. 3.46 – *Statistiques sur le nombre d'itérations du Bi-CG Stab pour (CDIRK4).*

	Nombre d'itérations					Nombre d'itérations			
$(t = 2T)$	Min	Max	Moy.	Ec-type	$(t = 10T)$	Min	Max	Moy.	Ec-type
cl	2	2	2	0	cl	2	2	2	0
(CDIRK4)	Résidu du système linéaire				(CDIRK4)	Résidu du système linéaire			
$(t = 2T)$	Min	Max	Moy.	Ec-type	$(t = 10T)$	Min	Max	Moy.	Ec-type
cl	5E-16	1E-15	7E-16	1E-16	cl	5E-17	1E-15	5E-16	3E-16

Les spectres de la fonction U_1 sont à $t = 2T$ et $t = 10T$:

FIG. 3.23 – *Spectre de Legendre pour (CDIRK4) à $t = 2T$ et $t = 10T$.*



Les temps calcul pour la méthode à 2 niveaux vont de 2/3 à près de 3 fois celui de la méthode classique. Le cas *maxiter = 6* qui fournit des résultats de qualité équivalente, ne nécessite environ que le double du temps calcul et cela à même Δt.

TAB. 3.47 – *Temps calcul pour (CDIRK4) classique et (CDIRK4) à 2 niveaux.*

	cl	2	4	6	8	10
$(t = 2T)$	4586 s	2756 s	5434 s	8205 s	10764 s	11524 s
$(t = 10T)$	20732 s	13776 s	27260 s	40582 s	54153 s	57738 s

3.7 Parallélisation des codes.

Nous considérons le parallélisme sur des machines de type MIMD (multiple instruction, multiple data) comme les Cray YMP et C90.
D'un point de vue algorithmique, on peut voir les trois étapes de chaque pas de temps du schéma (CDIRK4) comme trois applications du schéma (CN2). En effet, ce sont deux schémas semi-implicites dont les matrices ne différent que par la valeur du coefficient α :

$$\alpha = \begin{cases} \dfrac{1}{2} & \text{pour (CN2)} \\ \dfrac{1+\xi}{2} & \text{pour (CDIRK4)} \end{cases}$$

De plus, les opérations conduisant aux seconds membres des systèmes linéaires à résoudre sont semblables. Ainsi nous ne présentons la parallélisation des versions classique et à deux niveaux que pour le schéma (CN2).
La parallélisation sera effectuée à l'aide de deux techniques complémentaires :

– les portions de code pouvant se scinder en gros grains indépendants sont mis en évidence à l'aide de directives de compilation :

```
CMIC$ PARALLEL
CMIC$1SHARED  ( Variables accessibles par toutes les subroutines )
CMIC$2PRIVATE ( Variables duppliqu\'ees pour chaque subroutine )
CMIC$ CASE
      call sub1
CMIC$ CASE
      call sub2
CMIC$ CASE
      call sub3
CMIC$ CASE
      call sub4
CMIC$ END CASE
CMIC$ END PARALLEL
```

  pour exécuter les sous-programmes `sub1`,`sub2`, `sub3` et `sub4` en parallèle.

– le reste du code est parallélisé par le compilateur seul (*auto-tasking*).
  Les options de compilation sont alors `cf77 -Zp -Wf" ...`

3.7.1 Version (CN2) classique.

Nous avons vu au paragraphe 3 que le schéma (CN2) consiste en la résolution du système linéaire

$$\begin{pmatrix} I_1 & 0 & -\dfrac{\Delta t}{2} D_y^0 \\ 0 & I_2 & \dfrac{\Delta t}{2} D_x^0 \\ -\dfrac{\Delta t}{2} D_y & \dfrac{\Delta t}{2} D_x & Id \end{pmatrix} \begin{pmatrix} \widehat{U}_1^{k+1} \\ \widehat{U}_2^{k+1} \\ \widehat{U}_3^{k+1} \end{pmatrix} = \begin{pmatrix} \widehat{B}_1 \\ \widehat{B}_2 \\ \widehat{B}_3 \end{pmatrix}$$

avec

$$\begin{pmatrix} \widehat{B}_1 \\ \widehat{B}_2 \\ \widehat{B}_3 \end{pmatrix} = \begin{pmatrix} I_1 & 0 & \frac{\Delta t}{2} D_y^0 \\ 0 & I_2 & -\frac{\Delta t}{2} D_x^0 \\ \frac{\Delta t}{2} D_y & -\frac{\Delta t}{2} D_x & Id \end{pmatrix} \begin{pmatrix} \widehat{U}_1^k \\ \widehat{U}_2^k \\ \widehat{U}_3^k \end{pmatrix}$$

On peut décomposer cette résolution en trois étapes :

1. Construction des seconds membres des trois équations ;
2. Résolution de l'équation pour $\widehat{U}_3^{k+1}$;
3. Calcul de $\widehat{U}_1^{k+1}$ et $\widehat{U}_2^{k+1}$.

La part dominante dans l'évaluation du second membre est le calcul des quatre dérivées partielles à l'aide des matrices D_y, D_x et des vecteurs $\widehat{U}_1^k$, $\widehat{U}_2^k$ et $\widehat{U}_3^k$. Ces produits sont répartis sur deux ou quatre processeurs pour plus d'efficacité à l'aide des directives de parallélisation. Il en est de même pour les initialisations, les cumuls.

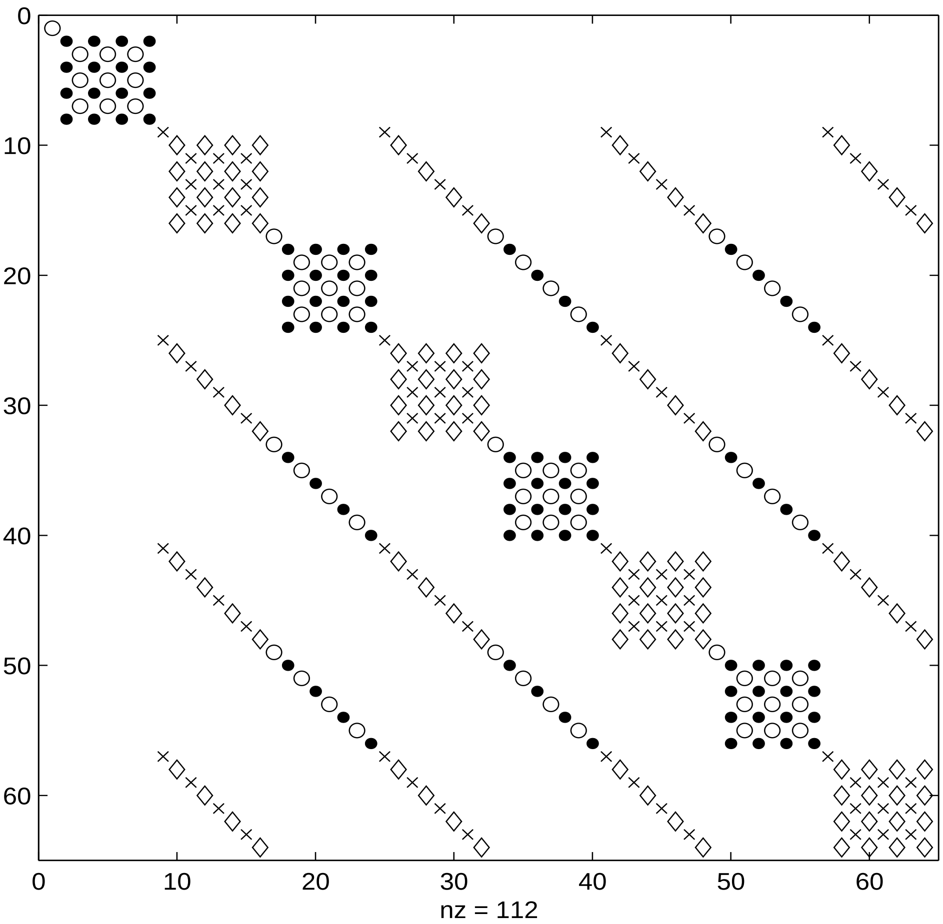
L'inversion du système linéaire se réalise à l'aide des propriétés de parité des polynômes de Legendre : nous remarquons que dans la matrice $\widetilde{M}_3$, pour chaque direction d'espace, les modes pairs et impairs forment des systèmes linéaires disjoints. Cela nous permet de transformer le système linéaire initial pour $\widehat{U}_3^{k+1}$ en quatre petits systèmes de dimension quasi-équivalente, ce qui donne un bon *load-balance* (équilibre des tâches sur les différents processeurs). La figure (3.24) représente les quatre sous-matrices avec différents symboles dans le cas $N = M = 8$ pour plus de lisibilité. La résolution est alors aussi effectuée sur plusieurs processeurs grâce aux directives précédentes.

La dernière étape consiste à déterminer $\widehat{U}_1^{k+1}$ et $\widehat{U}_2^{k+1}$ à l'aide de $\widehat{U}_3^{k+1}$. Là encore, on employe les propriétés de parité des polynômes de Legendre pour séparer $\widehat{U}_1^{k+1}$ et $\widehat{U}_2^{k+1}$ en modes pairs et impairs mais seulement dans la direction spatiale où sont imposées les conditions aux limites, i.e. la direction y pour $\widehat{U}_1^{k+1}$ et x pour $\widehat{U}_2^{k+1}$. On obtient encore 4 produits indépendants, de coût (en terme de nombre d'opérations) similaire.

3.7.2 Version (CN2) à deux niveaux.

L'algorithme est basé sur un point fixe de type Gauss-Seidel. Pour chacune de ses itérations, on doit résoudre un système pour les hautes fréquences puis un système pour les basses fréquences et les modes servant à l'imposition des conditions aux limites.
Ces deux parties peuvent se scinder en trois autres en suivant le découpage de la version classique.
Les valeurs de N_1 et M_1 choisies entraine une dominance écrasante (en terme de temps calcul ou de nombre d'opérations) pour les hautes fréquences par rapport aux autres termes. Il est donc important que la résolution de leur équation soit optimale sur plusieurs processeurs. Toujours grâce aux propriétés de parité des polynômes de Legendre, nous pouvons décomposer la matrice $\widetilde{M}_{\mathrm{w}_3}$ en quatre sous-matrices triangulaires supérieures avec des 1 sur la diagonale et inverser le système sur quatre processeurs de manière optimale.

FIG. 3.24 – *Squelette de la matrice* $\widetilde{M_3}$ *pour (CN2) avec* $(N, M) = (8, 8)$.



Les parties du code traitant la résolution de l'équation des basses fréquences et des modes"Tau" sont quant à elles compilées en mode *autotasking*.

3.7.3 Résultats.

Pour tester l'efficacité de cette décomposition, nous prenons le test 3 sur l'intervalle temporel $[0, 2T]$. Le logiciel *ja* permet d'établir le taux moyen d'occupation des processeurs alloués.

2 processeurs alloués

(CN2) cl	1,97
(CN2) 2n	1,94

(CDIRK4) cl	1,92
(CDIRK4) 2n	1,94

4 processeurs alloués

(CN2) cl	3,56
(CN2) 2n	3,74

(CDIRK4) cl	3,58
(CDIRK4) 2n	3,72

Ces tests ont été réalisés sur les Cray YMP des universités Blaise Pascal de Clermont-Ferrand et d'Orsay durant les périodes où la machine était le moins chargée, voire totalement dédiée pour le cas quadri-processeur.
Nous remarquons que le cas bi-processeur est quasi-optimal, avec plus de 95 % de temps sur les deux processeurs. Dans le cas quadri-processeurs, les résultats sont de qualité équivalente : pendant près de 90 % du temps CPU, les quatre processeurs travaillent en parallèle.

3.8 Conclusions.

Dans le cadre de la résolution des équations de Maxwell pour le problème modèle de la cavité résonnante, nous nous sommes intéressés au traitement de conditions aux limites de type Dirichlet homogène pour une méthode spectrale multi-niveaux. Pour cela, nous avons choisi une méthode spectrale utilisant une base orthogonale au sens L^2 pour un maniement aisé des projection des fonctions : la méthode Tau-Legendre.
Un algorithme de point-fixe de type Gauss-Seidel, accéléré par une méthode de sous-relaxation permet de résoudre ce problème de conditions aux limites.
La contrainte de convergence du point fixe, relativement sévère, nous impose de prendre des pas de temps Δt assez petits mais cela présente l'avantage d'avoir des erreurs globales d'approximation assez faibles. Nous notons que la méthode classique ne supporte pas une telle contrainte.
Pour des schémas en temps explicites, la linéarité des équations à résoudre enlève tout intérêt à l'usage de cette méthode. Par contre, pour des schémas en temps semi-implicites, cette méthode permet la résolution exacte des systèmes linéaires.

Chapitre 4

Étude de l'équation de Burgers.

4.1 Introduction

L'équation parabolique quasi-linéaire

$$\frac{\partial u}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2} - u \frac{\partial u}{\partial x} \tag{4.1}$$

est connue sous le nom d'équation de Burgers, où la quantité réelle inconnue $u(x,t)$, la vitesse, est fonction d'une variable de temps t et d'une variable d'espace x; on prescrit la condition initiale $u_0(x) = u(x,0)$ et des conditions aux limites (e.g. périodiques), le coefficient ν (viscosité) appartient à $\mathbb{R}_+$.
Forme asymptotique de nombreux systèmes à la fois non linéaires et dissipatifs ([45]), elle modélise avec succès certaine dynamique des gaz ([38]), des phénomènes accoustiques ([7]). Burgers, ([10],[11]), l'étudia de manière intensive comme modèle mathématique pour la turbulence.
Elle possède un intérêt physique certain du fait de ses propriétés statistiques et de son rôle dans la hiérarchie des approximations des équations de Navier-Stokes.
Cette équation est un très bon modèle pour les équations de Navier-Stokes puisqu'elle représente, de la manière la plus simple, l'équilibre entre le processus de convection non linéaire $\left(u\dfrac{\partial u}{\partial x}\right)$, dont l'importance peut être mesurée par un nombre de Reynolds, Re, proportionnel à ν^{-1} et le processus de dissipation $\left(\dfrac{\partial^2 u}{\partial x^2}\right)$.
En tant que modèle de certains aspects de la turbulence, l'équation de Burgers doit être étudiée à grand nombre de Reynolds ou même dans la limite $\nu \searrow 0$.
Bien que l'équation de Burgers soit non linéaire, elle possède une solution exacte pour de nombreuses combinaisons de conditions limites et initiales ([5]). Pour cette raison, elle a aussi souvent été employée pour établir la précision de schémas.

Solution générale : Transformation de Cole-Hopf.
Considérons le problème (de Cauchy) à résoudre :

$$\frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial u^2}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2} \quad \nu > 0 \ , \quad t \in (0,T), \ x \in \mathbb{R} \tag{4.2}$$

$$u(x,0) = u_0(x) \tag{4.3}$$

Cole et Hopf (resp. [15], [30]) ont défini une transformation qui permet de déterminer les solutions de (4.2) - (4.3) (sur un domaine infini) en fonction des solutions de l'équation de la Chaleur. Dans ce cas (dans le cas périodique en espace, on peut se ramener à des solutions sur un borné $x \in [a,b]$), une liste exhaustive des solutions connues à partir de cette transformation a été établie par Benton et Platzmann ([5]). Nous allons présenter cette transformation.
Si $\Theta(x,t)$ ($\Theta \not\equiv 0$) est solution de l'équation de la Chaleur :

$$\frac{\partial \Theta}{\partial t} = \nu \frac{\partial^2 \Theta}{\partial x^2} \tag{4.4}$$

alors

$$u(x,t) = \frac{-2\nu}{\Theta} \frac{\partial \Theta}{\partial x} \tag{4.5}$$

est solution de

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2}$$

Regardons la condition initiale :
soit $u_0(x)$ la condition initiale, d'après (4.5) on a $u_0(x) = \dfrac{-2\nu}{\Theta} \dfrac{\partial \Theta}{\partial x}$
donc

$$\Theta(x,0) = \Theta_0(x) = \exp\left\{-\frac{1}{2\nu}\int_0^x u_0(\tau)\,\mathrm{d}\tau\right\}$$

et la solution, $\Theta(x,t)$, de (4.4) vérifie :

$$\Theta(x,t) = \frac{1}{\sqrt{4\pi\nu t}} \int_{-\infty}^{+\infty} exp\left[-\frac{(x-\tau)^2}{4\nu t}\right] \Theta_0(\tau)\mathrm{d}\tau \tag{4.6}$$

Maintenant, sachant calculer la solution $\Theta(x,t)$ de (4.4) grâce à la relation (4.6), on peut déterminer la solution $u(x,t)$ avec la formule (4.5), ce qui nous donne de manière unique la solution du problème (4.2) - (4.3).

Nous considérons maintenant l'équation de Burgers forcée. Nous étudierons deux cas : le forçage déterministe et le forçage aléatoire. Nous commencerons l'étude par des résultats d'existence et d'unicité de solutions avant de présenter les résultats obtenus.

Dans le cadre de la modélisation de la turbulence, nous considérons ici l'équation de Burgers 1D comme un modèle simplifié des équations de Navier-Stokes. Nous prenons dans un premier temps un forçage déterministe similaire à ceux employés dans ([31, 23]) pour résoudre les équations de Navier-Stokes. Nous employons les projections de la solution sur les grandes et petites échelles introduites au chapitre 1.
Ensuite nous étudions l'équation de Burgers stochastique, i.e. munie d'un forçage aléatoire de type bruit blanc, dans les deux situations suivantes : conditions aux limites périodiques et de non glissement. L'intégration en temps doit être alors effectuée sur de longs intervalles de temps et nous considérons les moyennes temporelles des quantités provenant des projections sur les grandes et petites échelles.

4.2 Résolution de l'équation de Burgers déterministe.

On appelle équation de Burgers généralisée, l'équation (4.1) munie d'un second membre f appelée force extérieure. Nous allons d'abord étudier le cas où f est une force déterministe, i.e. indépendante du temps, dans le cas de conditions aux limites périodiques.

$$\frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} + u \ \frac{\partial u}{\partial x} = f$$

L'équation de Burgers généralisée est une forme simplifiée des équations de Navier-Stokes qu'elle approche qualitativement (le lecteur peut se référer à ([47, 48]) pour ces dernières). Elle conserve les parties convective et dissipative mais les termes du gradient de pression et de la divergence de la vitesse ne sont pas retenus. Ainsi la solution peut présenter de forts gradients dûs à l'interaction du terme non linéaire et du terme dissipatif.

4.2.1 Résultats d'existence et d'unicité.

4.2.1.1 Cadre fonctionnel.

Conditions aux limites.
On pose $\Omega = (0, 2\pi)$ et $\Gamma = \partial\Omega$.
On supposera dorénavant que u est périodique de période 2π :

$$u(x + 2\pi, t) = u(x, t)$$

Espaces de Lebesgue.
Nous introduisons d'abord l'espace $L^2(\Omega)$ des classes de fonctions réelles sur Ω de carré intégrable pour la mesure de Lebesgue, i.e. $\displaystyle\int_\Omega |u|^2 \,\mathrm{d}x < +\infty$. $L^2(\Omega)$ est muni du produit scalaire

$$(u, v) = \int_\Omega u\, v \,\mathrm{d}x$$

et de la norme associée

$$|\, u \,| = (u, u)^{\frac{1}{2}} = \left\{ \int_\Omega |u|^2 \,\mathrm{d}x \right\}^{\frac{1}{2}}$$

Dans le cas de conditions limites périodiques, nous supposons que la moyenne de l'écoulement est nulle, i.e.

$$\int_\Omega u \,\mathrm{d}x = 0$$

Ensuite nous introduisons le sous-espace de $L^2(\Omega)$

$$\dot{L}^2_{\text{per}}(\Omega) = \left\{ u \in L^2(\Omega) : u \text{ est périodique sur } \Omega \text{ et } \int_\Omega u \,\mathrm{d}x = 0 \right\}$$

L'espace de Sobolev $H^m(\Omega)$, pour $m \geqslant 0$, est l'espace des fonctions appartenant, ainsi que leurs dérivées jusqu'à l'ordre m, à $L^2(\Omega)$.
L'espace $H^m(\Omega)$ est muni du produit scalaire

$$(\!(u, v)\!)_m = \sum_{|\alpha| \leqslant m} \left(\frac{\partial^\alpha u}{\partial \mathrm{x}^\alpha}, \frac{\partial^\alpha v}{\partial \mathrm{x}^\alpha} \right)$$

La norme associée est définie par

$$\| u \|_m = ((u, u))_m^{\frac{1}{2}}$$

$H^m(\Omega)$ est clairement un espace de Hilbert.
Nous introduisons $\dot{H}^m_{\text{per}}(\Omega)$ le sous-espace de $H^m(\Omega)$ des fonctions dans $\dot{L}^2_{\text{per}}(\Omega)$ ainsi que leurs dérivées jusqu'à l'ordre m, qui est un espace de Hilbert pour le produit scalaire $((.,.))_m$ et la norme $\| . \|_m$.

Nous considérons maintenant les deux espaces de fonctions H et V qui jouerons un rôle essentiel pour la suite.
Nous définissons d'abord l'espace de Hilbert H qui est un sous-espace fermé de $L^2(\Omega)$:

$$H = \dot{L}^2_{\text{per}}(\Omega)$$

Le second espace de Hilbert, V, est un sous-espace fermé de $H^1(\Omega)$:

$$V = \dot{H}^1_{\text{per}}(\Omega)$$

V est un espace de Hilbert pour le produit scalaire

$$((u, v)) = \left(\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x} \right)$$

et la norme associée :

$$\| u \| = ((u, u))^{\frac{1}{2}}$$

L'opérateur A.
A une forme bilinéaire a

$$a : \begin{array}{ccc} V \times V & \longrightarrow & \mathbb{R} \\ (u, v) & \longmapsto & a(u, v) = ((u, v)) \end{array}$$

nous pouvons associer un opérateur linéaire non borné dans H défini par

$$(Au, v) = ((u, v)) \ , \ \forall \, u, \, v \in V$$

Nous désignons par $D(A)$ le domaine de A

$$D(A) = \{ u \in V, \, Au \in H \}$$

Dans le cas de conditions aux limites périodiques, $D(A)$ peut être désigné par

$$D(A) = \left(\dot{H}^2_{\text{per}}(\Omega) \right) \cap V$$

De plus l'application $u \mapsto |Au|$ est une norme sur $D(A)$ équivalente à celle induite par $H^2(\Omega)$.
Nous avons, dans le cas de conditions aux limites périodiques en espace

$$Au = - \frac{\partial^2 u}{\partial x^2} \ , \ \forall u \in D(A)$$

On peut se référer à [49] pour les propriétés de l'opérateur A.

Formulation variationnelle.
Supposons que u soit une solution régulière de l'équation de Burgers, e.g. nous supposons que, à chaque instant t, $u(.,t)$ définie par

$$u(.,t): \begin{array}{ccc} \Omega & \longrightarrow & \mathbb{R} \\ x & \longmapsto & u(x,t) \end{array}$$

appartienne à V. Soit v une fonction test de V.
En multipliant l'équation de Burgers par v et en intégrant sur le domaine Ω nous obtenons

$$\frac{d}{dt}(u,v) + \nu\,((u,v)) + b(u,u,v) = (f,v)\ ,\ \forall v \in V \tag{4.7}$$

où

$$b(u,v,w) = \int_\Omega \frac{2}{3} u \frac{\partial v}{\partial x} w \,\mathrm{d}x + \int_\Omega \frac{1}{3} v \frac{\partial u}{\partial x} w \,\mathrm{d}x$$

L'équation (4.7) suggère alors la formulation suivante du problème

<u>Solutions faibles :</u>
Pour f et u_0 données,

$$f \in L^2(0,T;V') \text{ et } u_0 \in H,$$

trouver u satisfaisant

$$\begin{cases} \frac{d}{dt}(u,v) + \nu\,((u,v)) + b(u,u,v) = (f,v) \ \ \forall v \in V \\ u(t=0) = u_0 \end{cases} \tag{4.8}$$

Formulation abstraite.
En utilisant les propriétés de l'opérateur linéaire A défini précédemment et en introduisant un opérateur bilinéaire B, induit par la forme trilinéaire b, nous pouvons écrire l'équation de Burgers comme une équation différentielle dans V' (V' est l'espace dual de V).
La forme trilinéaire b est continue sur $V \times V \times V$ et satisfait l'inégalité suivante :

$$|b(u,v,w)| \ \leqslant\ c\,|\,u\,|^{\frac{1}{2}}\,\|\,u\,\|^{\frac{1}{2}}\,\|\,v\,\|^{\frac{1}{2}}\,|\,w\,|^{\frac{1}{2}}\,\|\,w\,\|^{\frac{1}{2}} \quad \forall u,v,w \in V$$

Pour u, v, $w \in V$, nous définissons $B(u,v)$ élément de V' en posant

$$< B(u,v), w >_{V',V} = b(u,v,w)$$

Puisque b est une forme trilinéaire, B est un opérateur continu et bilinéaire de $V \times V$ dans V' :
$B\,:\,V \times V \to V'$
Nous posons $B(u) = B(u,u)\,,\ \ \forall u \in V$
Si nous supposons que $u \in L^2(0,T;V)$ alors la fonction $B(u)\ :\ t \mapsto B(\,u(t)\,)$ appartient à $L^1(0,T;V)$. De plus $Au \in V'$, $f \in V'$ et l'équation (4.7) est équivalente à

$$\frac{d}{dt} < u,v >_{V',V} = < f - \nu Au - B(u)\ ,v >_{V',V}$$

Comme $(f - \nu Au - B(u)) \in L^1(0,T;V')$ alors $\frac{du}{dt} \in L^1(0,T;V')$.
Nous obtenons finalement l'égalité suivante dans V'

$$\frac{du}{dt} + \nu Au + B(u) = f$$

qui est une formulation forte de l'équation de Burgers.
On peut prouver que u est p.p. égale à une fonction continue de $[0,\ T]$ dans V', ce qui donne un sens à $u(0)$.
Nous pouvons proposer une version forte du problème original.

Solutions fortes :
Pour f et u_0 données,

$$f \in L^2(0,T;H) \text{ et } u_0 \in V,$$

trouver u dans $L^2(0,T;D(A)) \cap L^\infty(0,T;V)$ satisfaisant

$$\begin{cases} \frac{du}{dt} + \nu Au + B(u) = f \\ u(t=0) = u_0 \end{cases}$$

4.2.1.2 Théorèmes d'existence et d'unicité.

Dans cette section, nous rappelons les résultats classiques d'existence et d'unicité ([47]) pour l'équation de Burgers dans un domaine borné de $\mathbb{R}$. Commençons par la formulation faible.

Théorème *4.8*
Pour f et u_0 données,

$$f \in L^2(0,T;V') \text{ et } u_0 \in H,$$

il existe une solution faible u de l'équation de Burgers satisfaisant

$$u \in L^2(0,T;V) \cap L^\infty(0,T;H).$$

De plus, u est unique et

$$\begin{cases} u \in C([0,\ T];H) \\ \frac{du}{dt} \in L^2(0,T;V') \end{cases}$$

Voici maintenant un résultat équivalent concernant les solutions fortes.

Théorème *4.9*

Pour f et u_0 données,

$$f \in L^2(0, T; H) \text{ et } u_0 \in V,$$

Il existe une unique solution u de l'équation de Burgers vérifiant

$$u \in L^2(0, T; D(A)), \frac{du}{dt} \in L^2(0, T; H).$$

et

$$u \in C\left([0,\, T]; V\right).$$

Enfin, rappelons que dans le cas de conditions aux limites périodiques en espace, la régularité de u est complètement déterminée par la régularité des données f et u_0 :

Théorème *4.10*

si $u_0 \in C^\infty(\mathbb{R})$ et $f \in C^\infty(\mathbb{R} \times [0, T])$ alors u appartient à $C^\infty(\mathbb{R} \times [0, T])$.

4.2.2 Présentation du problème.

Le but de l'application de cette force est d'apporter suffisamment d'énergie pour, qu'associée à une viscosité suffisamment petite, elle entretienne un mouvement chaotique pendant une phase transitoire avant d'obtenir la convergence vers une solution stationnaire. L'écoulement est entrainé par une force extérieure indépendante du temps qui agit, dans l'espace spectral, sur seulement quelques modes de basse fréquence de la vitesse. La condition initiale est choisie de telle sorte que le spectre d'énergie de la vitesse ait une forme donnée et ses phases sont déterminées aléatoirement.
Ainsi à $t = 0$, l'écoulement n'a pas de réelles structures.

4.2.2.1 Description de la condition initiale.

On part de la condition initiale donnée sur le spectre de la vitesse, u.
On définit

$$E(k) = \sum_{\substack{|l|=k \\ l \in \mathbb{Z}}} |\hat{u}_l|^2 = |\hat{u}_{-k}|^2 + |\hat{u}_k|^2$$

Cela nous donne

$$\begin{aligned} E(0) &= E(N/2) = 0 \text{ (u fonction réelle à moyenne nulle)} \\ E(k) &= c_1\, k^{-1}\, e^{-k} \text{ pour } k \in \left[\!\left[1,\, \frac{N}{2} - 1\right]\!\right] \end{aligned}$$

pour la condition initiale;
c_1 est une constante à déterminer en imposant $|\, u_0\, |_{L^2(0,2\pi)} = c_U$, c_U donnée.

Des relations

$$\begin{array}{lcl} E(k) & = & |\hat{u}_k(t_0)|^2 + |\hat{u}_{-k}(t_0)|^2, \quad k \in \left[\!\left[1, \dfrac{N}{2}-1\right]\!\right] \\ \hat{u}_k(t_0) & = & |\hat{u}_k(t_0)|\, e^{i\theta_k}, \quad \theta_k \in [0, 2\pi], \ k \in \left[\!\left[1, \dfrac{N}{2}-1\right]\!\right] \\ |\, u_0\, |_{L^2(0,2\pi)} & = & c_U \end{array}$$

on déduit les égalités

$$c_1 = \frac{(c_U)^2}{2\pi \displaystyle\sum_{k=1}^{N/2-1} k^{-1} e^{-k}}, \quad |\hat{u}_k(t_0)| = \sqrt{\frac{c_1}{2}\, k^{-1} e^{-k}} \ , \ k \in \left[\!\left[1, \frac{N}{2}-1\right]\!\right]$$

4.2.2.2 La force extérieure.

On construit la force extérieure f comme étant indépendante du temps, ce qui signifie que son spectre de Fourier est constant. D'autre part, on privilégie un nombre restreint de modes où sera concentrée son énergie. On se donne $c_F = |\, f\, |_{L^2(0,2\pi)}$ et on définit le spectre de f par

$$\hat{f}_k = \begin{cases} c_2 e^{i\theta_k} & , \ k \in [\![4,\, 6]\!] \ , \ \ \theta_k \in [0,\, 2\pi] \text{ aléatoire} \\ 0 & , \ k \in \left[\!\left[0,\, \dfrac{N}{2}\right]\!\right] \setminus [\![4,\, 6]\!] \end{cases}$$

c_2 est déterminée par $c_2 = \dfrac{c_F}{\sqrt{12\pi}}$ en utilisant des relations analogues à celles qui nous ont permis de déterminer c_1.

4.2.2.3 Choix des paramètres.

On travaille avec les paramètres suivants :
- la fréquence de coupure N égale à 512 pour $\nu = 10^{-2}$ et 3072 pour $\nu = 10^{-3}$,
- la norme de f, $|\, f\, |_{L^2(0,2\pi)}$, choisie égale à 1.

La norme de la condition initiale $|\, u(t=0)\, |_{L^2(0,2\pi)}$ est déterminée de la manière suivante : considérons l'équation de Burgers

$$\frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} + u\, \frac{\partial u}{\partial x} = f \tag{4.9}$$

En prenant le produit scalaire de (4.9) avec u elle-même et en utilisant la propriété d'orthogonalité :

$$\left(u\, \frac{\partial u}{\partial x}, u \right) = 0$$

FIG. 4.1 – *Evolution en temps du nombre de Courant (gauche) et des quantités* (f_N, u_N) *(1) et* $\nu \,\| u_N \|^2_{H^1(0,2\pi)}$ *(2) (droite).*

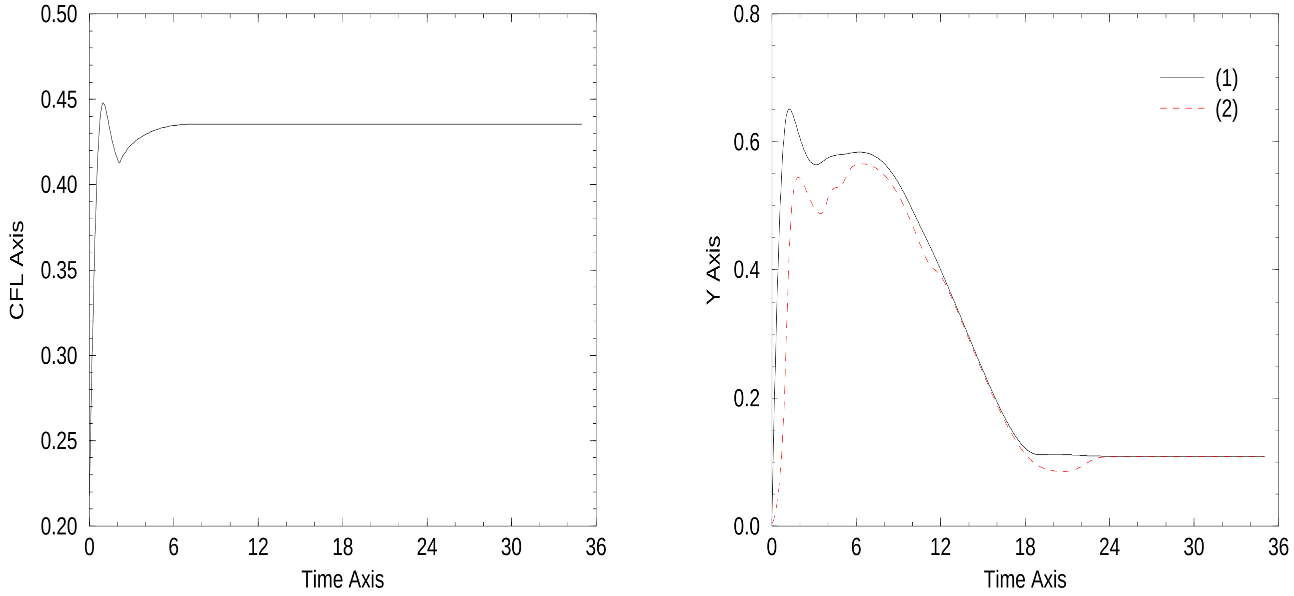


FIG. 4.2 – *Evolution en temps des normes* $| u_N |_{L^2(0,2\pi)}$ *et* $| u_N |_{L^\infty(0,2\pi)}$ *de la vitesse.*

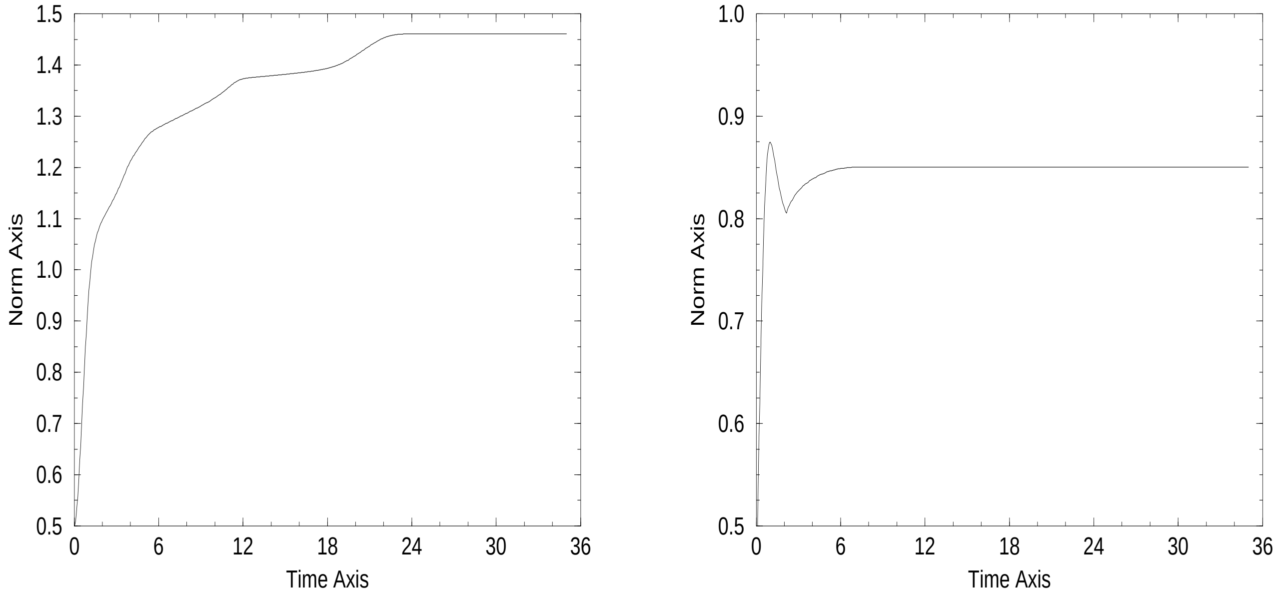


FIG. 4.3 – *Evolution en temps des normes* $|\,y_{N_1}\,|_{L^2(0,2\pi)}$ *(gauche) et* $|\,z_{N_1}\,|_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* N_1 *= 12 (1), 32 (2), 64 (3), 80 (4) avec* N *= 512.*

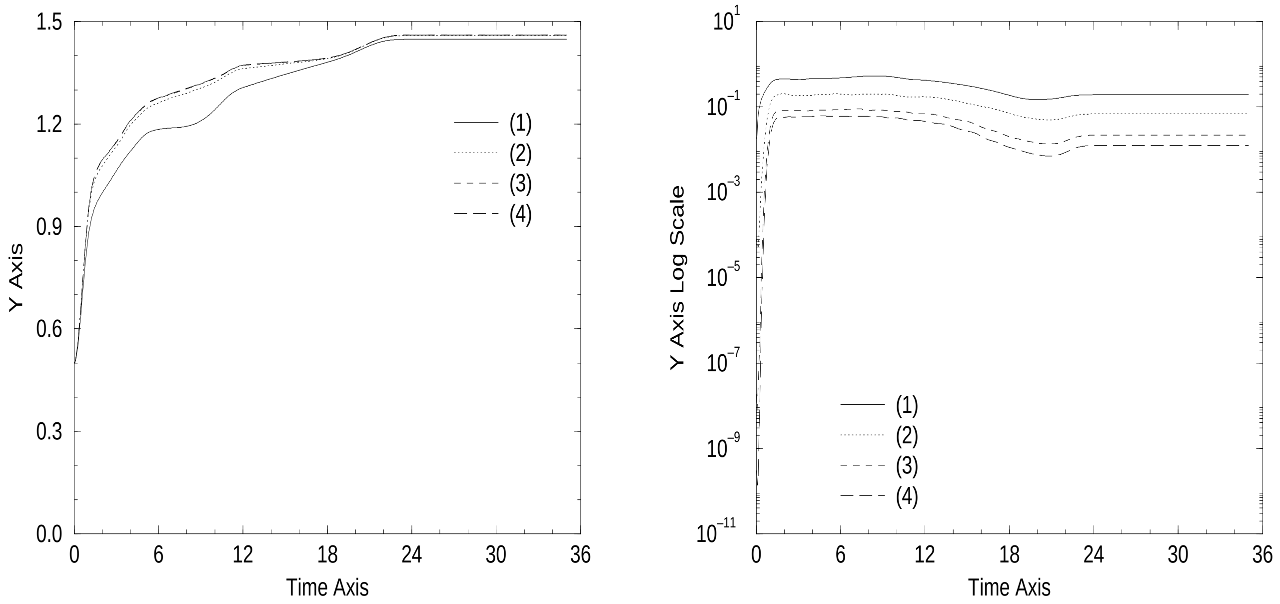


FIG. 4.4 – *Evolution en temps des normes* $|\,\dot{y}_{N_1}\,|_{L^2(0,2\pi)}$ *(gauche) et* $|\,\dot{z}_{N_1}\,|_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* N_1 *= 12 (1), 32 (2), 64 (3), 80 (4) avec* N *= 512.*

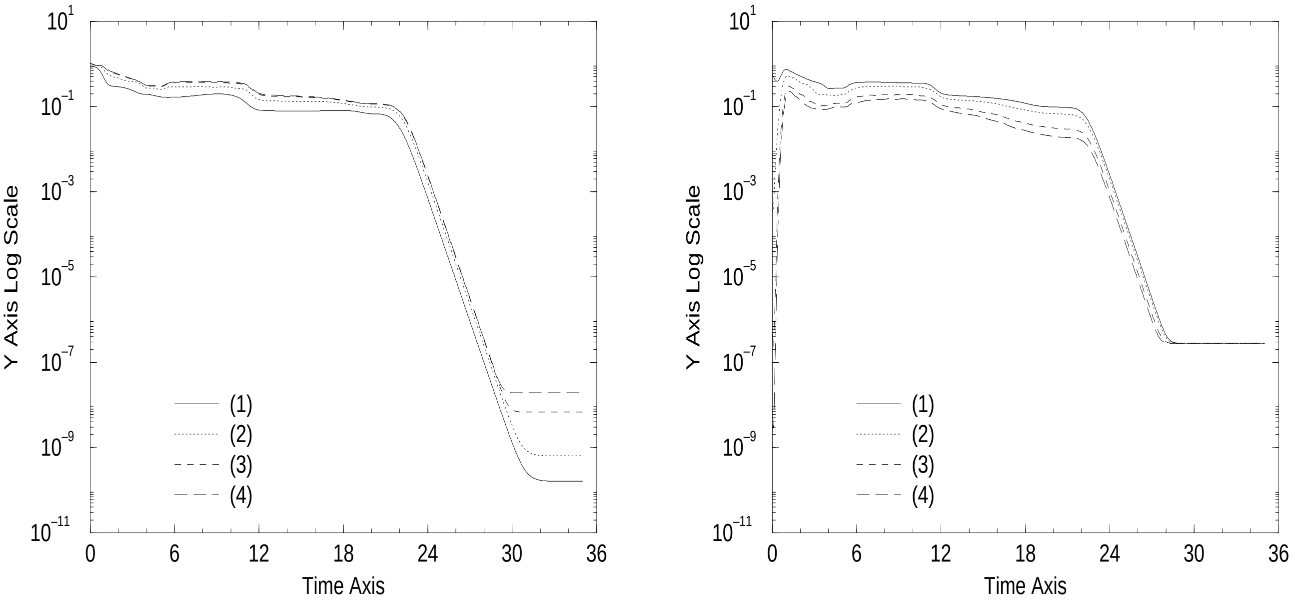


FIG. 4.5 – *Evolution en temps des normes* $\nu\left|\dfrac{\partial^2 y_{N_1}}{\partial x^2}\right|_{L^2(0,2\pi)}$ *(gauche) et* $\nu\left|\dfrac{\partial^2 z_{N_1}}{\partial x^2}\right|_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* $N_1 = 12$ *(1), 32 (2), 64 (3), 80 (4) avec* $N = 512$.

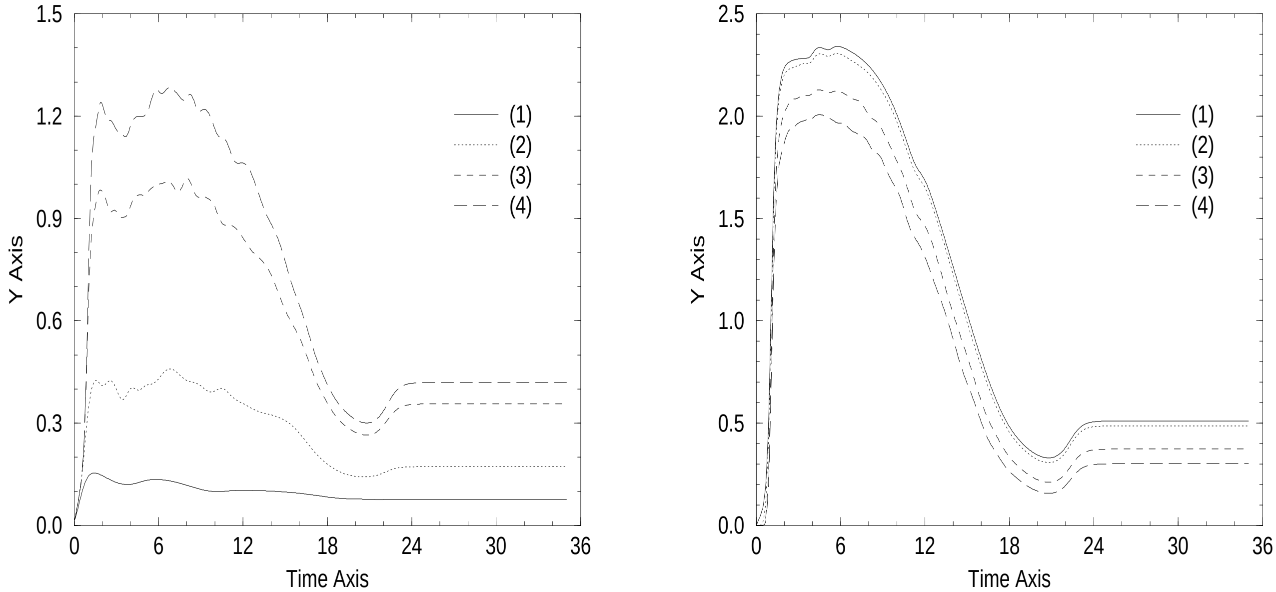
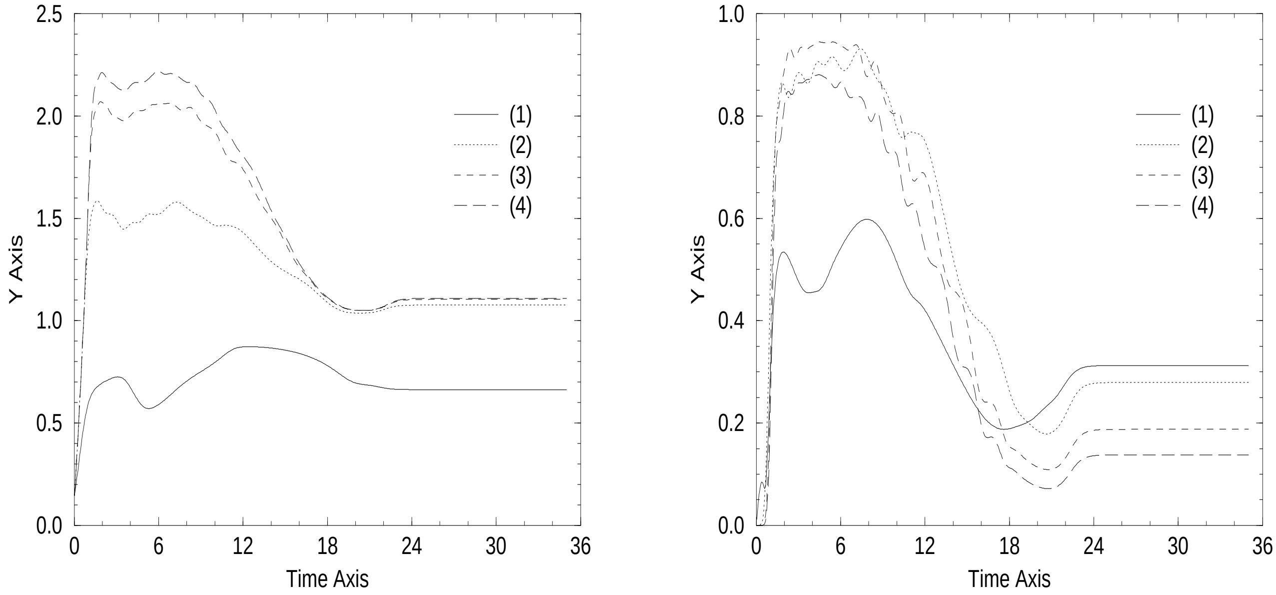


FIG. 4.6 – *Evolution en temps des quantités* $\mid P_{N_1}B(y_{N_1}, y_{N_1}) \mid_{L^2(0,2\pi)}$ *(gauche) et* $\mid P_{N_1}B_{int}(y_{N_1}, z_{N_1})\mid_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* $N_1 = 12$ *(1), 32 (2), 64 (3), 80 (4) avec* $N = 512$.



nous obtenons l'égalité d'énergie bien connue

$$\frac{1}{2}\frac{d}{dt}\left(|\,u\,|_{L^2(0,2\pi)}\right) + \nu\left(\|\,u\,\|_{H^1(0,2\pi)}\right)^2 = (f,u) \tag{4.10}$$

que l'on reécrit

$$\frac{1}{2}\frac{d}{dt}\left(|\,u\,|_{L^2(0,2\pi)}\right) = (f,u) - \nu\left(\|\,u\,\|_{H^1(0,2\pi)}\right)^2$$

On choisit $u_0 = u(t=0)$ de telle sorte que

$$(f,u_0) \approx \frac{3}{2}\nu\left(\|\,u_0\,\|_{H^1(0,2\pi)}\right)^2 \tag{4.11}$$

Ce qui implique

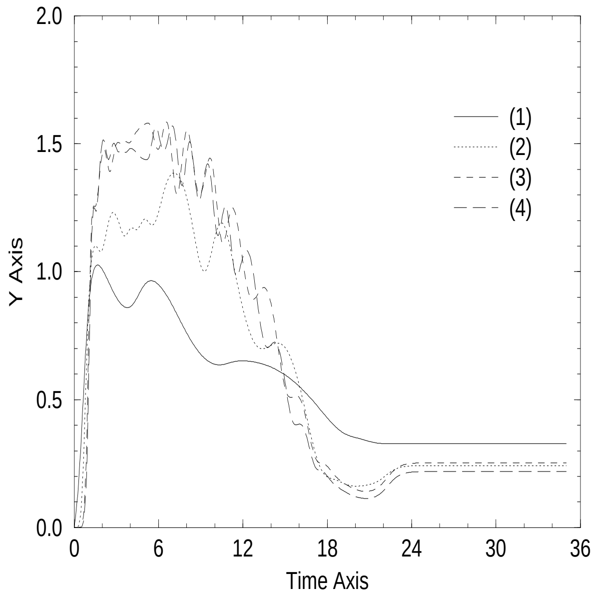
$$\frac{d}{dt}\left(|\,u_0\,|_{L^2(0,2\pi)}\right) = \nu\left(\|\,u_0\,\|\right)^2 > 0$$

On choisit ainsi $|\,u_0\,|_{L^2(0,2\pi)}= 0.5$ pour $\nu = 10^{-2}$ et $|\,u_0\,|_{L^2(0,2\pi)}= 1$ pour $\nu = 10^{-3}$. Enfin, on adapte le pas de temps Δt pour que le schéma vérifie la condition de type (C.F.L.).

4.2.2.4 Présentation des résultats.

Simulation pour $\nu = 10^{-2}$.
Le schéma est stable, figure (4.1) avec une condition de stabilité $N\Delta t\,|\,u_N\,|_{L^\infty(0,2\pi)} < \alpha$ avec $\alpha = 0,5$ qui, après une forte croissance dûe à la perturbation initiale, se stabilise pour demeurer constante jusqu'à convergence. La condition de stabilité étant proportionnelle à la norme $\|\,u_N\,\|_{L^\infty}$, on en déduit un comportement similaire pour celle-ci, figure (4.2). La condition initiale a été choisie de telle sorte que la relation (4.11) soit réalisée.
Ainsi $|\,u_N\,|_{L^2(0,2\pi)}$ est initialement croissante, figure (4.2), le temps que l'énergie de la condition initiale se propage dans le spectre. Ensuite, la croissance est moins forte : par "pallier". Ce comportement s'explique à l'aide de l'évolution des quantités de l'équation (4.10), figure (4.1). Les deux quantités (f_N, u_N) et $\nu\left(\|\,u_N\,\|\right)^2$ présentent des allures similaires avec des pentes identiques, ainsi durant ces périodes la quantité $\frac{d}{dt}\,|\,u_N\,|^2_{L^2(0,2\pi)}$ est constante, ce qui implique le comportement observé.
Nous considérons maintenant deux niveaux dans le spectre de u_N, à l'aide des opérateurs P_{N_1} et Q_{N_1} définis au §1.3.5. Pour $N = 512$, nous avons choisi les valeurs suivantes pour le second niveau : $N_1 = 12,\ 32,\ 64,\ 80$.
Les grandes échelles, figure (4.3) présentent un comportement similaire au champ complet de la vitesse (figure 8), seul $|\,y_{N_1}\,|_{L^2(0,2\pi)}$ pour $N_1 = 12$ se distingue des autres. La figure (4.3) nous montre aussi l'importance relative des petites échelles par rapport aux grandes.
Les dérivées temporelles, figure (4.4), après une forte augmentation dûe à l'apport initial d'énergie, gardent une valeur constante en restant très proches les unes des autres, avant d'entamer une décroissance exponentielle qui se stabilise avec la convergence de la solution calculée. Par contre, pour le terme dissipatif, on peut distinguer les grandes des petites échelles et les différents niveaux de ces deux quantités, figure (4.5). La dissipation se fait essentiellement dans les petites échelles, même si l'apport des grandes échelles n'est pas négligeable.
Les figures (4.6) et (4.7) nous renseignent sur l'évolution des quantités formant le terme non linéaire :

FIG. 4.7 – *Evolution en temps des quantités* $| Q_{N_1} B(y_{N_1}, y_{N_1}) |_{L^\infty(0,2\pi)}$ *(gauche) et* $|Q_{N_1} B_{int}(y_{N_1}, z_{N_1})|_{L^\infty(0,2\pi)}$ *(droite) pour différentes valeurs de* N_1 = *12 (1), 32 (2), 64 (3), 80 (4) avec* N = *512.*



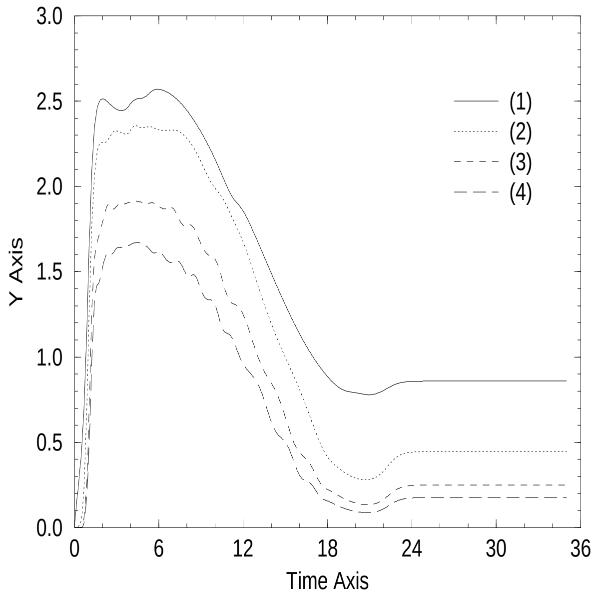
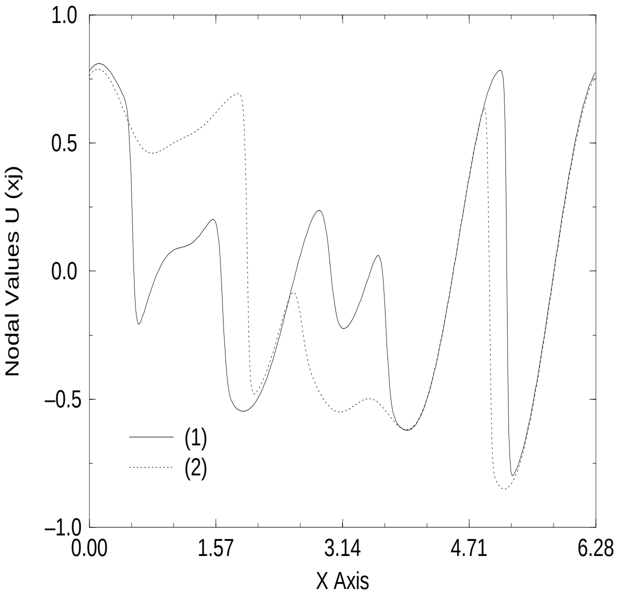


FIG. 4.8 – *Valeurs instantannées de la vitesse à différents temps :* t = *2 (1), 10 (2), 18 (3), 26 (4).*



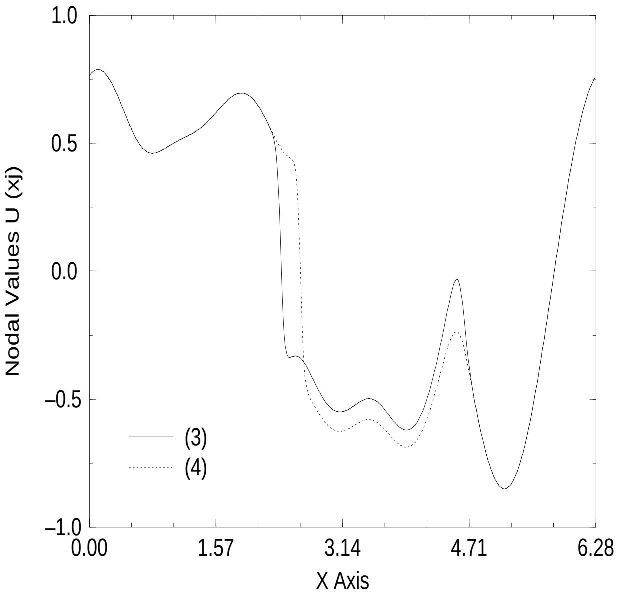


FIG. 4.9 – *Valeurs instantannées de la vitesse à $t = 35$ (5) et de la force extérieure (6) (gauche), spectre d'énergie à $t = 6$ (droite).*

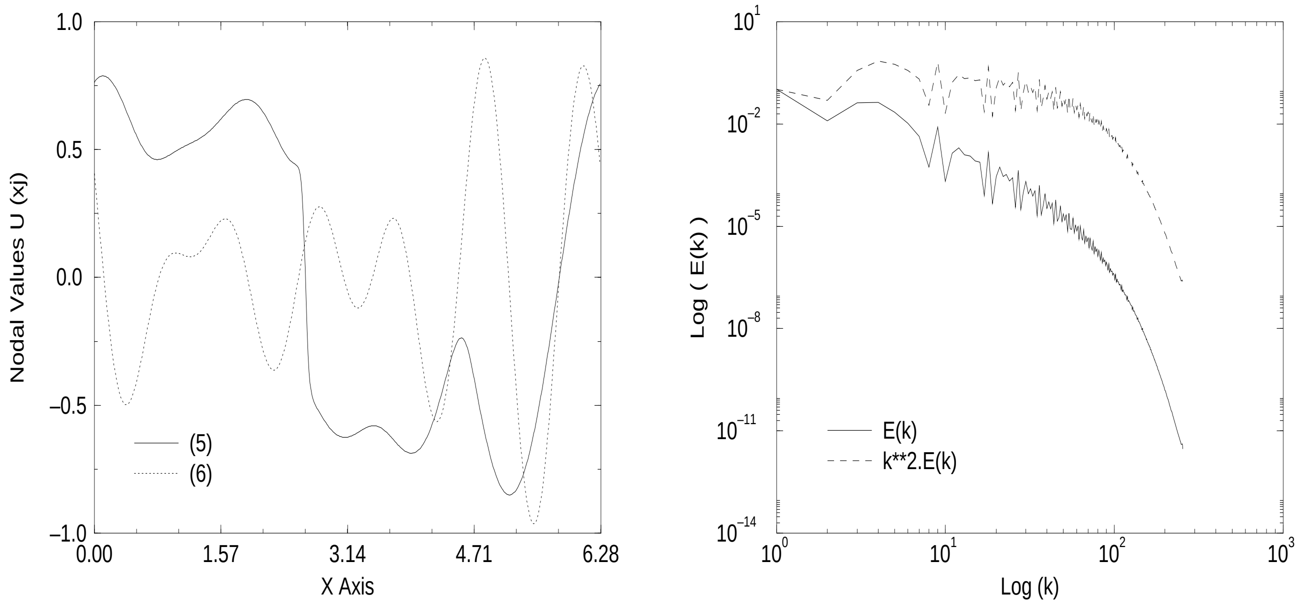


FIG. 4.10 – *Spectres d'énergie de la vitesse à $t = 14$ et 35.*

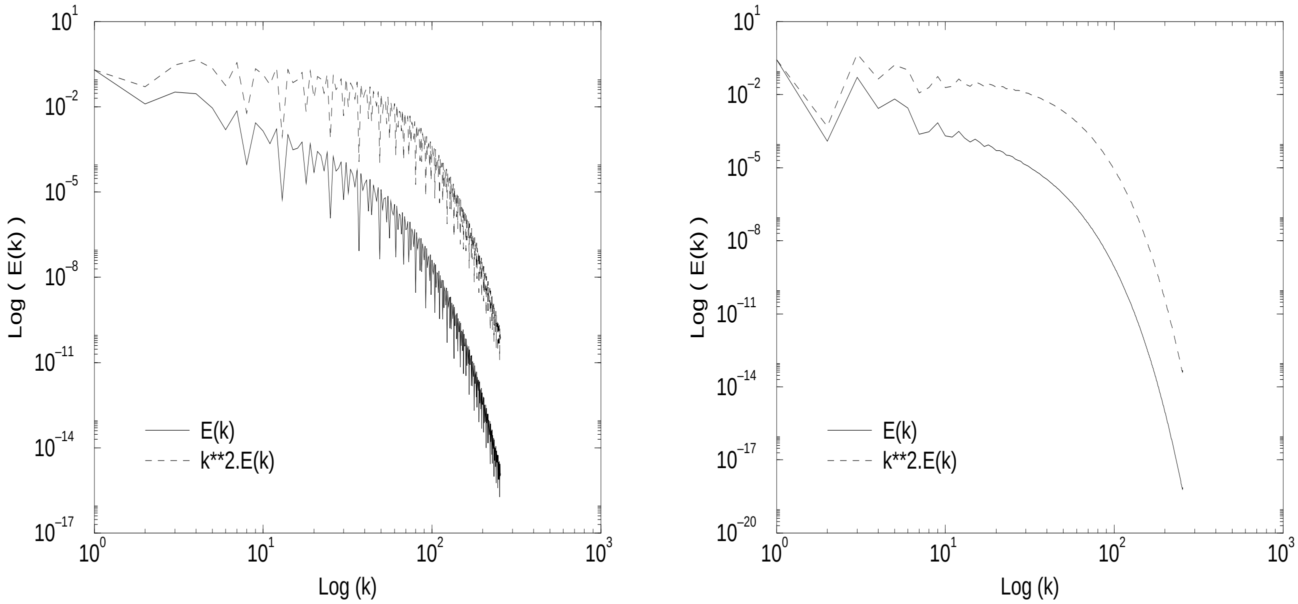


FIG. 4.11 – *Evolution en temps du nombre de Courant (gauche) et des quantités* (f_N, u_N) *(1) et* $\nu(\| u_N \|^2_{H^1(0,2\pi)}$ *(2) (droite).*

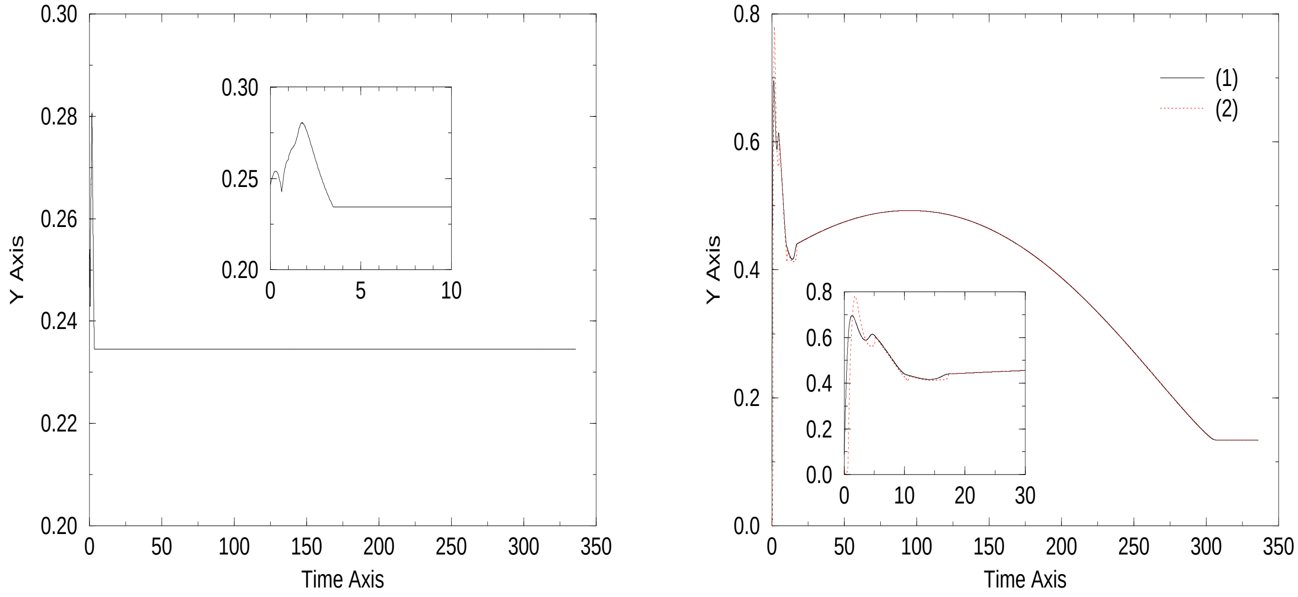
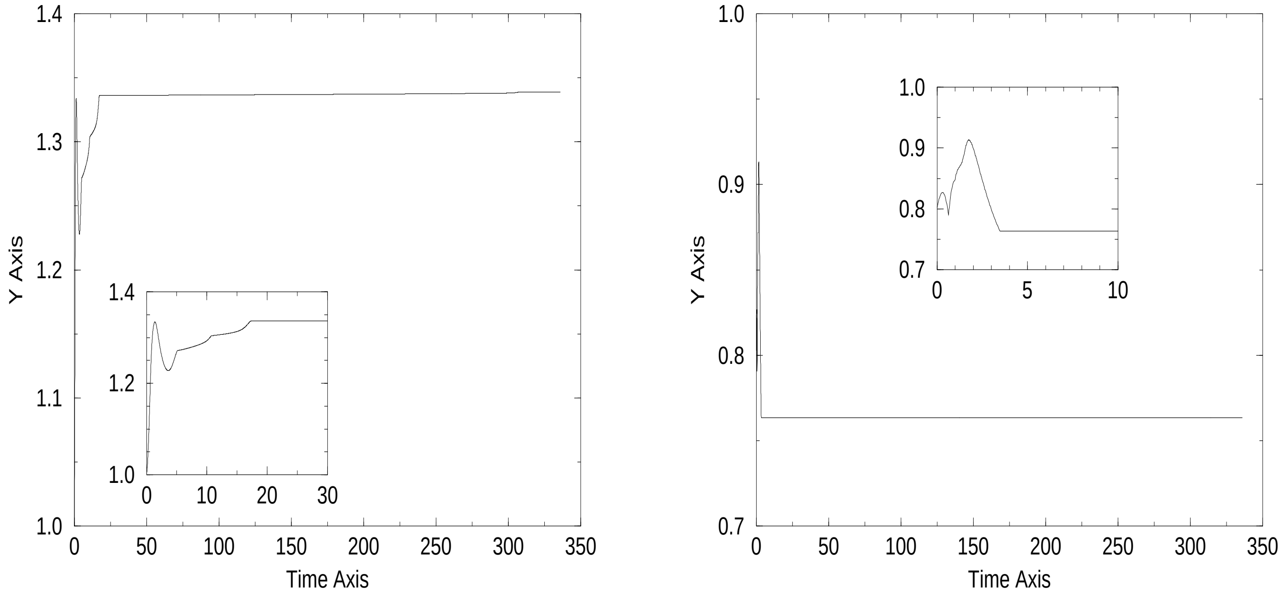


FIG. 4.12 – *Evolution en temps des normes* L^2 *(gauche) et* L^∞ *(droite) de la vitesse* u_N.



$$\begin{aligned} B(u_N, u_N) = B(y_{N_1} + z_{N_1}, y_{N_1} + z_{N_1}) &= B(y_{N_1}, y_{N_1}) + B_{int}(y_{N_1}, z_{N_1}) \\ \text{avec } B_{\text{int}}(y_{N_1}, z_{N_1}) = B(y_{N_1}, z_{N_1}) &+ B(z_{N_1}, y_{N_1}) + B(z_{N_1}, z_{N_1}) \end{aligned}$$

Le terme $P_{N_1} B(y_{N_1}, y_{N_1})$ correspond aux contributions du terme non linéaire dépendant uniquement des grandes échelles. L'allure de ces contributions dépend fortement du niveau N_1 de coupure, (figure 4.6).
Le terme $P_{N_1} B_{\text{int}}(y_{N_1}, z_{N_1})$ désigne les contributions des petites échelles dans l'équation régissant les grandes. Elles présentent une allure chahutée jusqu'à la fin de la période transitoire, soit $t = 24$, où elles deviennent constantes en norme.
Les figures (4.4, 4.5, 4.6, 4.7) nous permettent de comparer les quantités des équations régissant l'évolution des grandes échelles d'une part :

$$\frac{\partial y_{N_1}}{\partial t} - \nu \frac{\partial^2 y_{N_1}}{\partial x^2} + P_{N_1} B(y_{N_1}, y_{N_1}) + P_{N_1} B_{\text{int}}(y_{N_1}, z_{N_1}) = P_{N_1} f_N \tag{4.12}$$

et des petites échelles d'autre part :

$$\frac{\partial z_{N_1}}{\partial t} - \nu \frac{\partial^2 z_{N_1}}{\partial x^2} + Q_{N_1} B(y_{N_1}, y_{N_1}) + Q_{N_1} B_{\text{int}}(y_{N_1}, z_{N_1}) = 0 \tag{4.13}$$

Dans l'équation gouvernant les grandes échelles (4.12), le terme de couplage des grandes structures entre-elles ($P_{N_1} B(y_{N_1}, y_{N_1})$) prédomine, alors que dans l'équation régissant les petites structures (4.13), le terme de diffusion $\nu \left| \frac{\partial^2 z_{N_1}}{\partial x^2} \right|_{L^2(0,2\pi)}$ est du même ordre que le terme non linéaire de couplage $| Q_{N_1} B_{int}(y_{N_1}, z_{N_1}) |_{L^2(0,2\pi)}$.
Nous avons représenté l'allure dans l'espace physique de la solution à différents instants ainsi que celle de la force, figures (4.8) et (4.9). A partir de $t = 22$, le profil de la vitesse n'évolue plus : les grandes structures se sont formées.
Les tracés des spectres d'énergie le confirme (figure 16), les modes correspondants aux hautes fréquences ne sont pas encore stabilisés.

Simulation pour $\nu = 10^{-3}$.
Avec une viscosité $\nu = 10^{-3}$, la perturbation initiale nécessite 3072 modes pour capter convenablement la solution calculée. Suite à la perturbation initiale, pour $t \in [0, 30]$, les normes L^2 et L^∞ présentent une évolution visible puis semblent se stabiliser jusqu'à convergence. Le comportement de $| u_N |_{L^2(0,2\pi)}$, figure (4.12) est dicté par les termes de l'équation (4.10). Après de fortes amplitudes dûes à la perturbation initiale, ceux-ci adoptent une allure plus régulière et leur différence (de l'ordre de 10^{-4}) reste constante pendant la quasi-totalité du temps d'intégration.
Cela implique la très faible croissance de $| u_N |_{L^2(0,2\pi)}$. Pour la résolution sur deux niveaux $u_N = y_{N_1} + z_{N_1}$ nous avons choisi les différentes valeurs $N_1 = 8, 32, 128, 512$ pour la fréquence de coupure $N = 3072$.
Les grandes échelles $| y_{N_1} |_{L^2(0,2\pi)}$, figure (4.13) reprennent l'allure générale de la vitesse. Seul le niveau $N_1 = 8$ présente une évolution différente des autres niveaux. Toutefois, la remontée qui s'opère tout au long de l'intégration montre que l'essentiel de l'énergie se concentre peu à peu dans les premiers modes. Cela se trouve conforté par l'évolution des spectres d'énergie, figures (4.19,4.20).
L'énergie apportée par la perturbation initiale se répand sur l'ensemble du spectre, puis au cours de la phase transitoire, les petites échelles se désexcitent pour finalement converger.

FIG. 4.13 – *Evolution en temps des quantités* $|y_{N_1}|_{L^2(0,2\pi)}$ *(gauche) et* $|z_{N_1}|_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* N_1 *= 8 (1), 32 (2), 128 (3), 512 (4) avec* N *= 3072.*

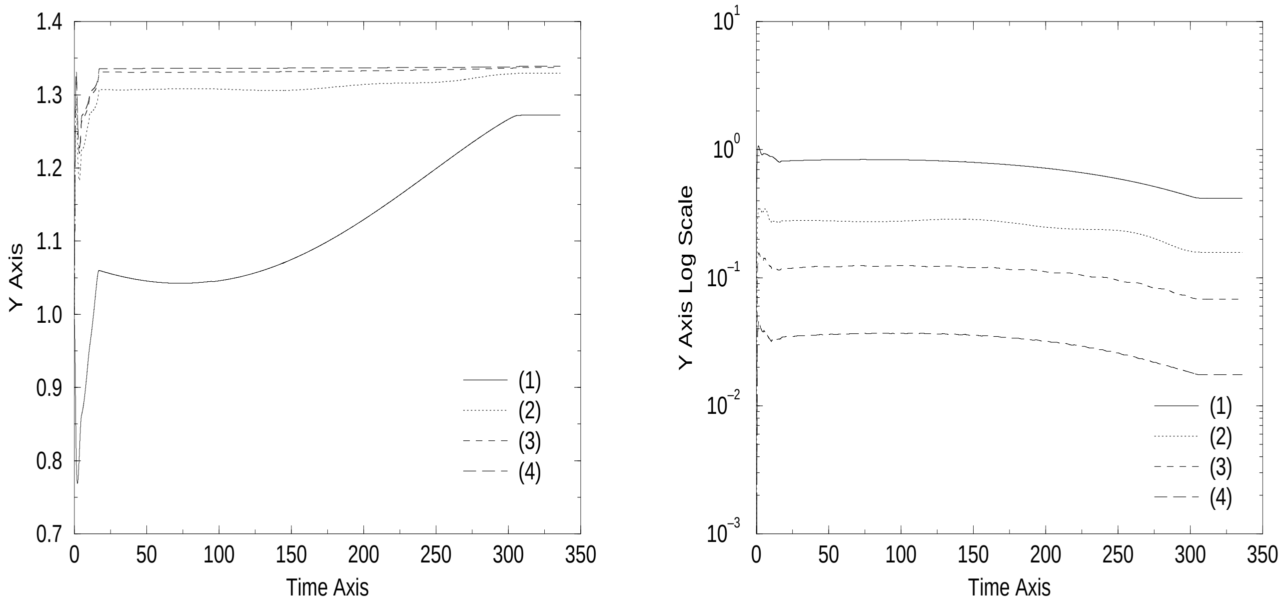


FIG. 4.14 – *Evolution en temps des quantités* $|\dot{y}_{N_1}|_{L^2(0,2\pi)}$ *(gauche) et* $|\dot{z}_{N_1}|_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* N_1 *= 8 (1), 32 (2), 128 (3), 512 (4) avec* N *= 3072.*

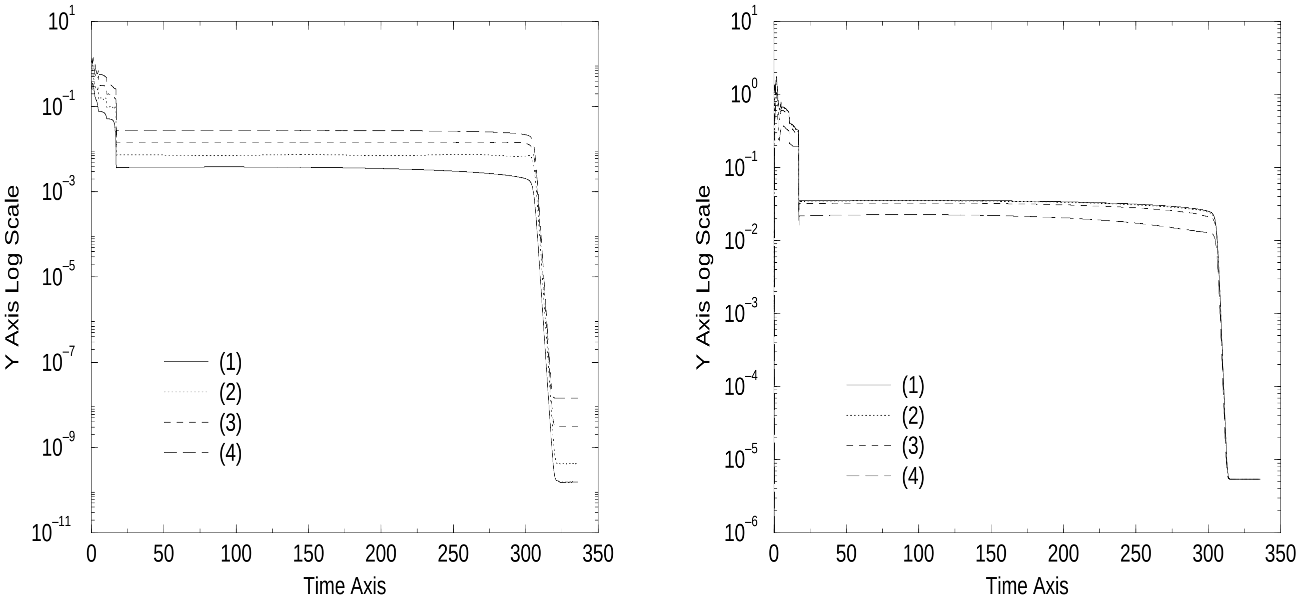


FIG. 4.15 – *Evolution en temps des normes* $\nu\left|\dfrac{\partial^2 y_{N_1}}{\partial x^2}\right|_{L^2(0,2\pi)}$ *(gauche) et* $\nu\left|\dfrac{\partial^2 z_{N_1}}{\partial x^2}\right|_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* N_1 *= 8 (1), 32 (2), 128 (3), 512 (4) avec* N *= 3072.*

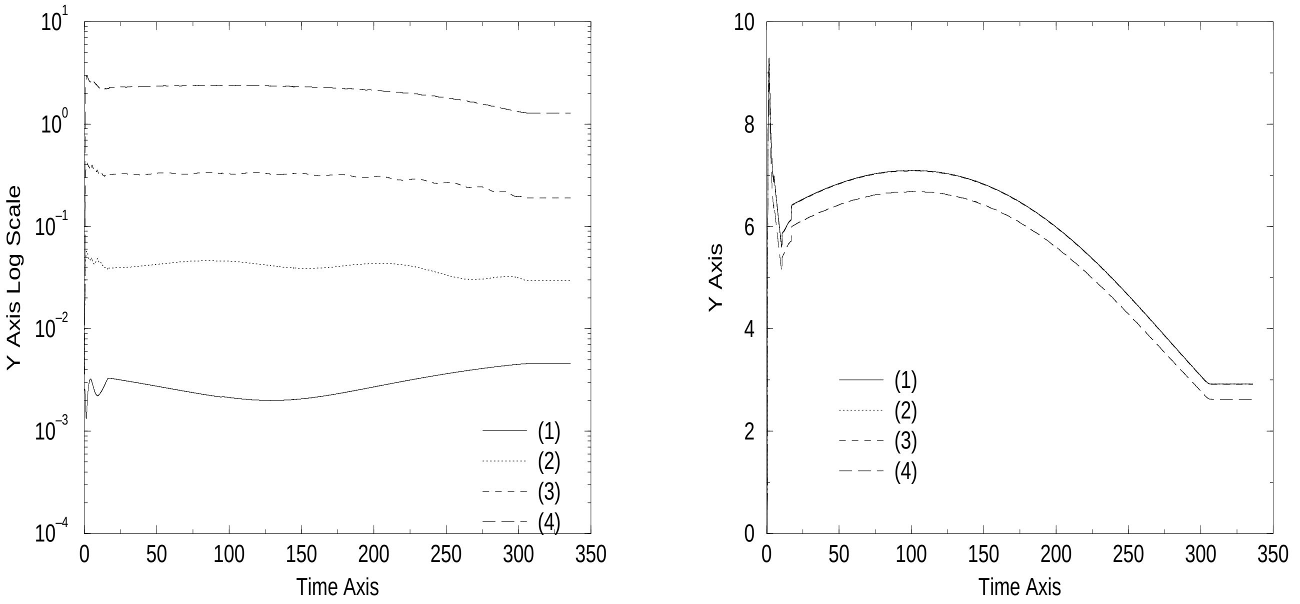


FIG. 4.16 – *Evolution en temps des quantités* $|\, P_{N_1} B(y_{N_1}, y_{N_1})\,|_{L^2(0,2\pi)}$ *(gauche) et* $|\, P_{N_1} B_{int}(y_{N_1}, z_{N_1})\,|_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* N_1 *= 8 (1), 32 (2), 128 (3), 512 (4) avec* N *= 3072.*

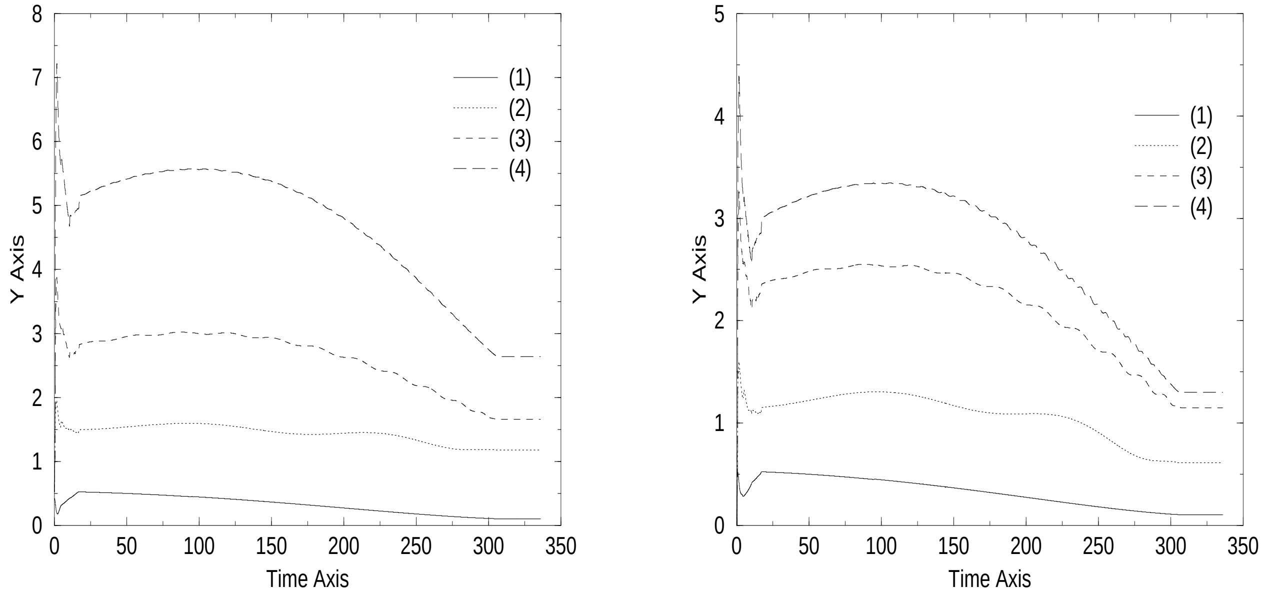


FIG. 4.17 – *Evolution en temps des quantités* $\mid Q_{N_1} B(y_{N_1}, y_{N_1}) \mid_{L^\infty(0,2\pi)}$ *(gauche) et* $|Q_{N_1} B_{int}(y_{N_1}, z_{N_1})|_{L^\infty(0,2\pi)}$ *(droite) pour différentes valeurs de* N_1 = *8 (1), 32 (2), 128 (3), 512 (4) avec* N = *3072.*

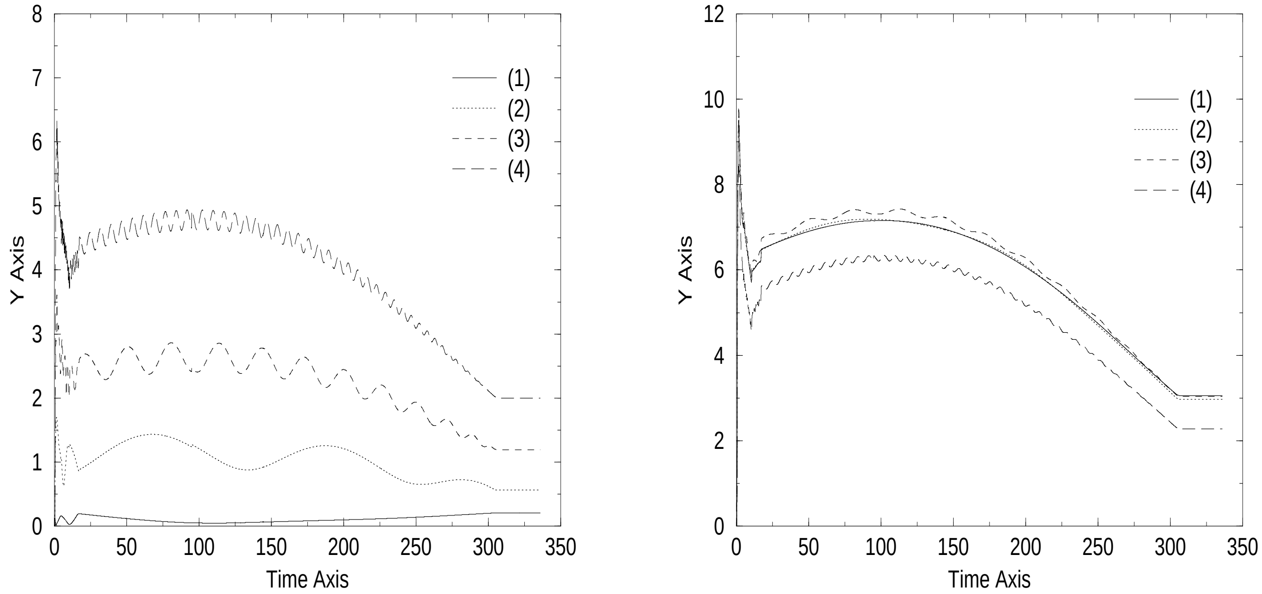
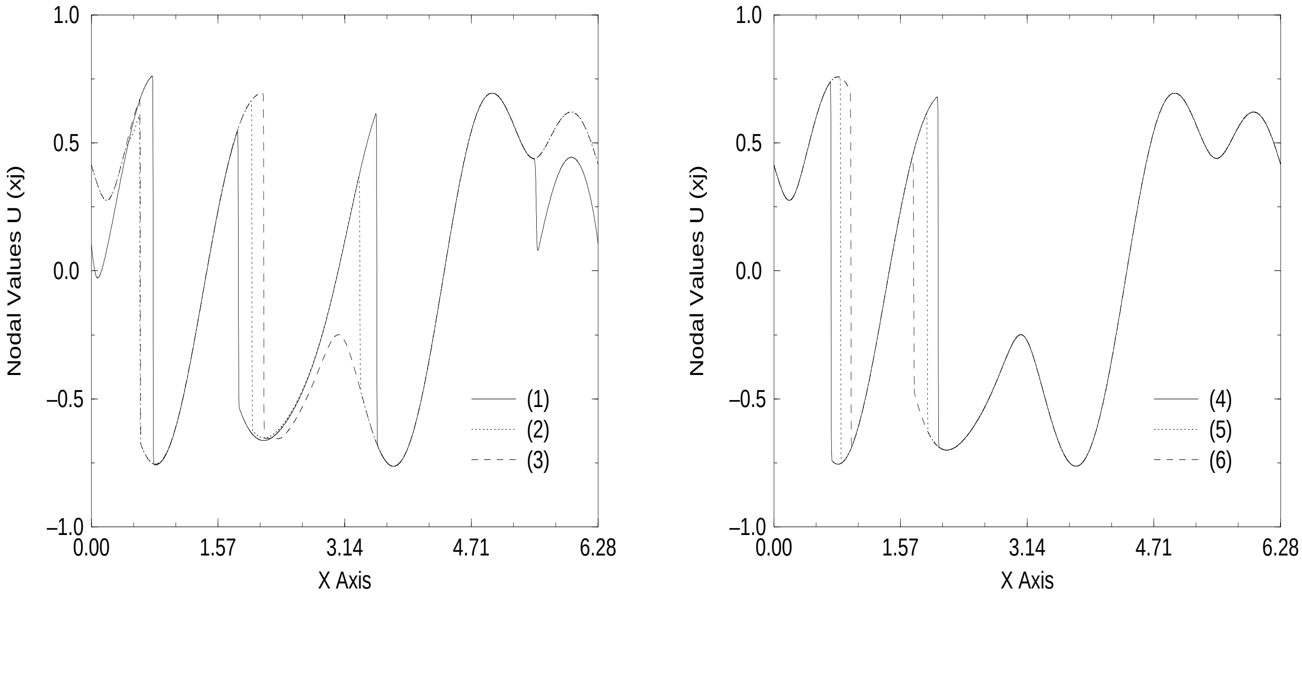


FIG. 4.18 – *Valeurs instantannées de la vitesse à différents temps :* t = *3.5 (1), 10.5 (2), 17 (3), 80 (4), 160 (5), 240 (6).*



La figure (4.13) nous montre aussi l'importance relative des petites échelles $|\, z_{N_1} \,|_{L^2(0,2\pi)}$ qui gardent un ordre de grandeur constant. On peut faire la même remarque pour les dérivées temporelles, figure (4.14) qui, après la perturbation initiale, présentent un régime d'amplitude constante pendant une très longue période. C'est seulement durant les dernières unités de temps précédent la convergence de la solution calculée, qu'elles accusent une brutale décroissance.
La figure (4.16) indique clairement que la dissipation a lieu dans les petites échelles.
Dans l'équation gouvernant les grandes échelles (4.12), le terme de couplage des grandes structures entre-elles $(P_{N_1}B(y_{N_1}, y_{N_1}))$ est presque rattrapé par le terme d'interaction $|\, Q_{N_1}B_{int}(y_{N_1}, z_{N_1}) \,|_{L^2(0,2\pi)}$, alors que dans l'équation régissant les petites structures (4.13), le terme de diffusion $\nu \left| \dfrac{\partial^2 z_{N_1}}{\partial x^2} \right|_{L^2(0,2\pi)}$ est du même ordre que le terme non linéaire de couplage $|\, Q_{N_1}B_{int}(y_{N_1}, z_{N_1}) \,|_{L^2(0,2\pi)}$.
Dans l'espace physique, figures (4.18, 4.19), la solution présente de forts gradients qui nécessitent beaucoup de points pour les capter convenablement. Ceux-ci disparaissant au cours du temps, sauf au voisinage de $x \approx 1$.

Comparaison.
Nous avons réalisé deux simulations numériques dans le cas périodique avec une force déterministe n'agissant que sur quelques modes de basse fréquence; avec une viscosité de 10^{-2} pour l'une et 10^{-3} pour l'autre.
Nous avons adapté la fréquence de coupure pour capter convenablement la solution calculée ainsi que le pas de temps pour avoir la stabilité et la convergence.
Tout d'abord le rapport de 10 présent entre les deux valeurs de la viscosité se retrouve dans le temps nécessaire pour obtenir la convergence.
Ensuite, pour les deux simulations, nous observons le même phénomène sur l'intervalle d'intégration $[0, 20]$, à savoir une augmentation de la norme de la vitesse par étapes, chacune étant liée au comportement des quantités de l'équation (4.10). Jusqu'à la convergence de la solution, ces deux dernières restent très proches et présentent une évolution très lente et régulière.
il est à noter que dans les deux cas, les solutions convergées ont de forts gradients.

FIG. 4.19 – *Valeurs instantannées de la vitesse à t=300 (7), 336 (8) et de la force extérieure (9) (gauche), spectre d'énergie à t = 5 (droite).*

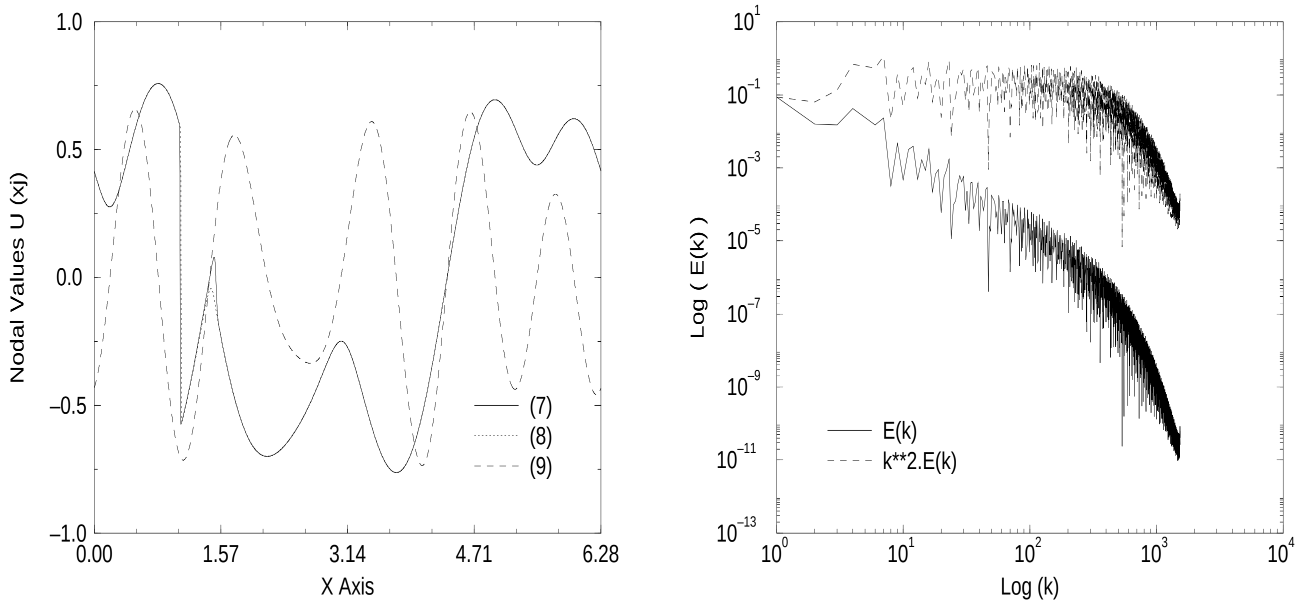


FIG. 4.20 – *Spectres d'énergie à t = 160 et 272.*

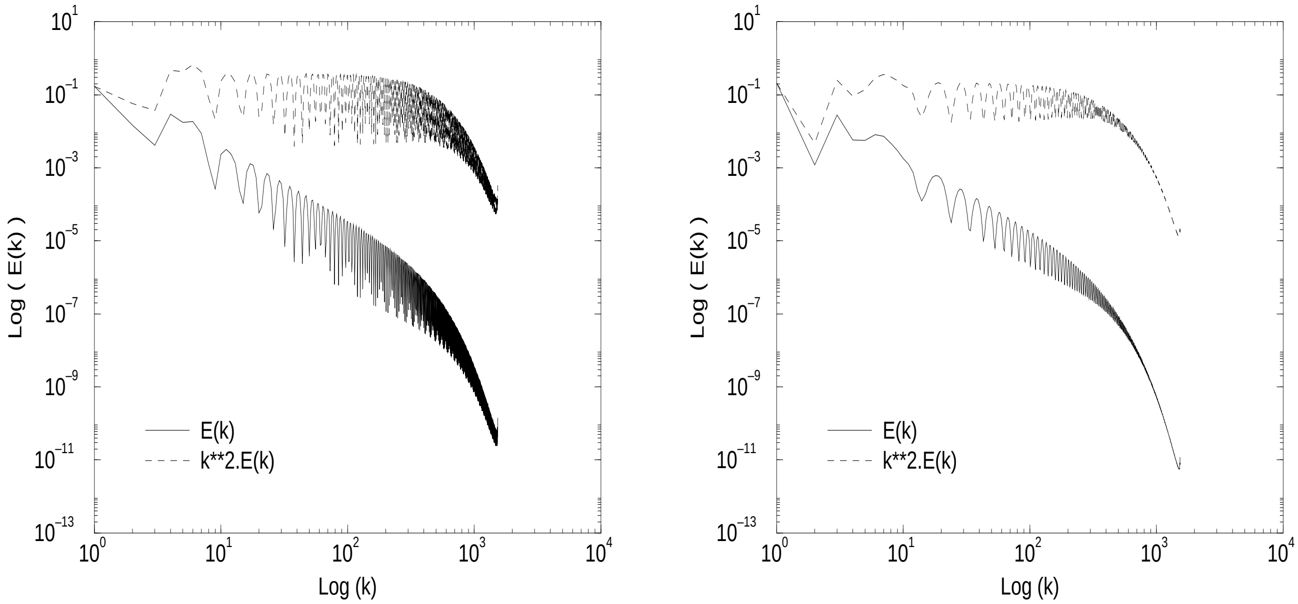


FIG. 4.21 – *Evolution en temps de l'erreur pour* $\Delta t = 10^{-2}$ *(1),* $\Delta t = 10.10^{-2}$ *(2) (gauche) et pour* $\Delta t = 13.10^{-2}$ *(3),* $\Delta t = 14.10^{-2}$ *(4) (droite).*

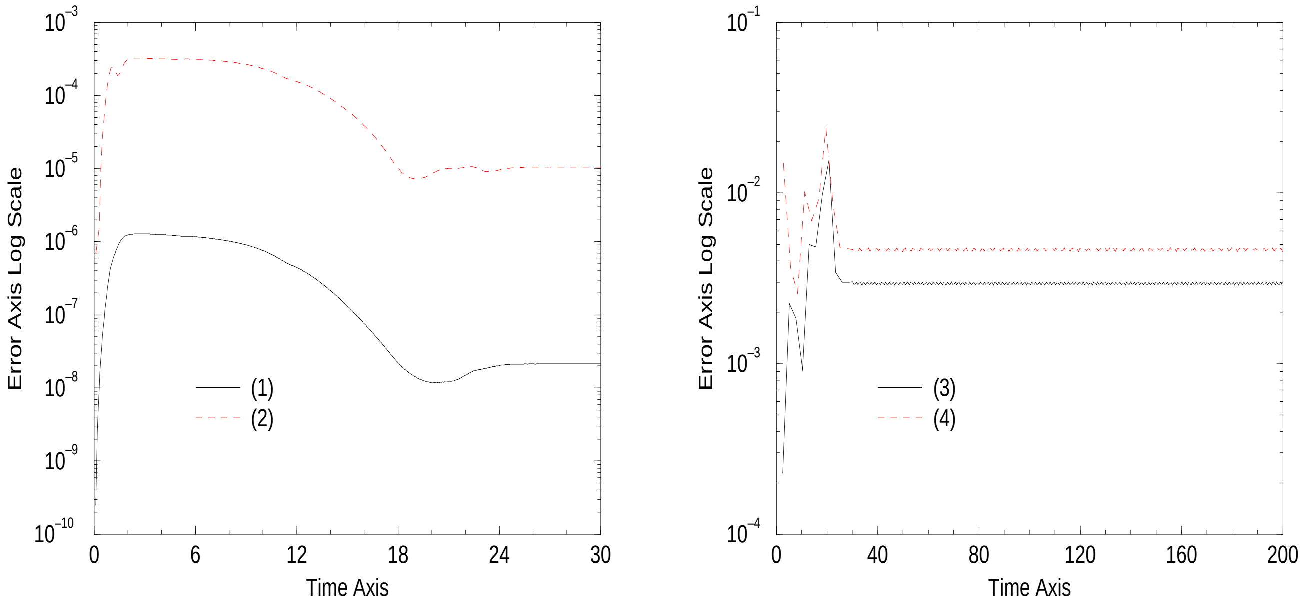
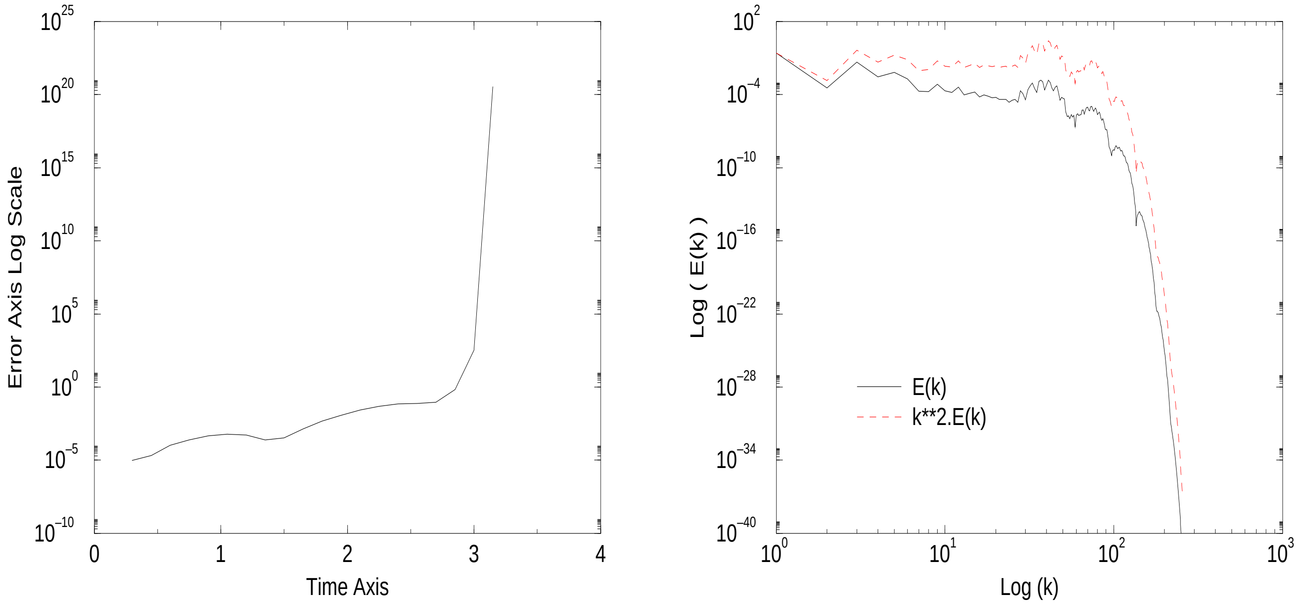


FIG. 4.22 – *Evolution en temps de l'erreur pour* $\Delta t = 15.10^{-2}$ *(gauche) et spectre d'énergie pour* $\Delta t = 13.10^{-2}$ *à* $t = 26$ *(droite).*



4.2.2.5 Test de stabilité numérique.

A partir de la simulation effectuée pour $\nu = 10^{-2}$, nous avons décidé d'étudier le mécanisme explosif qui survient lorsque l'on augmente Δt à N fixé.
Pour cela, nous avons pris comme référence le calcul effectué avec $\Delta t = 10^{-3}$ qui confère la stabilité et une bonne précision. Puis nous avons testé différents pas de temps en comparant les valeurs obtenues avec celles de référence pour les normes.
Nous considérons deux situations :
- départ à $t = 0$ avec la solution initiale (test 1),
- départ à $t = 30$ avec la solution convergée (test 2).

Test 1.
Nous avons effectué les calculs avec les pas de temps suivants :
$\Delta t = 10^{-2},\ 10.10^{-2},\ 13.10^{-2},\ 14.10^{-2},\ 15.10^{-2}$, figures (4.21,4.22,4.23).
Les résultats sont présentés dans le tableau suivant. Nous dirons qu'un calcul est stable lorsqu'il n'explose pas (i.e. la norme de la solution reste bornée), de même il sera convergent si la norme n'oscille pas autour d'une valeur, mais converge.

TAB. 4.1 – *Comparaison des résultats pour le test 1.*

Δt	$N\Delta t\ \lvert u_N \rvert_{L^\infty(0,2\pi)}$	stable	convergent	"erreur "
10^{-3}	$\approx 0,435$	oui	oui	référence
10^{-2}	$\approx 4,35$	oui	oui	$\approx 2.10^{-8}$
10.10^{-2}	$\approx 43,5$	oui	oui	$\approx 10^{-5}$
13.10^{-2}	≈ 90	oui	non	oscillante $\approx 3.10^{-3}$
14.10^{-2}	≈ 100	oui	non	oscillante $\approx 3.10^{-3}$
15.10^{-2}	—	non	non	explose à $t = 3,30$

Le tableau (4.1) montre que le nombre de Courant peut être pris 100 fois plus grand sans que la simulation en pâtisse. Bien entendu, la précision des résultats est moindre.

Test 2.
Les calculs ont été effectués avec les mêmes valeurs pour Δt que précédemment :
$\Delta t = 10^{-2},\ 10.10^{-2},\ 13.10^{-2},\ 14.10^{-2},\ 15.10^{-2}$ (figures 30, 31, 32).
Nous partons à $t = 30$ avec la solution convergée obtenue pour $\Delta t = 10^{-3}$.

FIG. 4.23 – *Spectres d'énergie pour* $\Delta t = 13.10^{-2}$ *au temps* $t = 52$ *(gauche) et 182 (droite).*

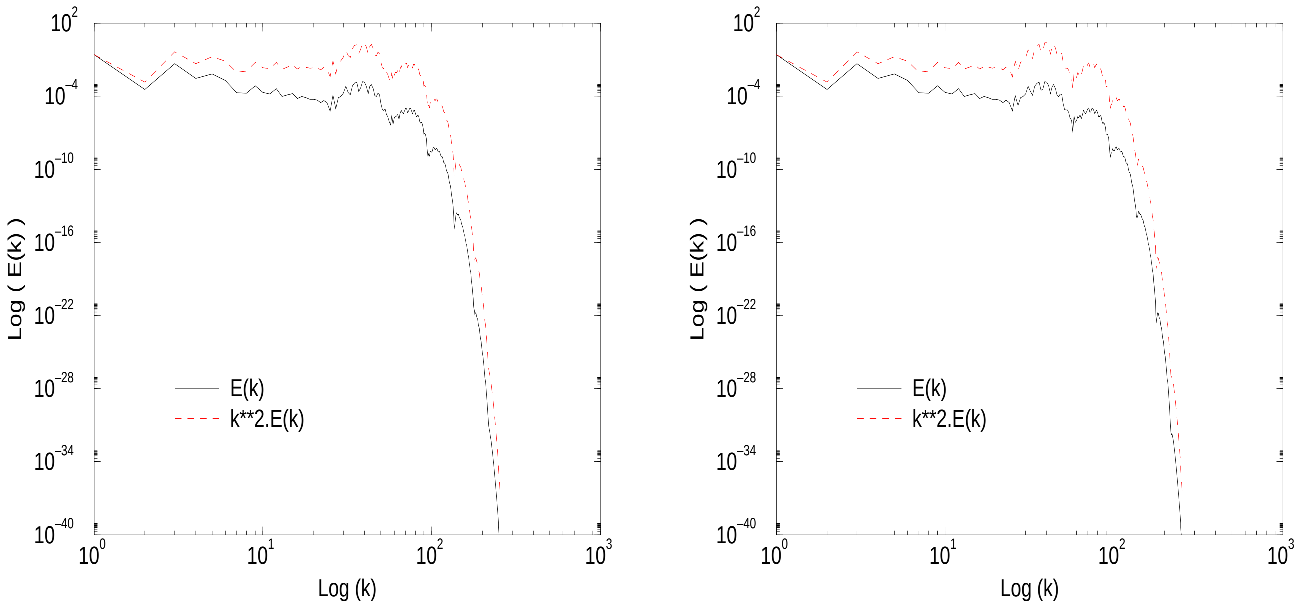


FIG. 4.24 – *Evolution en temps de l'erreur pour* $\Delta t = 10^{-2}$ *(1),* $\Delta t = 10.10^{-2}$ *(2) (gauche) et pour* $\Delta t = 13.10^{-2}$ *(droite).*

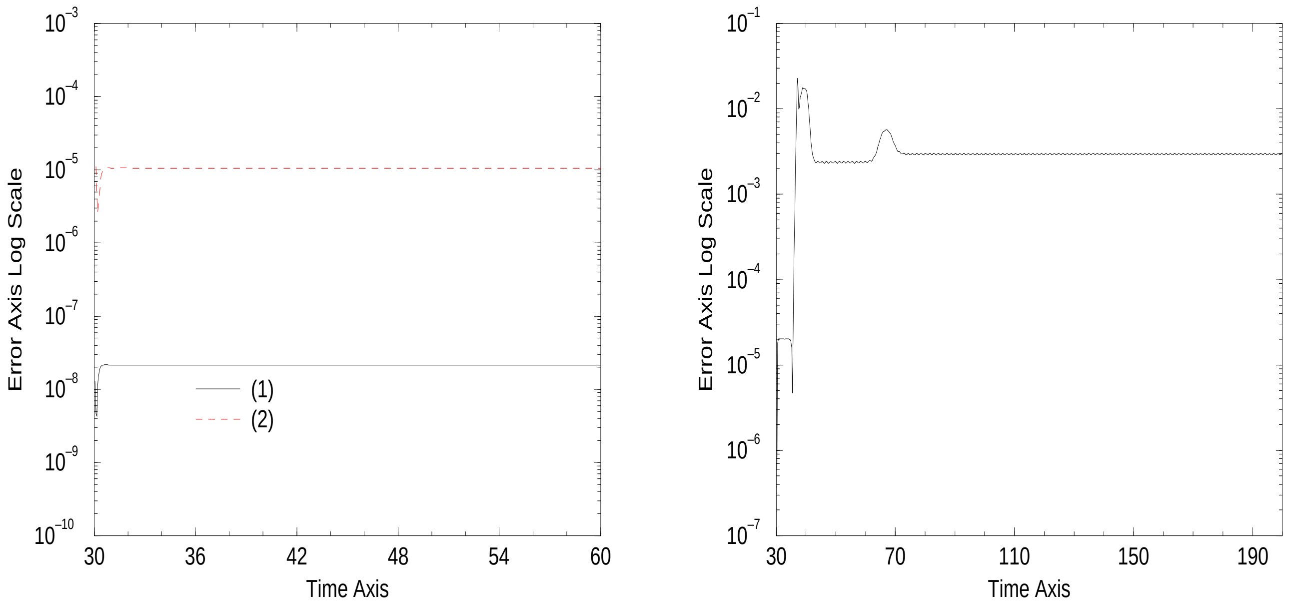


FIG. 4.25 – *Evolution en temps de l'erreur pour* $\Delta t = 14.10^{-2}$ *(gauche),* $\Delta t = 15.10^{-2}$ *(droite).*

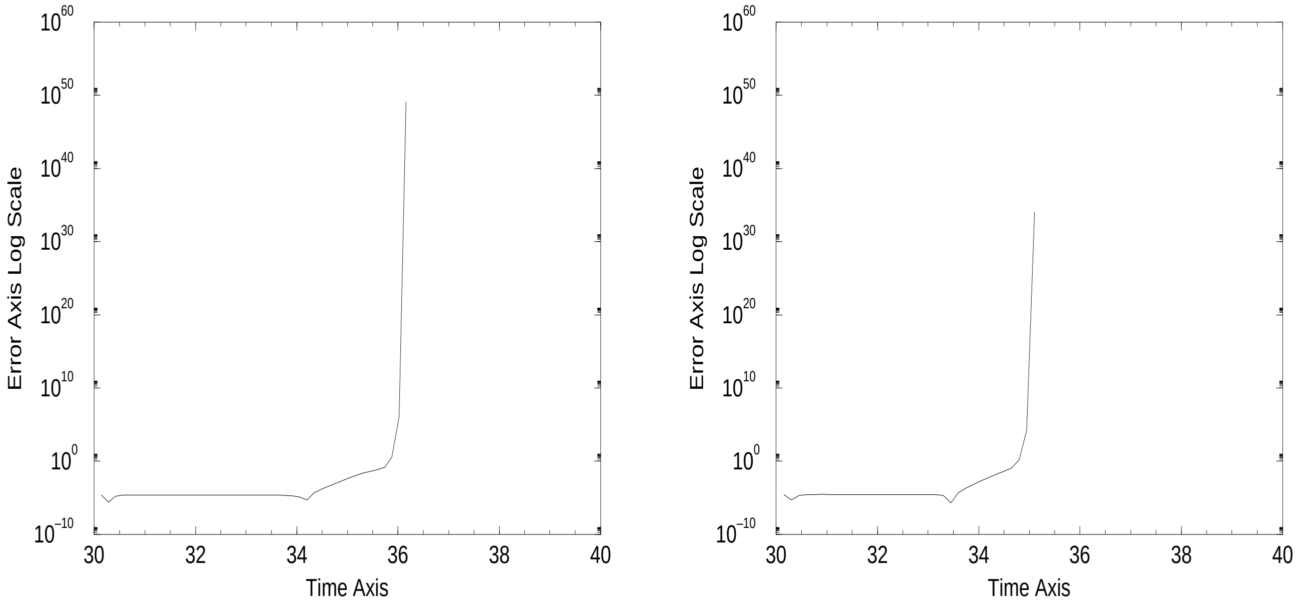
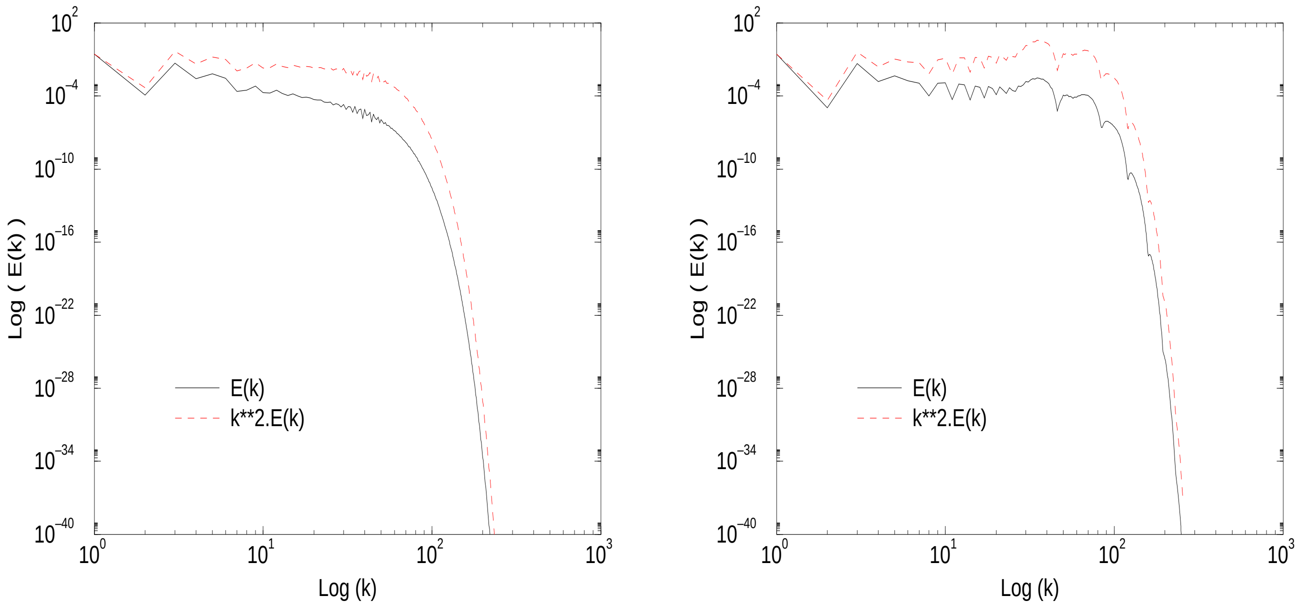


FIG. 4.26 – *Spectres d'énergie pour* $\Delta t = 14.10^{-2}$ *au temps* t = *34,2 (gauche) et* t = *35,6 (droite).*



TAB. 4.2 – *Comparaison des résultats pour le test 2.*

Δt	$N\Delta t\ \lvert u_N \rvert_{L^\infty(0,2\pi)}$	stable	convergent	"erreur "
10^{-3}	$\approx 0,435$	oui	oui	référence
10^{-2}	$\approx 4,35$	oui	oui	$\approx 2.10^{-8}$
10.10^{-2}	$\approx 43,5$	oui	oui	$\approx 10^{-5}$
13.10^{-2}	≈ 90	oui	non	oscillante $\approx 3.10^{-3}$
14.10^{-2}	—	non	non	explose à $t = 36,30$
15.10^{-2}	—	non	non	explose à $t = 35,10$

Analyse des tests.
Les différentes valeurs pour le pas de temps mettent en évidence l'existence d'une valeur critique Δt^* se situant vers 14.10^{-2} pour cette simulation. Pour des pas de temps plus petits ($\Delta t = 10^{-3}, 10^{-2}, 10.10^{-2}$) malgré un nombre de Courant très grand (0,43 pour $\Delta t = 10^{-3}$ d'où 43 pour $\Delta t = 10.10^{-2}$) nous observons numériquement la convergence du schéma vers la solution.
Cette étude nous amène à faire deux remarques :
- tout d'abord, le fait d'augmenter le pas de temps amplifie l'erreur d'intégration commise sur le terme non linéaire (la partie linéaire est intégrée exactement). Celle-ci étant d'ordre 3 par rapport au pas de temps (nous utilisons le schéma explicite d'ordre 3 de type Runge-Kutta), ce facteur théorique est confirmé numériquement;
- ensuite, le temps nécessaire pour converger est sensiblement le même, bien que cela se fasse de moins en moins facilement.

Le test avec le pas de temps $\Delta t = 13.10^{-2}$ est très intéressant, car il montre le processus de destabilisation sans que celui-ci ait lieu : le schéma reste stable mais n'est pas convergent. La solution présente même un comportement oscillant en temps. Cela est visible sur les différents tracés : les spectres présentent une accumulation d'énergie à la limite des zones inertielles et dissipatives. Pour un pas de temps $\Delta t = 13.10^{-2}$ celle-ci se maintient, un apport d'énergie supplémentaire est dissipé, alors que pour des pas de temps plus grands, on assiste à une accumulation d'énergie au même endroit dans le spectre sans que la dissipation soit suffisante. Plus le pas de temps est grand, plus rapide est l'explosion qui s'en suit.

4.3 Résolution de l'équation de Burgers stochastique.

Cette section est consacrée à la description d'un écoulement turbulent, i.e. l'équation de Burgers perturbée par une force aléatoire.
Après avoir précisé la nature de cette force, nous donnerons un théorème d'existence et d'unicité de solutions pour ensuite présenter les résultats obtenus.

4.3.1 Force aléatoire - bruit blanc.

Nous commençons par définir le mouvement brownien.

4.3.1.1 Mouvement brownien.

Soit E un espace de Banach séparable muni de la tribu des boréliens $\mathcal{E}$ et $(\Omega,\, \mathcal{F},\, P)$ un espace probabilisé.

Définition *2*

- $(\, X(t)\,)_{t\in I}$ est un processus stochastique si et seulement si
$\forall t \in I$, $X(t)(\omega) : (\Omega, \mathcal{F}) \longrightarrow (E, \mathcal{E})$ est une variable aléatoire
(i.e. à t fixé, $X(t)$ est mesurable par rapport à ω)
I quelconque en général, en pratique $I = [0,\, T]$ ou $\mathbb{R}_+$.
- Pour ω fixé, l'application $t \longmapsto X(t)(w)$ s'appelle une trajectoire.

On se place dans $(\Omega,\, \mathcal{F},\, P)$.

Définition *3*

Un processus stochastique $(\, \beta(t)\,)_{t\in\mathbb{R}_+}$ est un mouvement brownien (ou processus de Wiener cylindrique) si et seulement si
- $\beta(t)$ est à valeurs réelles, $\beta(0) = 0$;
- $(\, \beta(t)\,)_{t\in\mathbb{R}_+}$ est un processus stochastique ;
- β a des trajectoires continues p.p. ;
- si $0 \leqslant t \leqslant s$ fixés alors $\beta(s) - \beta(t) \sim \mathcal{N}(0,\, s-t)$,
i.e. $(\beta(s) - \beta(t)$ est une variable aléatoire de loi normale, moyenne nulle et variance égale à $s-t)$;
- β est à incréments indépendants,
i.e. $0 = t_0 \leqslant t_1 \leqslant \ldots \leqslant t_n$ alors les variables aléatoires $\beta(t_{i+1}) - \beta(t_i)$ sont indépendantes.

Remarque *13*

Cette définition est naturelle, le mouvement brownien décrit le mouvement d'une particule soumise à des chocs aléatoires
$\beta(t)$ indique la position de la particule à l'instant t. La probabilité d'être déplacée vers la gauche est la même que celle d'être déplacée vers la droite, soit $1/2$.
Beaucoup de chocs successifs, indépendants les uns des autres entrainent une loi normale par le théorème Central-Limite.
Les chocs indépendants impliquent des positions indépendantes donc des incréments (amplitude du déplacement) indépendants.
Le modèle mathématique possède, entre autres, comme propriétés :

Propriété *1*

- $E\,(\,\beta(t)\beta(s)\,) = \min(t,s) = (t \wedge s)$
- *$\beta(t)$ n'est pas dérivable sur un ensemble de mesure 1.*

4.3.1.2 Bruit blanc.

On définit un bruit blanc comme la dérivée (en temps) d'un mouvement brownien $(\,\beta(t)\,)_t$ mais pas au sens usuel des dérivées, plutôt sous la forme intégrale

$$\beta(t) = \int_0^t W(s)\,\mathrm{d}s$$

Ce processus est appelé bruit blanc, car à l'instar de la lumière blanche, son spectre est uniformément excité, il n'y a pas de fréquence privilégiée.

4.3.2 Résultats d'existence et d'unicité.

Nous énonçons les résultats d'existence et d'unicité de solutions pour l'équation de Burgers stochastique perturbée par un bruit blanc en espace et en temps (drap brownien). Une étude complète peut être trouvée dans ([18]).
L'équation de Burgers s'écrit

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{1}{2}\frac{\partial (u^2)}{\partial x} + \frac{\partial^2 \tilde{W}}{\partial t \partial x} \tag{4.14}$$

Nous rappelons que $\tilde{W}(t,x)$, $t \geqslant 0$ et $x \in \mathbb{R}$ est un processus gaussien à moyenne nulle dont la fonction de covariance est donnée par

$$E\left[\,\tilde{W}(t,x)\,\tilde{W}(s,y)\right] = (t \wedge s)\,(x \wedge y)\ ;\ t,s \geqslant 0 \text{ et } x,y \in \mathbb{R}$$

Nous pouvons considérer un processus de Wiener cylindrique W en posant

$$W(t) = \frac{\partial \tilde{W}}{\partial x} = \sum_{h=1}^{\infty} \beta_h\, e_h \tag{4.15}$$

où $\{e_h\}_h$ est une base orthogonale de $L^2(0,1)$ et $\{\beta_h\}_h$ est une suite de mouvements browniens réels mutuellement indépendants dans un espace probabilisé fixé $(\Omega,\ \mathcal{F},\ P)$ adaptés à un filtrage $\{\mathcal{F}_t\}_{t\geqslant 0}$. La série (4.15) définissant W ne converge pas dans $L^2(0,1)$ mais est convergente dans n'importe quel espace U tel que l'injection $L^2(0,1) \hookrightarrow U$ soit Hilbert-Schmidt.
Dans la suite, nous écrirons (4.14) sous la forme

$$\mathrm{d}u = \left\{ \frac{\partial^2 u}{\partial x^2} + \frac{1}{2}\frac{\partial(u^2)}{\partial x} \right\} \mathrm{d}t + \mathrm{d}W \ , \quad x \in [0,\ 1]\,,\ t > 0 \tag{4.16}$$

où W est défini par (4.15).
L'équation (4.16) est complétée avec des conditions limites de type Dirichlet homogène

$$u(0,\ t) = u(1,\ t) = 0 \tag{4.17}$$

et la condition initiale

$$u(x,\ 0) = u_0(x)\,,\ x \in [0,\ 1] \tag{4.18}$$

Nous allons maintenant énoncer le résultat d'existence et d'unicité.

Théorème *4.11*

Soit u_0 donné qui soit $\mathcal{F}_0$-mesurable et tel que pour un $p \geqslant 0$, $u_0 \in L^p(0,1)$ p.s. . Alors il existe une unique solution douce de l'équation (4.16) - (4.18) qui appartienne à $C([0,\ T];\ L^p(0,1))$.

Remarque *14*

Des résultats similaires peuvent être obtenus dans le cas d'écoulements périodiques en espace en prenant un écoulement à moyenne nulle.
De plus, il a été montré qu'il existe une unique mesure invariante, ergotique, pour le problème qui nous intéresse. Cela a pour conséquence la convergence des moyennes en temps lorsque $t \longrightarrow +\infty$.

4.3.3 Simulation numérique avec des conditions aux limites périodiques.

Dans cette simulation, nous considérons un écoulement entrainé par un bruit blanc

$$\mathrm{d}u = \left\{ \frac{\partial^2 u}{\partial x^2} + \frac{1}{2}\frac{\partial(u^2)}{\partial x} \right\} \mathrm{d}t + \mathrm{d}W$$

où W est défini par (4.15), dans le cas où la solution est périodique en espace.

FIG. 4.27 – *Moyenne temporelle du nombre de Courant (gauche) et des normes* $|u_N|_{L^2(0,2\pi)}$ *(1),* $|u_N|_{L^\infty(0,2\pi)}$ *(2) de la vitesse (droite).*

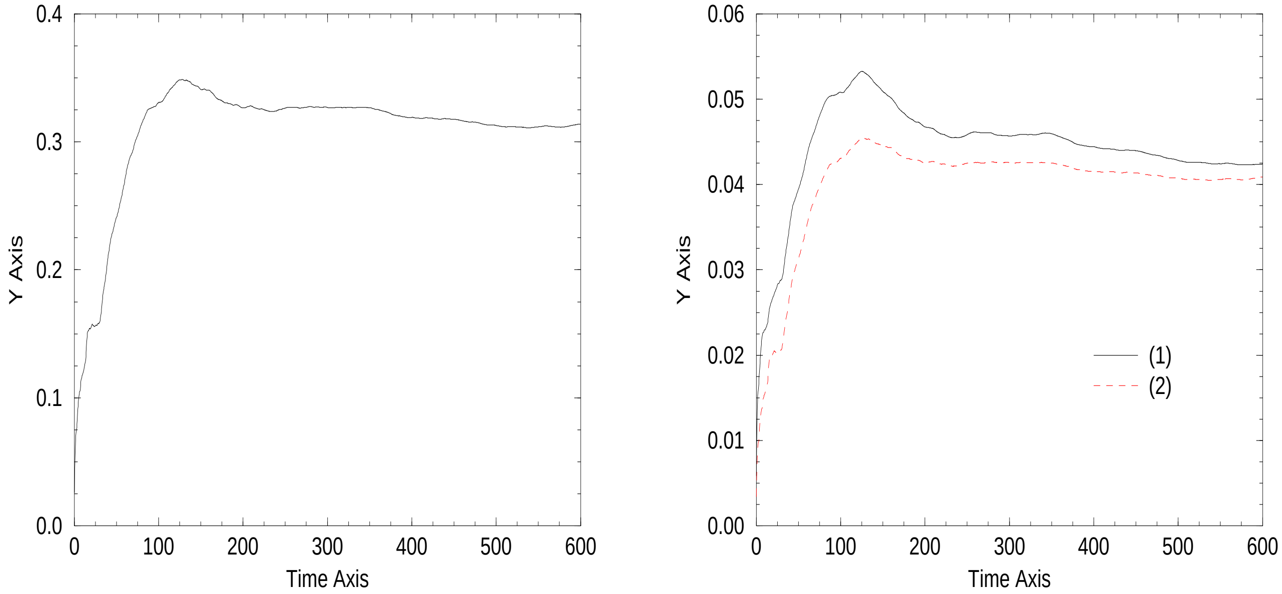


FIG. 4.28 – *Moyenne temporelle des quantités* $|y_{N_1}|_{L^2(0,2\pi)}$ *(gauche) et* $|z_{N_1}|_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* $N_1 = 8$ *(1), 32 (2), 128 (3), 512 (4) avec* $N = 2560$.

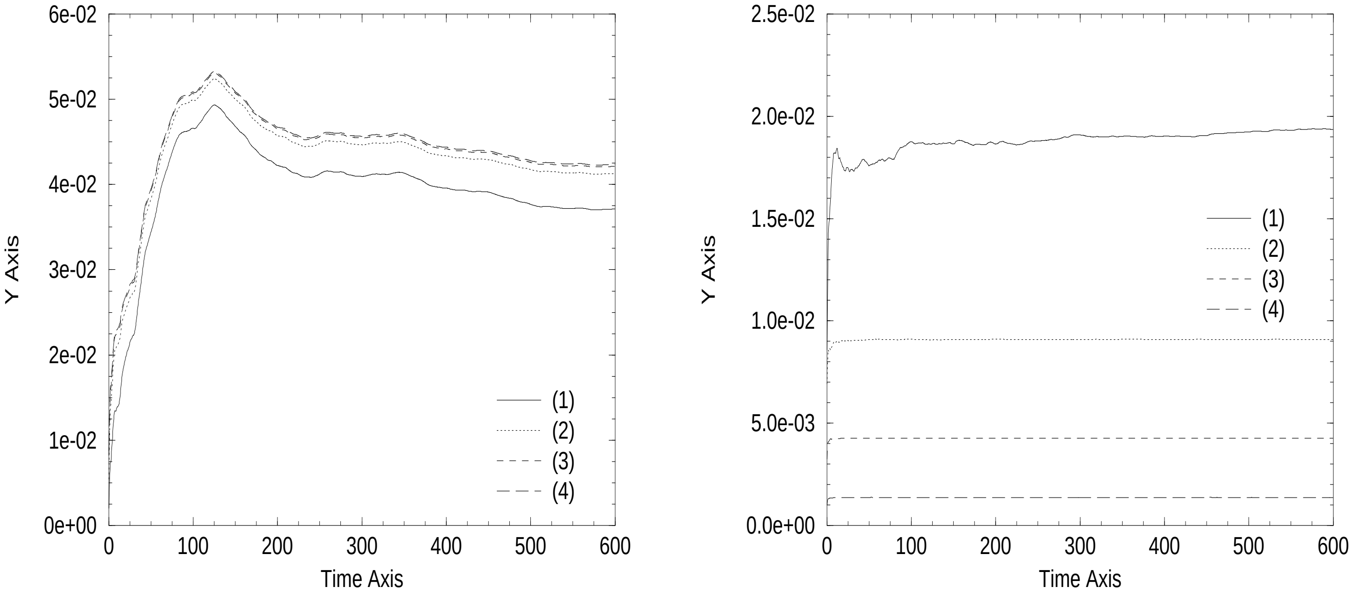


FIG. 4.29 – *Moyenne temporelle de* $\nu\left|\dfrac{\partial^2 y_{N_1}}{\partial x^2}\right|_{L^2(0,2\pi)}$ *(gauche) et* $\nu\left|\dfrac{\partial^2 z_{N_1}}{\partial x^2}\right|_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* N_1 *= 8 (1), 32 (2), 128 (3), 512 (4) avec* N *= 2560.*

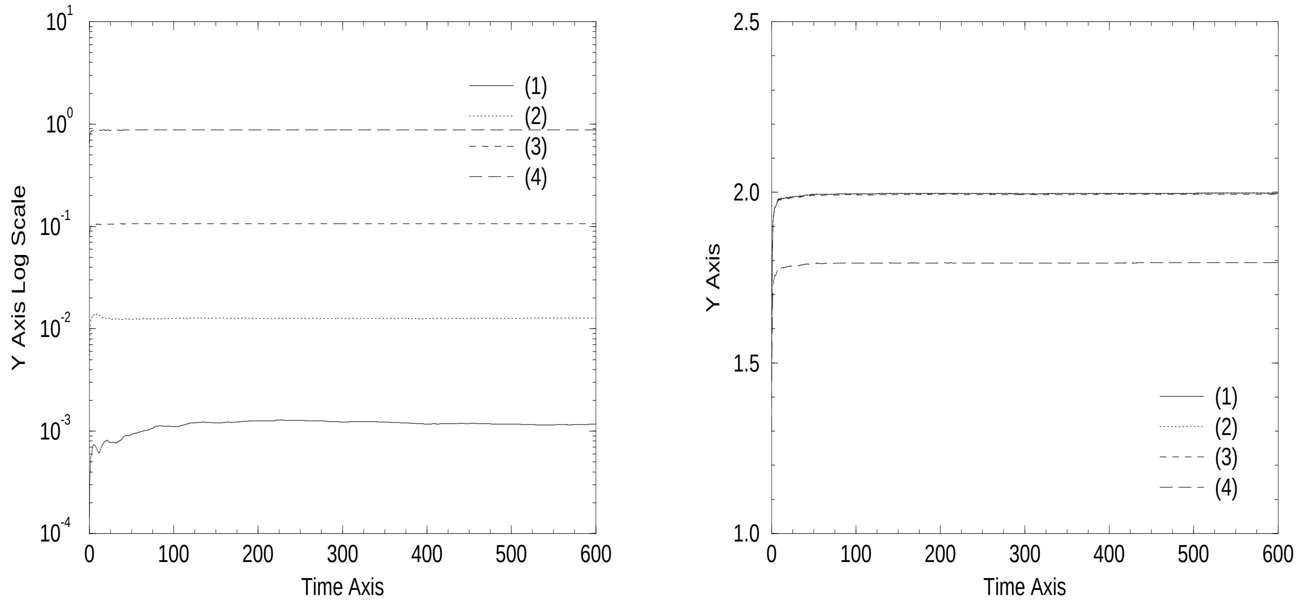
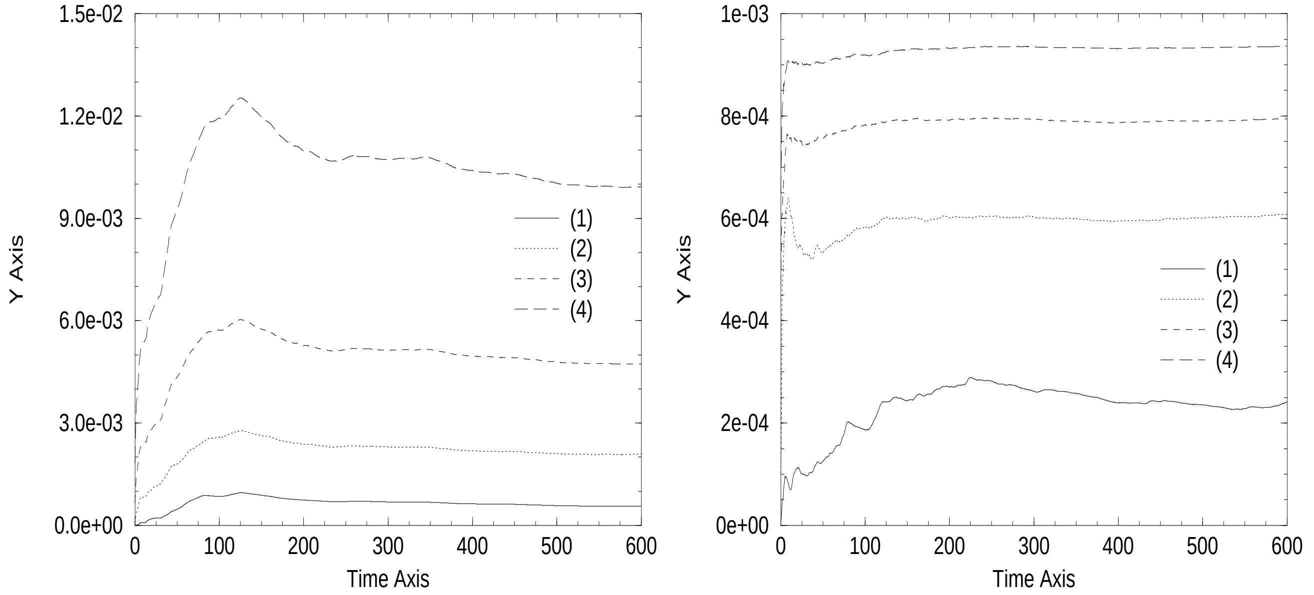


FIG. 4.30 – *Moyenne temporelle des quantités* $|\ P_{N_1}B(y_{N_1}, y_{N_1})\ |_{L^2(0,2\pi)}$ *(gauche) et* $|Q_{N_1}B(y_{N_1}, y_{N_1})|_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* N_1 *= 8 (1), 32 (2), 128 (3), 512 (4) avec* N *= 2560.*



4.3.3.1 Cadre de l'étude.

4.3.3.1.1 Description de la condition initiale.

On part de la condition initiale donnée sur le spectre de u.
On choisit $u_0 \equiv 0$, ce qui revient à

$$\hat{u}_k(0) = 0, \ \forall k \in \mathbb{I}_N = \left[\!\left[1 - \frac{N}{2}, \frac{N}{2} \right]\!\right]$$

Cela permet d'être immédiatement sous la totale influence du bruit blanc sans avoir une phase de transition.

4.3.3.1.2 Description de la force extérieure.

On construit la force extérieure comme un bruit blanc. Cela signifie que $W(t)$ s'écrit sous la forme

$$W(t) = \sum_{k \in \mathbb{Z}} \beta_k(t) \, \Phi_k$$

avec

$(\Phi_k)_k$	la base des exponentielles de Fourier,
$(\beta_k)_k$	une suite de mouvements browniens réels indépendants dans un espace probabilisé $(\Omega, \mathcal{F}, P)$.

Partant de (4.16), on applique la méthode de Galerkin et on obtient alors

$$\mathrm{d}u_N - \nu \frac{\partial^2 u_N}{\partial x^2} \mathrm{d}t = -P_N \left\{ u_N \frac{\partial u_N}{\partial x} \right\} \mathrm{d}t + \mathrm{d}W_N \tag{4.19}$$

avec

$$u_N = \sum_{k \in \mathbb{I}_N} \hat{u}_k(t) \, e^{ikx}, \quad W_N = \sum_{k \in \mathbb{I}_N} \beta_k(t) \, e^{ikx}$$

4.3.3.1.3 Discrétisation en temps.

Nous utilisons un schéma de Runge-Kutta d'ordre 3 explicite en temps pour discrétiser (4.19).
Chaque étape se décompose en 3 sous-étapes de pas de temps respectif Δt_i $i = 1, 2, 3$ avec $\sum_{i=1}^{3} \Delta t_i = \Delta t$, pour Δt le pas de temps.
On est donc ramené à considérer l'intégration de (4.19) sur l'intervalle $[t, \, t + \tau]$ avec $\tau = \Delta t_i, \ i = 1, 2, 3$.
Puisque $\{\Phi_k\}_{k \in \mathbb{I}_N}$ forme une base de $\mathcal{S}_N$, l'équation (4.19) s'écrit sous la forme du système suivant

$$\mathrm{d}\hat{u}_k(t) + \nu k^2 \hat{u}_k(t) \mathrm{d}t = - \left[P_N \left\{ \widehat{u_N \frac{\partial u_N}{\partial x}} \right\} \right]_k (t) \mathrm{d}t + \mathrm{d}\beta_k(t) \, , \ k \in \mathbb{I}_N$$

FIG. 4.31 – *Moyenne temporelle des quantités* $| P_{N_1} B_{int}(y_{N_1}, z_{N_1}) |_{L^2(0,2\pi)}$ *(gauche) et* $|Q_{N_1} B_{int}(y_{N_1}, z_{N_1})|_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* N_1 *= 8 (1), 32 (2), 128 (3), 512 (4) avec* N *= 2560.*

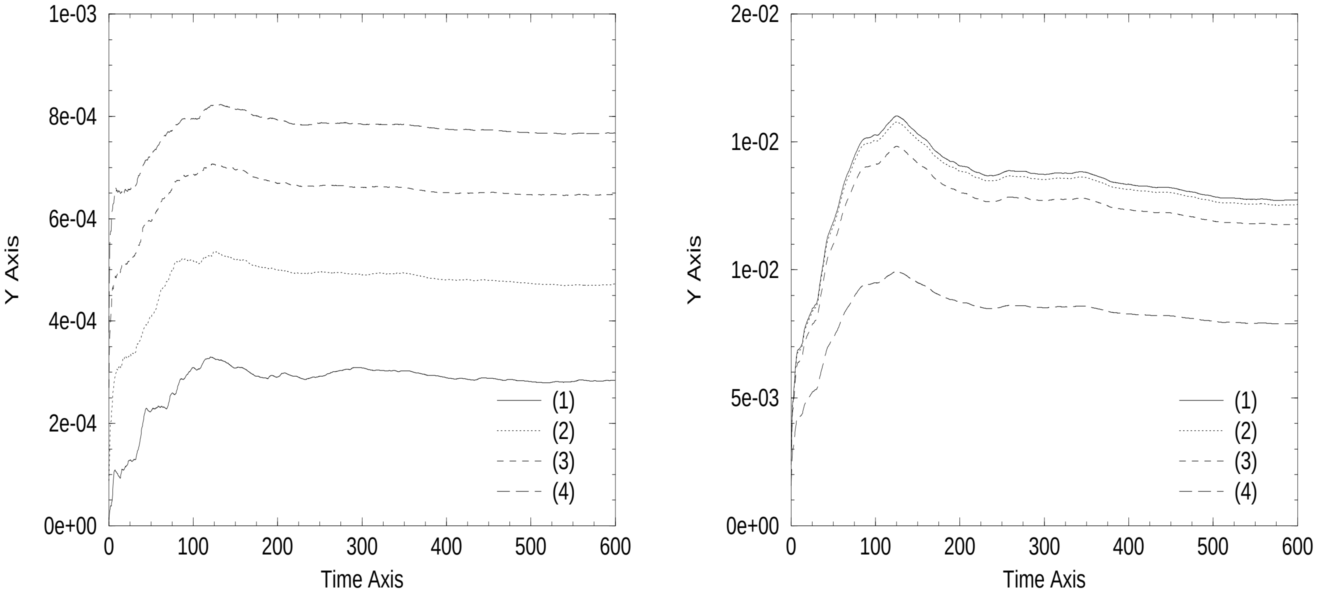


FIG. 4.32 – *Moyenne temporelle des valeurs nodales de la vitesse à différents temps :* $t = 240,\ 360$.

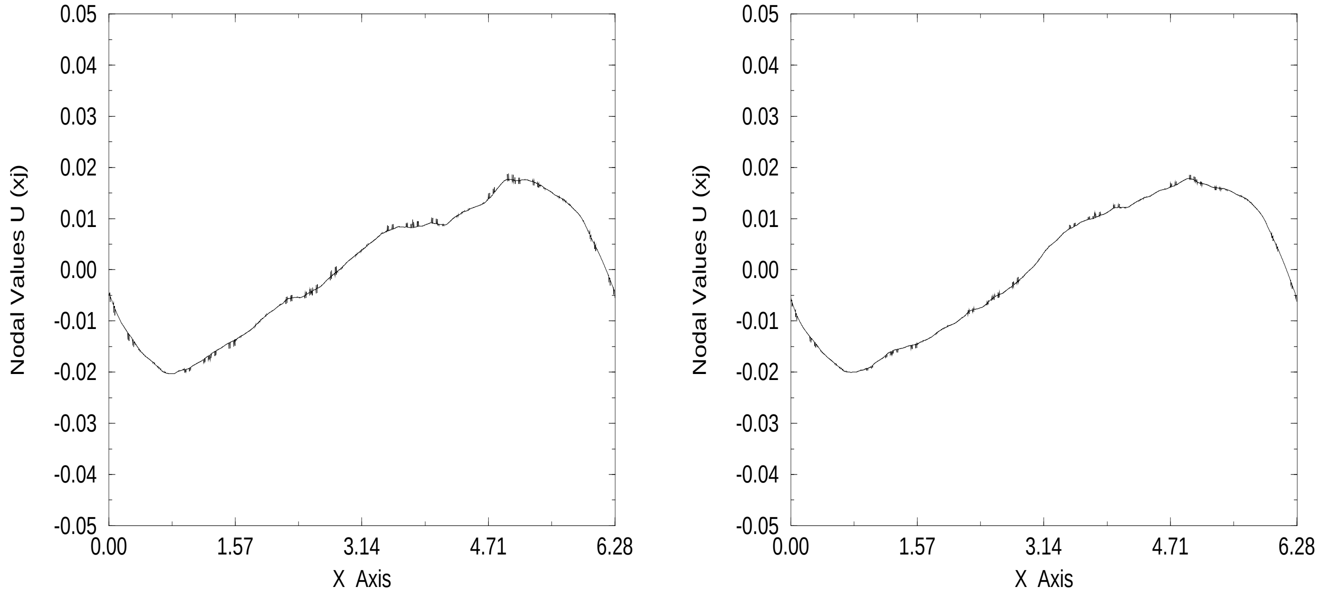


FIG. 4.33 – *Moyenne temporelle des valeurs nodales de la vitesse à différents temps : $t = 480, 600$.*

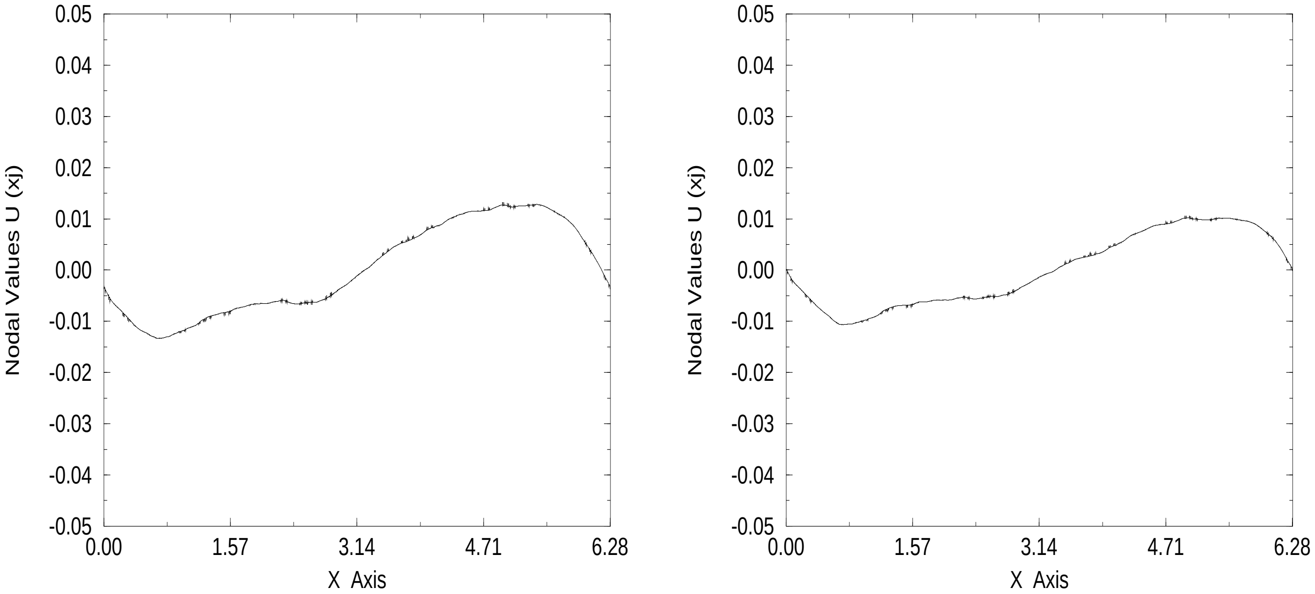


FIG. 4.34 – *Valeurs nodales instantannées de la vitesse (gauche) et spectre d'énergie (droite) à $t = 600$.*

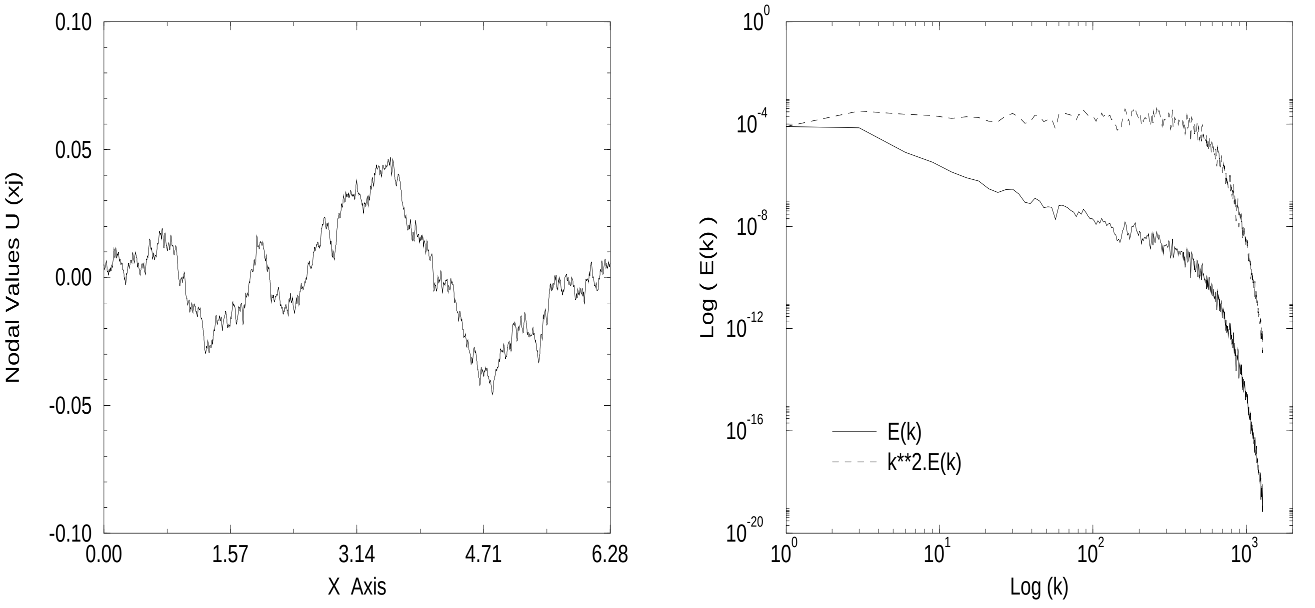
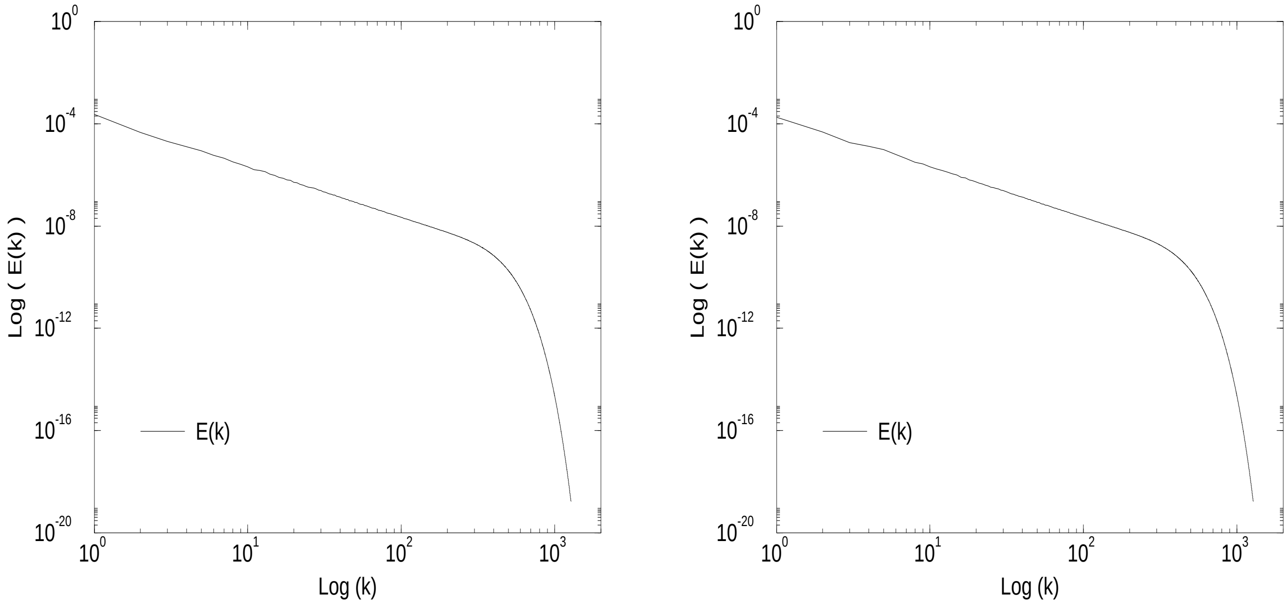


FIG. 4.35 – *Moyenne temporelle du spectre d'énergie à différents temps : t = 300, 600.*



La méthode de quadrature donne ensuite

$$\mathrm{d}\left\{e^{\nu k^2 t}\hat{u}_k(t)\right\} = \left\{-P_N\left\{\widehat{u_N\frac{\partial u_N}{\partial x}}\right\}(t)\mathrm{d}t + \mathrm{d}\beta_k(t)\right\} e^{\nu k^2 t} \tag{4.20}$$

Intégrons (4.20) sur $[t,\, t+\tau]$, nous obtenons

$$\begin{aligned}\hat{u}_k(t+\tau) &= e^{-\nu k^2\tau}\hat{u}_k(t+\tau) - \int_t^{t+\tau} e^{-\nu k^2[(t+\tau)-s]}P_N\left\{\widehat{u_N\frac{\partial u_N}{\partial x}}\right\}(s)\,\mathrm{d}s \\ &+ \int_t^{t+\tau} e^{-\nu k^2[(t+\tau)-s]}\,\mathrm{d}\beta(s)\end{aligned} \tag{4.21}$$

La dernière intégrale de (4.21) se discrétise $e^{-\nu k^2\tau}\ [\beta_k(t+\tau)-\beta_k(t)]$. La variable aléatoire $[\beta_k(t+\tau)-\beta_k(t)]$ suit, par définition, une loi $\mathcal{N}(0,\tau)$ ce que l'on peut reécrire sous la forme $\sqrt{\tau}\chi_k$ avec χ_k une variable aléatoire de loi $\mathcal{N}(0,1)$.

Remarque *15*

A chaque sous-pas de temps, il faut regénérer des variables aléatoires de loi $\mathcal{N}(0,1)$ *puis les multiplier par des coefficients constants,* $\sqrt{\Delta t_i}$, *ne dépendant que de la longueur du sous-pas de temps pour obtenir le spectre du bruit blanc.*

4.3.3.2 Simulation pour $\nu = 10^{-2}$.

Pour étudier convenablement l'évolution d'un flux soumis à une force aléatoire, donc très oscillante, la discrétisation en espace doit avoir un nombre de degrés de liberté suffisant pour contenir l'essentiel de l'information pour la simulation. Différents tests m'ont conduit à choisir $N = 2560$ modes pour une viscosité ν égale à 10^{-2}.

Remarque *16*

La moyenne en temps d'une quantité est calculée de la manière suivante

$$\overline{f(t)} = \frac{1}{t}\int_0^t f(s)\,ds\,,\ t > 0$$

Les paramètres de la simulation sont les suivants :
- la fréquence de coupure N fixée à 2560,
- la viscosité ν égale à 10^{-2},
- le pas de temps Δt égal à 3.10^{-3}.

L'intégration en temps de l'équation de Burgers a été effectuée sur l'intervalle [0, 600].
Nous partons d'une condition initiale nulle : $u_0 \equiv 0$.
Nous avons représenté graphiquement les moyennes des différentes quantités. L'absence de condition initiale perturbatrice fait entrer directement l'écoulement dans un régime chaotique. il en est de même pour le comportement de $|u_N|_{L^2(0,2\pi)}$, figure (4.27). A partir de $t = 300$, nous constatons une stabilisation des moyennes des normes.
Pour $N = 2560$, nous avons choisi les valeurs suivantes pour le second niveau
$N1 = 8, 32, 128, 512$.
Les grandes échelles, y_{N_1}, figure (4.28), présentent une évolution semblable à celle de la solution u_N durant toute la période d'intégration en temps : une forte croissance initiale puis une stabilisation en moyenne après $t = 300$. Par contre, les petites échelles, z_{N_1}, figure (4.28), les premiers instants passés, atteignent et conservent un régime stationnaire quel que soit le niveau N_1.
La dissipation de l'énergie du flux se produit essentiellement dans les petites échelles, figure (4.29), les grandes échelles n'ont ici qu'un apport secondaire.
Nous avons représenté les contributions du terme non linéaire aux figures (4.30) et (4.31). On remarque immédiatement que les interactions entre les petites et les grandes échelles n'ont pas toutes un régime stationnaire. On peut répartir ces quatre termes en deux catégories : ceux dont l'évolution est imposée par les grandes échelles ($P_{N_1}B(y_{N_1}, y_{N_1})$ et $Q_{N_1}B_{int}(y_{N_1}, z_{N_1})$) et ceux où les petites échelles dominent ($Q_{N_1}B(y_{N_1}, y_{N_1})$ et $P_{N_1}B_{int}(y_{N_1}, z_{N_1})$).
Dans l'équation gouvernant les grandes échelles, les termes dominants sont la dissipation et la contribution des grosses structures entre-elles. Dans l'équation gouvernant les petites échelles, la dissipation domine tous les autres termes.

FIG. 4.36 – *Moyenne temporelle du nombre de Courant (gauche) et des normes* $|u_N|_{L^2(0,2\pi)}$ *(1),* $|u_N|_{L^\infty(0,2\pi)}$ *(2) de la vitesse (droite).*

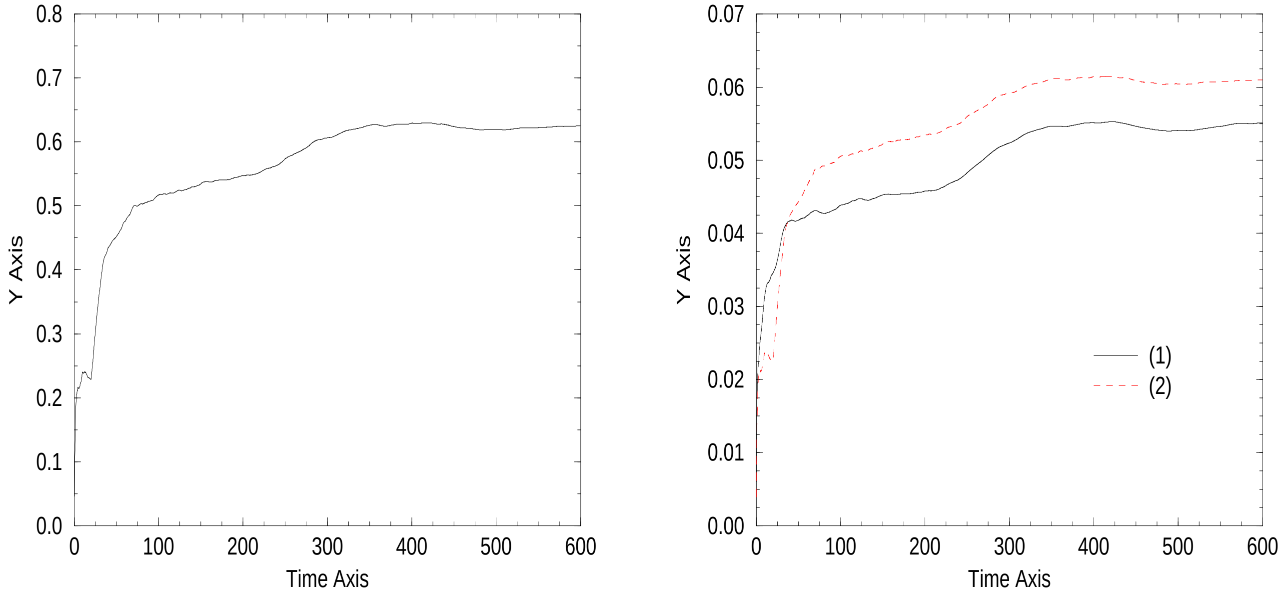
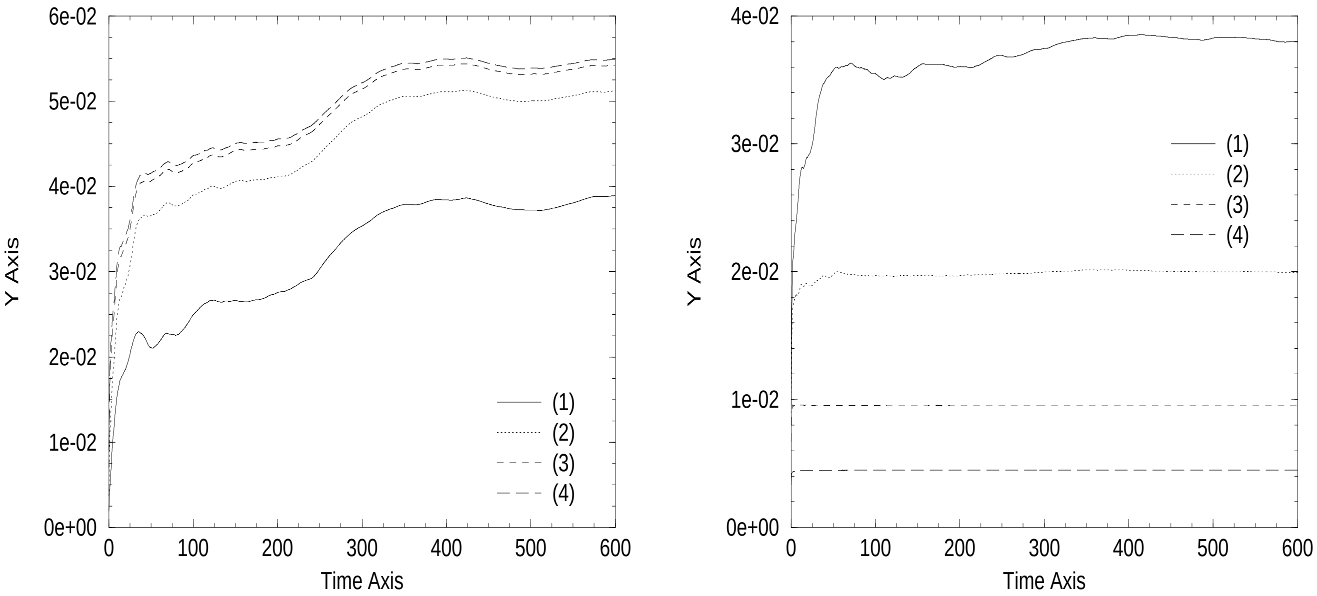


FIG. 4.37 – *Moyenne temporelle des quantités* $|y_{N_1}|_{L^2(0,2\pi)}$ *(gauche) et* $|z_{N_1}|_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* $N_1 = 8$ *(1), 32 (2), 128 (3), 512 (4) avec* $N = 5120$*.*



Nous avons tracé les valeurs instantannées et moyennées en temps de la vitesse dans l'espace physique, figures (4.32), (4.33), (4.34), ainsi que dans l'espace spectral, figures (4.34) et (4.35).
Comme on pouvait s'y attendre, l'allure de la vitesse dans l'espace physique, pour un écoulement turbulent, est très hachée (4.34). Par contre les moyennes en temps sont plus régulières, figures (4.32) et (4.33), avec de petites oscillations locales, réparties dans tout le domaine qui vont en s'atténuant. Ainsi ces petites zones sont sujettes à des perturbations dont les effets sont permanents puisqu'elles affectent de manière visible les moyennes en temps.
Nous remarquons que la zone inertielle du spectre d'énergie est très longue. D'autre part, la décroissance du spectre est en k^{-2}, figure (4.34), visible grâce au spectre du gradient), ce qui correspond à la décroissance prévue par la théorie.

4.3.3.3 Simulation pour $\nu = 10^{-3}$.

Les paramètres de la simulation sont les suivants :
- la fréquence de coupure N fixée à 5120,
- la viscosité ν égale à 10^{-3},
- le pas de temps Δt égal à 2.10^{-3}.

L'intégration en temps a été effectuée sur l'intervalle $[0, 600]$. Nous partons d'une condition initiale nulle : $u_0 \equiv 0$. Des tests préliminaires nous ont amenés à prendre 5120 modes en espace pour avoir suffisamment de précision lorsque de fortes variations apparaissent.
Les normes de la vitesse présentent une forte augmentation initiale puis semblent converger à partir de $t = 400$.
Pour $N = 5120$ modes, nous avons gardé les valeurs pour le second niveau
$N_1 = 8, 32, 128, 512$.
Les grandes échelles, figure (4.37), présentent un comportement similaire à celui de u_N. Les petites échelles atteignent rapidement un régime stationnaire en moyenne et le garde durant toute la période d'intégration. Il en est de même pour les projections du terme dissipatif, figure(4.38). Celle-ci a lieu essentiellement dans les petites échelles.
Les figures (4.39) et (4.40) représentent les différentes parties du terme de couplage $B(u_N, u_N)$.
Dans l'équation gouvernant les grandes échelles, le terme dissipatif et celui décrivant l'interaction des grandes échelles entre-elles ($P_{N_1}B(y_{N_1}, y_{N_1})$) prédominent; alors que celle régissant les petites échelles est totalement dominée par le terme de dissipation.
Les tracés instantannés, figure (4.43) et moyennés, figure (4.44), des spectres d'énergie montrent la présence d'une très grande zone inertielle dont la pente est en k^{-2}. Les tracés des moyennes n'offrent pas d'allure évoluant au cours du temps, il n'y a donc pas de grandes variations dans le spectre de la solution.

FIG. 4.38 – *Moyenne temporelle de* $\nu\left|\dfrac{\partial^2 y_{N_1}}{\partial x^2}\right|_{L^2(0,2\pi)}$ *(gauche) et* $\nu\left|\dfrac{\partial^2 z_{N_1}}{\partial x^2}\right|_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* $N_1 = 8$ *(1), 32 (2), 128 (3), 512 (4) avec* $N = 5120$*.*

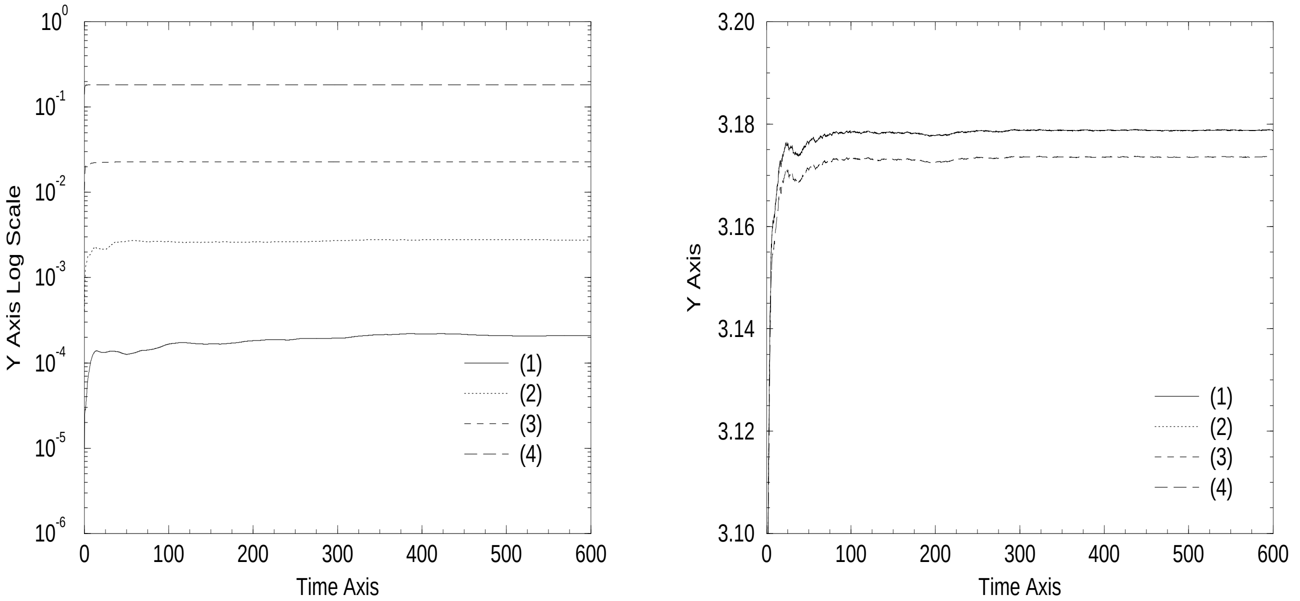


FIG. 4.39 – *Moyenne temporelle des quantités* $|\ P_{N_1}B(y_{N_1}, y_{N_1})\ |_{L^2(0,2\pi)}$ *(gauche) et* $|Q_{N_1}B(y_{N_1}, y_{N_1})|_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* $N_1 = 8$ *(1), 32 (2), 128 (3), 512 (4) avec* $N = 5120$*.*

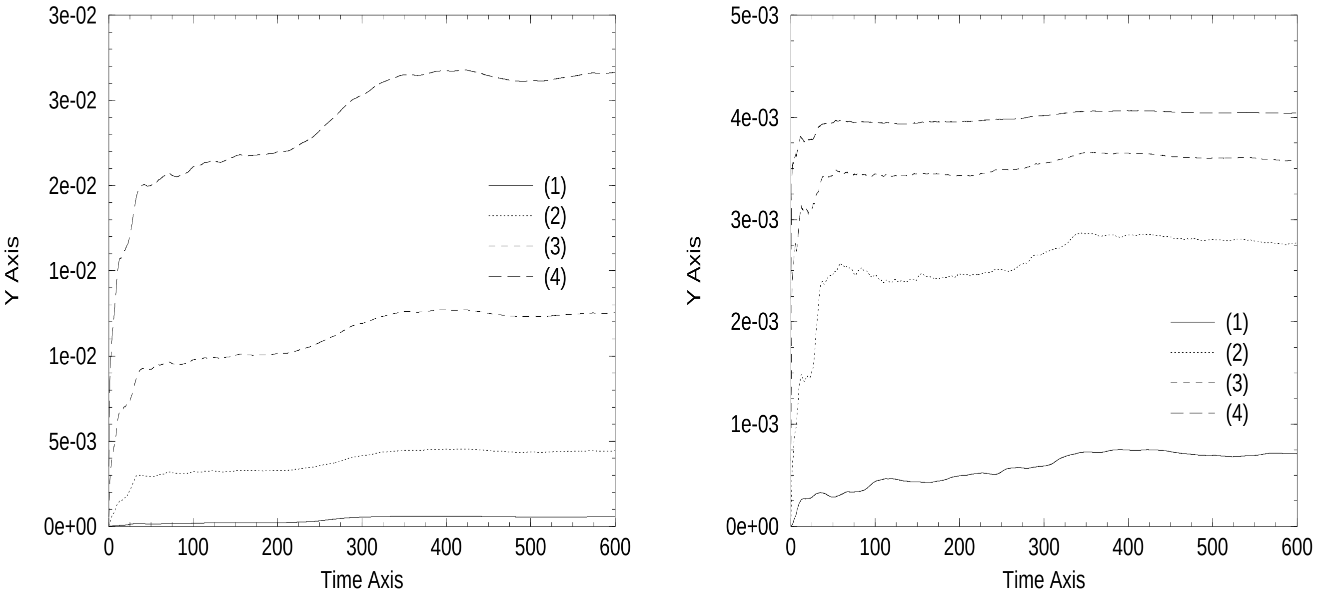
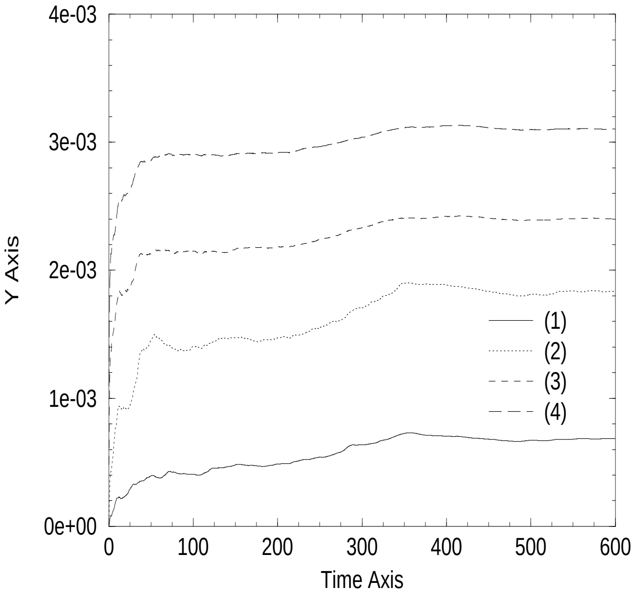
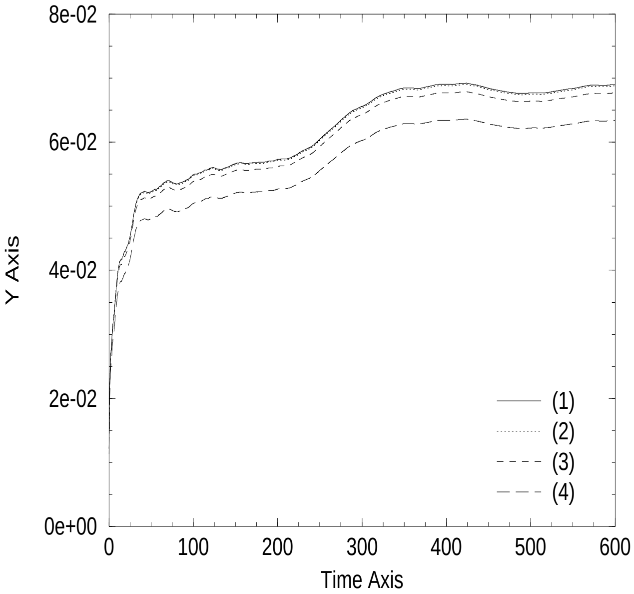


FIG. 4.40 – *Moyenne temporelle des quantités* $|\ P_{N_1} B_{int}(y_{N_1}, z_{N_1})\ |_{L^2(0,2\pi)}$ *(gauche) et* $|Q_{N_1} B_{int}(y_{N_1}, z_{N_1})|_{L^2(0,2\pi)}$ *(droite) pour différentes valeurs de* N_1 *= 8 (1), 32 (2), 128 (3), 512 (4) avec* N *= 5120.*





Intéressons-nous maintenant aux valeurs de la solution dans l'espace physique. La figure (4.43) nous indique les forts gradients de diverses amplitudes répartis dans tout le domaine. Si nous considérons les moyennes en temps des valeurs nodales, figures (4.41) et (4.42), les forts gradients générés par le bruit blanc sont visibles sur de longues périodes, il y a donc des zones de forte turbulence qui se développent et sont entretenues.
Toutefois, au fil du temps, nous remarquons une régularisation, une uniformisation de la moyenne de la vitesse.

4.3.3.4 Comparaison des deux simulations.

Nous avons effectué deux simulations d'écoulement turbulent, l'une avec une viscosité ν égale à 10^{-2}, l'autre avec $\nu = 10^{-3}$. Le fait de prendre une viscosité dix fois plus petite nous a obligé à doubler le nombre de modes N pour pouvoir capter convenablement la solution.
Dans les deux cas, nous voyons une forte croissance de l'énergie du flux puis une évolution plus lente, moins intense. Nous avons constaté que les grandes échelles contiennent la majeure partie de l'énergie de l'écoulement. les petites échelles entrant presque immédiatement dans un régime stationnaire.
Par contre, la dissipation se produit essentiellement dans les petites échelles, les grandes échelles n'ont ici qu'un apport secondaire. Cela est plus frappant encore pour $\nu = 10^{-3}$ que pour $\nu = 10^{-2}$.
Mais dans les deux cas, nous constatons numériquement une convergence de la moyenne en temps de la vitesse, ce qui est en accord avec les résultats de la théorie.

FIG. 4.41 – *Moyenne temporelle des valeurs nodales de la vitesse à différents temps :* $t = 240, 360$.

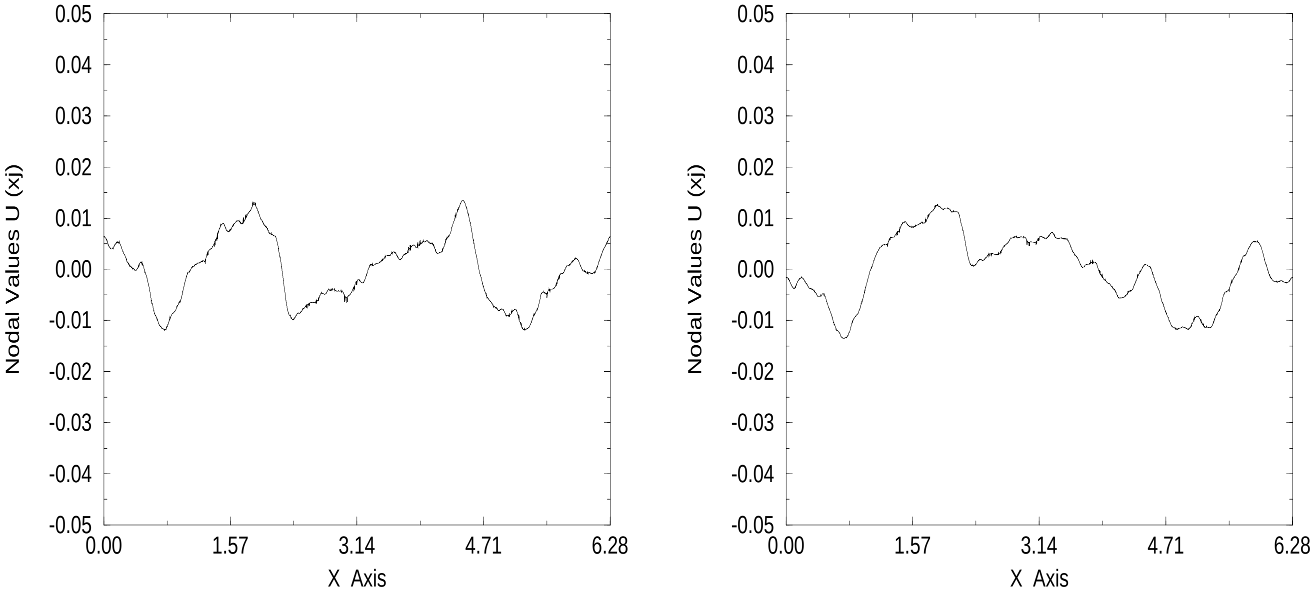


FIG. 4.42 – *Moyenne temporelle des valeurs nodales de la vitesse à différents temps* $t = 480,\ 600$.

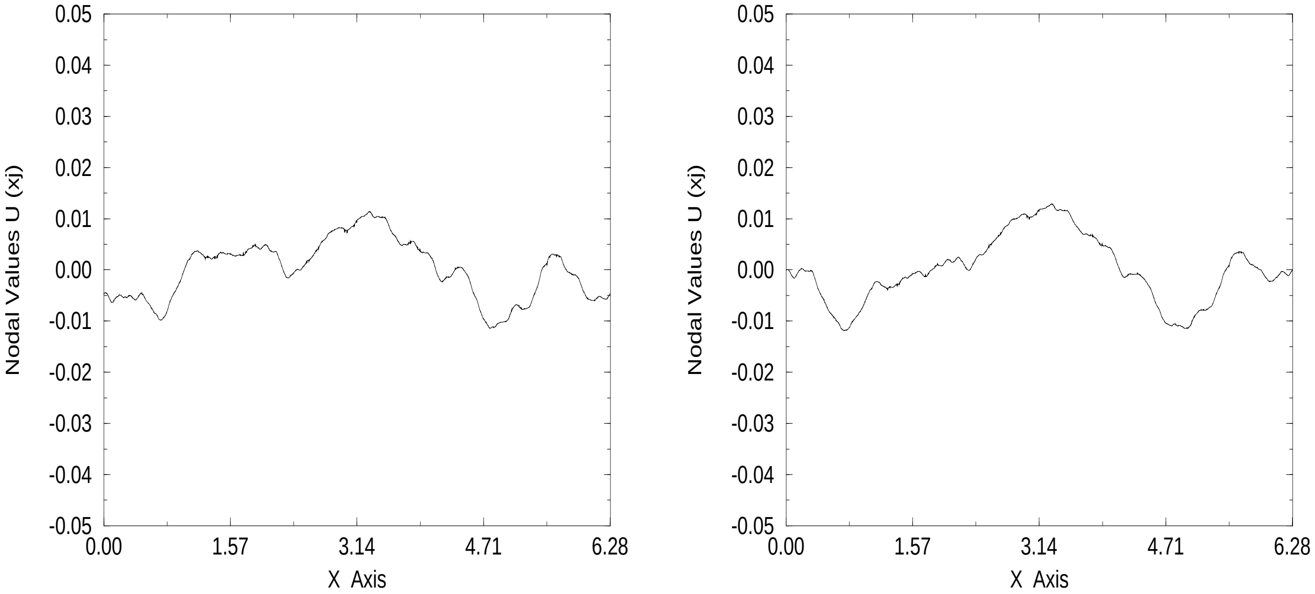
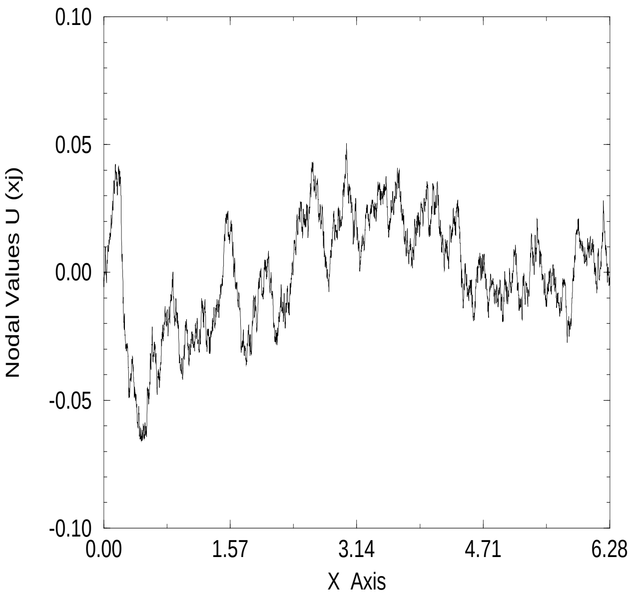


FIG. 4.43 – *Valeurs nodales instantannées de la vitesse (gauche) et spectre d'énergie (droite) à t = 600.*



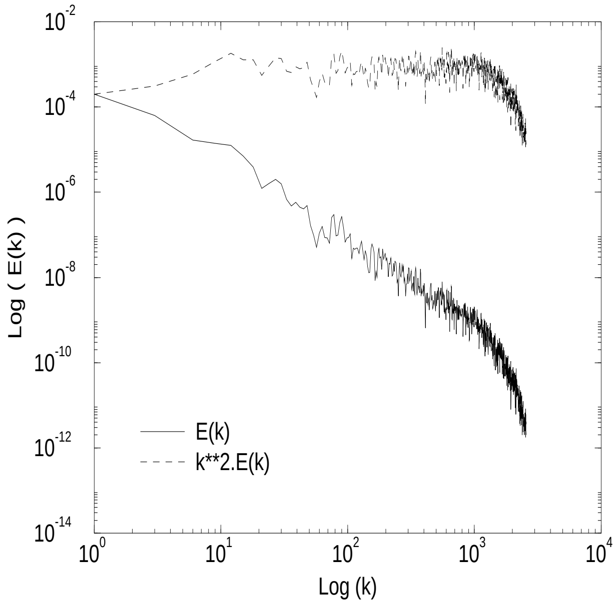
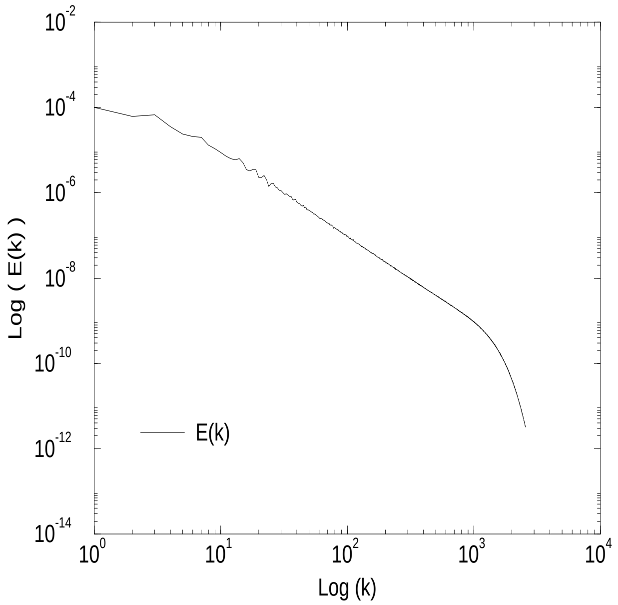
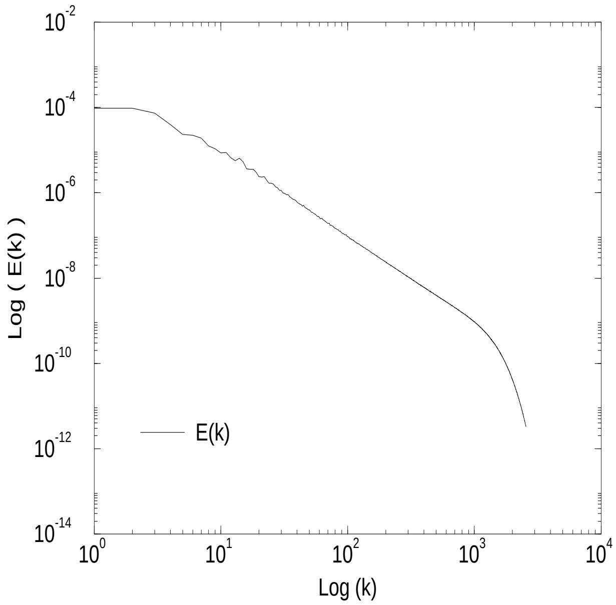


FIG. 4.44 – *Moyenne temporelle du spectre d'énergie à différents temps : t = 300, 600.*





4.3.4 Simulation numérique avec des conditions aux limites de non glissement.

Dans cette simulation, nous considérons un écoulement soumis à l'action d'un bruit blanc

$$\mathrm{d}u = \left\{ \frac{\partial^2 u}{\partial x^2} + \frac{1}{2}\frac{\partial(u^2)}{\partial x} \right\} \mathrm{d}t + \mathrm{d}W$$

où W est un processus de Wiener défini par (4.15) dans le cas de conditions aux limites de type Dirichlet homogène :

$$u(\pm 1,\, t) = 0$$

Le but de cette simulation est de comparer qualitativement les résultats obtenus avec ceux de ([14]) dans le cas d'une discrétisation en espace par méthode spectrale.

4.3.4.1 Cadre de l'étude.

4.3.4.1.1 Description de la condition initiale.

On part d'une condition initiale aléatoire de moyenne nulle et de variance un construite dans l'espace spectral, i.e.

$$< u_0 >_x = 0\,, \quad < u_0^2 >_x = 1$$

où $< . >_x$ désigne la moyenne en espace.

4.3.4.1.2 Description de la force extérieure.

On construit la force extérieure comme un bruit blanc. Cela signifie que $W(t)$ s'écrit sous la forme

$$W(t) = \sum_{k\in\mathbb{N}} \beta_k(t)\, T_k$$

avec

$\{T_k\}_k$	la famille des polynômes de Chebyshev,
$\{\beta_k\}_k$	une suite de mouvements browniens réels indépendants dans un espace probabilisé $(\Omega,\, \mathcal{F},\, P)$.

Partant de l'équation (4.16), nous appliquons la méthode Tau et on obtient alors

$$\mathrm{d}\hat{u}_k(t) - \nu \hat{u}_k^{(2)}(t)\mathrm{d}t = -\left\{ P_N \left[\widehat{u_N \frac{\partial u_N}{\partial x}} \right] \right\}_k (t)\mathrm{d}t + \mathrm{d}\beta_k(t)\,,\ k \in [\![0,\, N-2]\!] \tag{4.22}$$

munie des conditions limites

$$u_N(\pm 1,\, t) = 0$$

4.3.4.1.3 Discrétisation en temps.

La discrétisation en temps est effectuée à l'aide des schémas d'ordre 2 suivants: Crank-Nicholson (semi-implicite) pour la partie linéaire, Adams-Bashforth (explicite) pour la partie non linéaire et le bruit blanc.
Soit $t \geqslant 0$ et $\tau > 0$. L'intégration de (4.22) sur $[t,\, t+\tau]$ donne

$$\begin{aligned} \hat{u}_k(t+\tau) + \frac{\nu\tau}{2}\hat{u}_k^{(2)}(t+\tau) &= \hat{u}_k(t) - \frac{\nu\tau}{2}\hat{u}_k^{(2)}(t) + \frac{\tau}{2}\left\{3\widehat{NL}_k(t) - \widehat{NL}(t-\tau)\right\} \\ &+ \int_t^{t+\tau} \mathrm{d}\beta_k(s)\,,\ k \in [\![0,\, N-2]\!] \end{aligned}$$

en posant

$$\widehat{NL}_k(t) = \left\{P_N\left[\widehat{u_N\frac{\partial u_N}{\partial x}}\right]\right\}_k (t)\,,\ k \in [\![0,\, N-2]\!]$$

L'intégrale du membre de droite est égale à

$$\int_t^{t+\tau} \mathrm{d}\beta_k(s) = \beta_k(t+\tau) - \beta_k(t) \sim \mathcal{N}_k(0,\, \tau)$$

où la notation $\mathcal{N}_k(0,\, \tau)$ désigne une loi normale de moyenne nulle et de variance τ pour $k \in [\![0,\, N-2]\!]$. Ces variables aléatoires sont non corrélées entre-elles et dans le temps.
Nous appliquons deux filtres à ce bruit blanc.

4.3.4.1.4 Filtre en espace.

On applique un filtre dans l'espace spectral de la manière suivante:
soit N la fréquence de coupure et N_1 un entier inférieur à N $\left((\text{e.g. } N_1 = \frac{N}{2}\right)$. On procède comme précédemment quant à la définition de ses modes pour $k \in [\![0,\, N_1-1]\!]$. Pour $k \in [\![N_1+1,\, N]\!]$, on impose 0 comme valeur pour $\mathcal{N}_k(0,\, \tau)$.
Les (N_1+1) premiers modes de W_N (discrétisation du bruit blanc) restent non corrélés. Ainsi, cela revient à définir W_N dans l'espace physique sur une grille moins fine.

$$W(t) = \sum_{k=0}^{N_1} \beta_k(t)\, T_k$$

4.3.4.1.5 Filtre en temps.

On applique un filtre en temps au bruit blanc en le figeant sur un intervalle de longueur Δt_r, avec $\Delta t_r \geqslant \Delta t$ le pas de la discrétisation en temps, i.e. sur un intervalle de la forme, $(p\Delta t_r,\, (p+1)\Delta t_r)$, pour p entier.
Si $n\Delta t$ et $n'\Delta t$ appartiennent à deux intervalles distincts de cette forme alors les variables aléatoires $\{\mathcal{N}_k(0,\tau)\}_{n\Delta t}$ et $\{\mathcal{N}_k(0,\tau)\}_{n'\Delta t}$ sont choisies non corrélées.

4.3.4.2 Simulations numériques.

Dans cette simulation, nous considérons un écoulement entrainé par une force extérieure aléatoire de type bruit blanc. L'article ([14]), qui correspond à une discrétisation spatiale de type différences finies, nous servira de comparatif pour les résultats.Par conséquent, nous nous plaçons dans le même cadre d'étude et utilisons les mêmes notations.

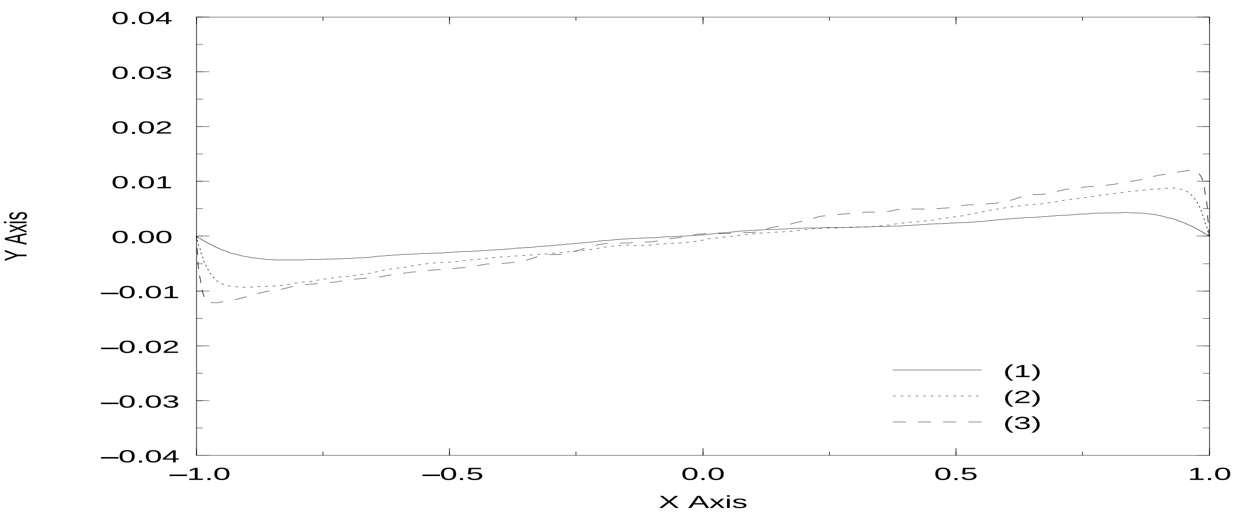
L'équation de Burgers s'écrit alors

$$\frac{\partial \tilde{u}}{\partial \tilde{t}} + \frac{\partial}{\partial \tilde{x}}\left(\frac{\tilde{u}^2}{2}\right) = \nu \frac{\partial^2 \tilde{u}}{\partial \tilde{x}^2} + \tilde{\chi}\left(\tilde{x}, \tilde{t}\right) \; , \; 0 < \tilde{x} < L \, , \; \tilde{t} \geqslant 0$$

munie des conditions limites

$$\tilde{u}\left(\tilde{x} = 0\right) = \tilde{u}\left(\tilde{x} = L\right) = 0$$

FIG. 4.45 – *Moyenne temporelle des valeurs nodales de la vitesse pour* $\Delta t_r = 10^{-1}$ *et* $Re = 1000$ *(1),* $Re = 3000$ *(2),* $Re = 9000$ *(3).*



en posant
$\tilde{u}$ est la vitesse,
ν la viscosité cinématique,
$\tilde{\chi}$ le forçage aléatoire de type bruit blanc,
L la longueur du domaine.

Le forçage est un processus aléatoire, de type bruit blanc en $\tilde{x}$ avec une moyenne nulle. La variance du forçage dimensionné, σ^2, définit une échelle de vitesse $U = (\sigma L)^{\frac{1}{2}}$. L'équation de Burgers, dans cette forme non dimensionnelle utilisant U et L s'écrit :

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = \frac{1}{Re^*}\frac{\partial^2 u}{\partial x^2} + \chi^*(x,t) \; ; \quad -1 < x < 1$$

avec

$$u(x = -1) = u(x = 1) = 0$$

FIG. 4.46 – *Moyenne temporelle des valeurs nodales de la vitesse pour* $Re = 9000$ *et* $\Delta t_r = 10^{-2}$ *(1)*, $\Delta t_r = 10^{-1}$ *(2)*, $\Delta t_r = 1$ *(3)*.

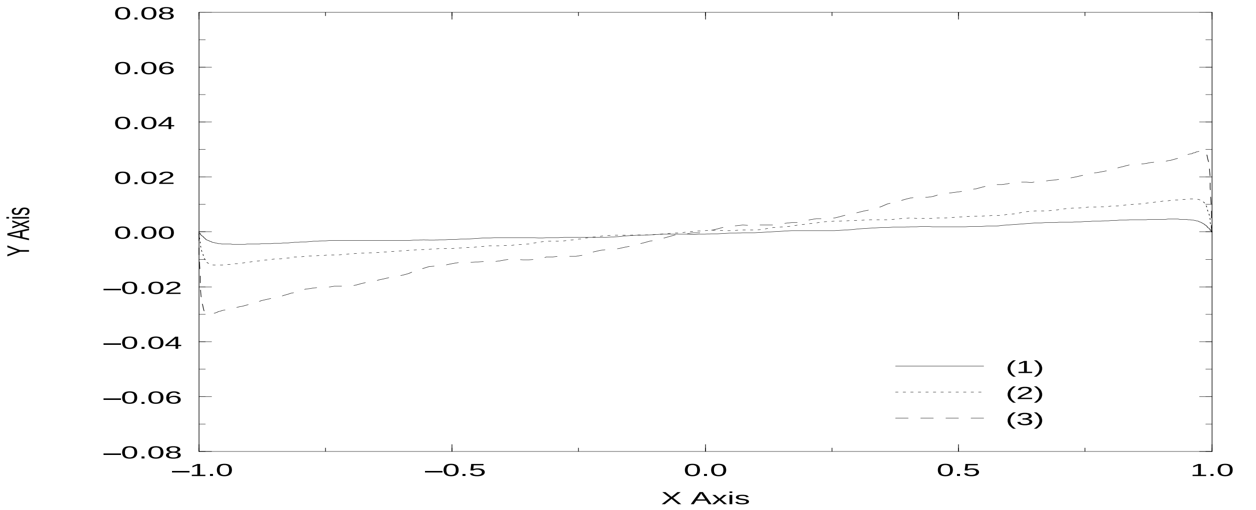
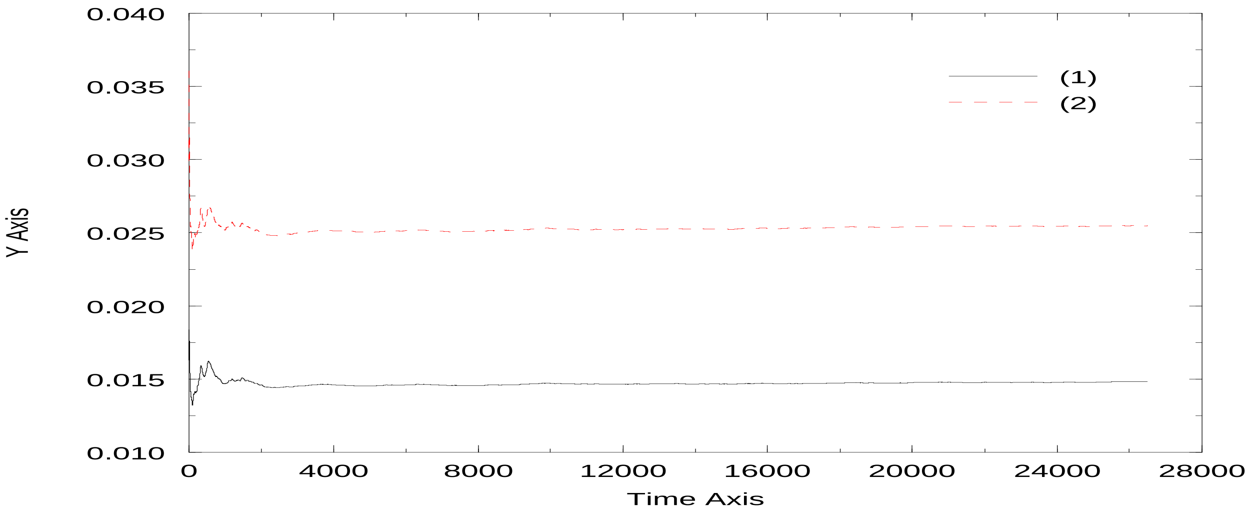


FIG. 4.47 – *Moyenne temporelle des normes* $|\, u_N \,|_{L^2(0,2\pi)}$ *(1) et* $|\, u_N \,|_{L^\infty(0,2\pi)}$ *(2) de la vitesse pour* $Re = 1000$ *et* $\Delta t_r = 10^{-1}$.



où u, x, t et χ^* sont des quantités adimensionnées; Re^* est le nombre de Reynolds $\dfrac{UL}{\nu}$ et $< \chi^* >_x = 0$, $< (\chi^*)^2 >_x = 1$.
Notons Re le nombre de Reynolds et χ le bruit blanc utilisé dans ([14]). Nous travaillons sur l'intervalle $(-1, +1)$ au lieu de $(0, 1)$. Pour obtenir ces paramètres nous avons les relations suivantes

$$\frac{2}{Re^*} = \frac{1}{Re} \quad \hookrightarrow \quad Re^* = 2Re$$

$$\frac{\chi^*}{2} \text{ de même loi que } \chi \quad \hookrightarrow \quad \chi^* \text{ est de loi } \mathcal{N}(0, 4)$$

Nous reprenons les paramètres de la discrétisation en temps utilisés dans ([14]); i.e. Δt le pas de temps égal à 10^{-3} pour toutes les simulations alors que Δt_r prend les valeurs 10^{-2}, 10^{-1} et 1. De même, la longueur du domaine nous oblige à prendre 1000, 3000 et 9000 resp. au lieu de 500, 1500 et 4500 comme nombre de Reynolds. Les calculs ont été effectués avec une fréquence de coupure N fixée à 256 et une fréquence de filtrage N_1 égale à 128.
Les figures (4.45) et (4.46) montrent les effets du nombre de Reynolds Re et de l'échelle en temps Δt_r sur la moyenne de la vitesse. L'intensité et les gradients de la vitesse moyenne augmentent visiblement avec Re et Δt_r croissants.
Dû à la nature convective de la solution de l'équation de Burgers, l'épaisseur de la couche limite tend à diminuer lorsque le nombre de Reynolds augmente, figure (4.45). Les gradients de la vitesse moyenne, près du milieu du domaine changent cependant peu avec le nombre de Reynolds.
De manière opposée, une échelle de temps croissante, à nombre de Reynolds fixé augmente sensiblement les gradients aussi bien au milieu du domaine que près du bord. Cependant, l'épaisseur de la couche limite n'est pas affectée par une échelle Δt_r croissante, figure (4.46).

Les intégrations en temps ont été faites sur différents intervalles de temps allant de 7000 à 26500 suivant le Reynolds et l'échelle de filtrage en temps. Le critère d'arrêt de l'intégration en temps est basé sur le profil de la vitesse moyenne dans l'espace physique. D'un point de vue qualitatif, nous avons observé les mêmes phénomènes de couche-limite en faisant varier les paramètres (Re et Δt_r). D'un point de vue quantitatif, nous n'obtenons pas les mêmes ordres de grandeur pour les moyennes. Cela peut être dû au fait que la force aléatoire n'est pas construite sur une base de $L^2(-1, +1)$ car la famille des polynômes de Chebyshev n'est pas orthogonale pour le produit scalaire usuel mais l'est pour le produit scalaire pondéré $(.,.)_w$.
Les figures 53, 54 et 55 représentent les moyennes en temps des normes L^2 et L^∞ de la vitesse dans les différents cas considérés.
Nous obtenons numériquement la convergence de ces quantités.

FIG. 4.48 – *Moyenne temporelle des normes L^2 (1) et L^∞ (2) de la vitesse pour $Re = 3000$, $\Delta t_r = 10^{-1}$ (gauche) et $Re = 9000$, $\Delta t_r = 10^{-2}$ (droite).*

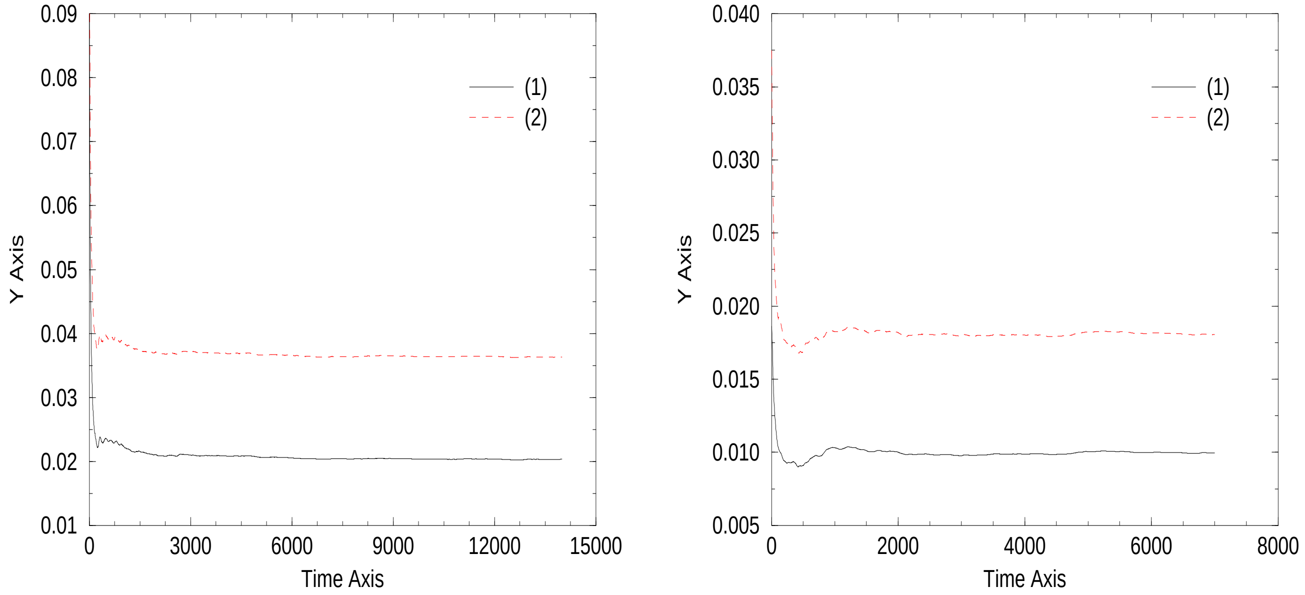
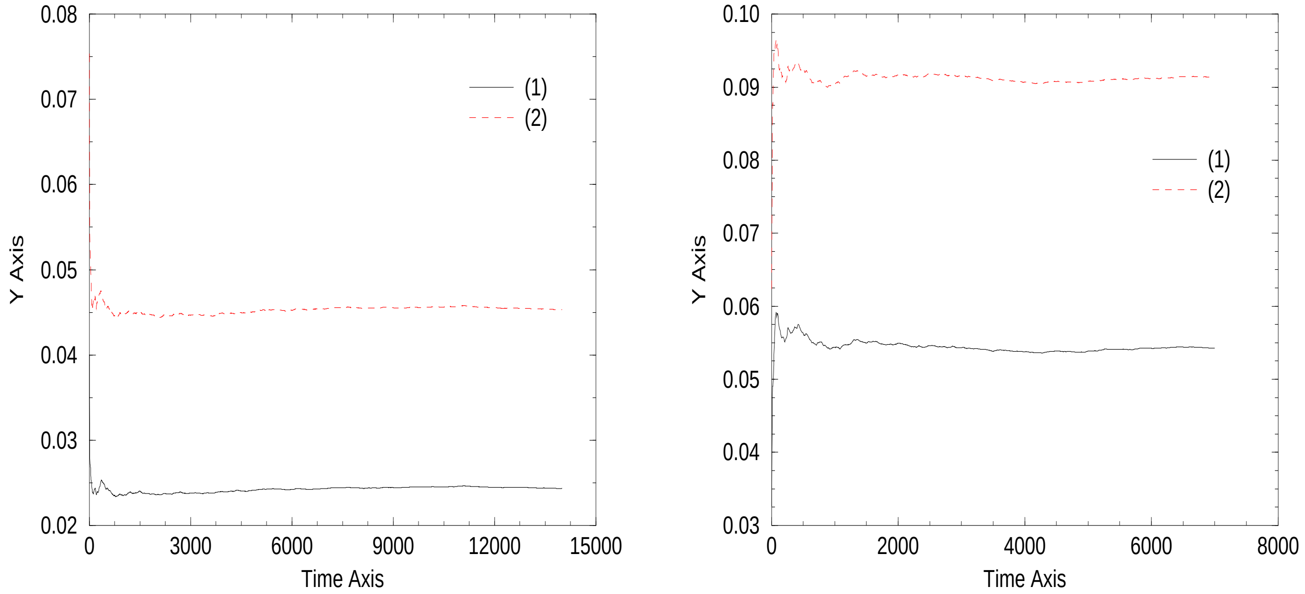


FIG. 4.49 – *Moyenne temporelle des normes L^2 (1) et L^∞ (2) de la vitesse pour $Re = 9000$, $\Delta t_r = 10^{-1}$ (gauche) et $Re = 9000$, $\Delta t_r = 1$ (droite).*



4.4 Conclusion.

Nous avons employé l'équation de Burgers comme modèle simplifié des équations de Navier-Stokes dans le cadre de la modélisation de la turbulence. La résolution numérique a été effectuée d'une part avec un forçage déterministe, d'autre part avec un forçage aléatoire. Nous considérons l'évolution des moyennes temporelles des quantités provenant des projections des différents termes de l'équation. Nous obtenons la convergence numérique de ces moyennes.

Pour les conditions aux limites périodiques, les résultats moyennés obtenus sont qualitativement comparables avec ceux obtenus dans le cas des équations de Navier-Stokes déterministes, en particulier en ce qui concerne l'importance relative des différents termes des équations projetées.

Les simulations effectuées dans le cas de conditions aux limites de type non glissement présentent des allures similaires aux résultats obtenus dans ([14]). Nous obtenons les mêmes phénomènes de couche-limite.

Toutefois le nombre de modes nécessaire pour contenir l'essentiel de l'information est important : plusieurs milliers en une dimension d'espace. La généralisation de cette méthode aux équations de Navier-Stokes avec plusieurs dimensions d'espace n'est pas concevable. Pour cela, il faudrait considérer un forçage s'exprimant à l'aide d'un nombre plus restreint de modes.

Conclusions

Dans ce travail, nous nous sommes intéressés à différents aspects des méthodes spectrales multi-niveaux. Ces méthodes sont basées sur le découpage de la solution en plusieurs parties.

La propagation de solitons nous offre la possibilité d'appliquer une telle décomposition. L'étude numérique des actions des opérateurs linéaire et non linéaire nous permet de calculer des hautes fréquences du signal à moindre coût mais à qualité égale.

Dans le cadre de l'électromagnétisme, le traitement de conditions aux limites de type non périodique constitue le deuxième aspect étudié. Une méthode spectrale Tau-Legendre nous fournit un système couplé d'équations projetées sur les basses et les hautes fréquences des champs électromagnétiques solutions du problème de la cavité résonnante. La séparation de la solution est réalisée indépendemment de toute contrainte; aisni elle peut s'appliquer de manière très générale. La résolution numérique d'un tel système est effectuée à l'aide d'un algorithme de point fixe sous-relaxé. Le bon conditionnement des matrices issues des différentes discrétisations nous permet d'obtenir ls convergence numérique en peu d'itérations.

Le troisième aspect concerne la convergence numérique des grandes et petites échelles d'un écoulement pour différents niveaux et cela dans le cadre de la modélisation de la turbulence. L'équation de Burgers, modèle simplifié des équations de Navier stokes gouverne l'évolution d'un écoulement soumis à un forçage aléatoire. Nous nous intéressons au comportement qualitatif des différents termes des équations projetées.

Annexe A

Dans ce dernier paragraphe, nous allons nous intéresser aux codes de calcul écrits pour implémenter les méthodes spectrales utilisées.
Ces différents codes ont, tout d'abord, été testés sur les machines locales (Stations Sun Sparc 2 et 10) disponibles au Laboratoire d'Analyse Numérique d'Orsay.
Ensuite, les validations des méthodes et les tests de stabilité numérique ont été effectués sur une machine vectorielle locale, un Kubota Titan biprocesseur.
Enfin, les simulations numériques d'écoulement turbulent ont été implémentés sur un super calculateur Cray YMP quadriprocesseur du Centre de Recherche en Informatique (C.R.I) du Campus d'Orsay.

La vectorisation d'un code consiste en fait à vectoriser les boucles d'instructions s'y trouvant. Cela signifie que l'on fait exécuter des boucles de manière plus rapide, les boucles qui n'ont pas de dépendances, i.e. de lien entre deux ou plusieurs valeurs successives de l'indice de boucle.
Pour analyser les dépendances dans chaque boucle du programme, nous avons d'abord testé le code avec le compilateur du Titan, puis celui du Cray avec son analyseur FPP. En étudiant les modifications apportées sur les fichiers sources par le compilateur, nous pouvons savoir exactement comment le code a été traduit et quelle partie a été ou non vectorisée. L'analyseur FPP supportent les options -Zv -Zp (vectorisation ou parallélisation). Cela consiste en la phase d'analyse des dépendances du compilateur CRAY CF-77. Les modifications sont visibles sur le programme source et le nouveau source Fortran modifié est généré alors que la version originale n'est pas modifiée. La nouvelle version peut être visualisée en étudiant les fichiers .l.
Ces analyseurs fournissent de bons outils pour vérifier si un programme est, ou n'est pas, bien écrit en considérant l'optimisation. Nous précisons aussi que des directives ont été placées lorsque cela était nécessaire afin d'améliorer la vectorisation effectuée par le compilateur.

Une analyse des performances des codes a montré que l'essentiel du temps CPU total (plus de 70 %) est consommé par le calcul des FFT ou des FCT.
Ces transformations discrètes, réelles ou complexes, entre l'espace physique et l'espace spectral sont effectuées à l'aide de bibliothèques de sous-programmes telles que NAG, IMSL.
Il nous a donc semblé essentiel de choisir avec soin les sous-programmes, dans les librairies, qui allaient calculer ces transformations. Pour cela, nous avons effectué, pour chacune d'entre-elles, un aller-retour entre l'espace physique et l'espace spectral, avec les options

de compilation appropriées. Ces tests ont été mesurés par des logiciels d'étude de performances comme "proview" disponible sur le CRAY-YMP ou "hpm" sur le CRAY 90 de l'I.D.R.I.S. (C.N.R.S.).
Cela nous a permis de déterminer de manière incontestable les transformations les mieux adaptées, parmi celles disponibles, aux calculs effectués.

Bibliographie

[1] ADAM J.C., GOURDIN SERVENIERE A., NEDELEC J.C. et RAVIART P.A.,
Study of an implicit scheme for integrating Maxwell's equations
(1974) Computer Methods in Applied Mechanics and Engineering, vol **22**, pp 327-346

[2] AGRAWAL G.P.,
Nonlinear Fiber Optics
Quantum Electronics, Principles and Applications

[3] ALPERT B.K. et ROKHLIN V.,
A fast algorithm for the evaluation of Legendre expansion
(1991) J. of Scientific and Statistical Computations, vol **12**, pp 158-179

[4] BEN BELGACEM F. et BERNARDI C.,
Spectral element discretization of the Maxwell equations
(1997) Publication du Laboratoire d'Analyse Numérique, Université Paris 6, vol **16**

[5] BENTON E. et PLATZMAN G.W.,
A table of solutions of the one-dimensional Burgers equation
(1972) Quarterly Journal of Mechanics and Applied Mathematics, vol **3**, pp 201-230.

[6] BERNARDI C. et MADAY Y.,
Approximations spectrales de problèmes aux limites elliptiques
(1992) Mathématiques et Applications, **10**, Springer Verlag

[7] BLACKSTOCK D.T.,
Convergence of the Keck-Boyer perturbation solution for plane waves of finite amplitude in a viscous fluid
(1966) J. Acoust. Soc. Am., vol **39**, pp 411-413.

[8] BOSSAVIT A.,
Electromagnétisme en vue de la modélisation.
(1993) Mathématiques et Applications, **14**, Springer Verlag

[9] BRIGHAM E.O.,
The Fast Fourier Transform
(1974) Prentice-Hall, Englewood Cliffs, NJ

[10] BURGERS J.M.,
A mathematica model illustrating the theory of turbulence
(1948) Advances in Applied Mechanics, vol **1**, pp 171-199.

[11] BURGERS J.M.,
The Nonlinear Diffusion Equation
(1974), Ed. Reidel Boston.

[12] BUTCHER J.C.,
The numerical analysis of ordinary differential equations
Runge-Kutta and general linear methods
(1987), Ed. John Wiley & Sons

[13] CANUTO C., HUSSAINI M.Y., QUARTERONI A. et ZANG T.A.,
Spectral Methods in Fluid Dynamics
(1987) Springer Series in Computational Physics, Springer Verlag

[14] CHOI H., TEMAM R., MOIN P., KIM J.
Feedback Control for unsteady flow and its application to the stochastic Burgers Equation
(1993) J. of Fluid Mechanics vol **253**, pp 509-544

[15] COLE J.D.,
On a quasi-linear parabolic equation occuring in Aerodynamics
(1951) Quarterly in Applied Mathematics, vol **9**, pp 225-236.

[16] COOLEY J.W. et TURKEY J.W.,
An algorithm for the Machine Calculation of complex Fourier Series
(1965) Mathematics of Computation, vol **19**, pp 297-301.

[17] CROUZEIX M.,
Sur l'approximation des équations différentielles opérationnelles linéaires.
(1975), Thèse Paris

[18] DA PRATO G., DEBUSSCHE A. et TEMAM R.,
Stochastic Burgers' Equation
(1994) Nonlinear Differential Equations and Applications, vol **1** pp 389-402.

[19] DAUTRAY R et LIONS J.L.,
Analyse numérique matricielle pour les sciences et les techniques, tomes 1,2,3
(1984,1985) Masson, Paris

[20] DEKKER K. et VERWER J.G.,
Stability of Runge Kutta methods for stiff nonlinear differential equations
(1984) CWI Monograpgs vol **2**, North Holland, Amsterdam

[21] DEVILLE M., HALDENWANG P. et LABROSSE G.,
Comparison of Time Integration (Finite Difference and Spectral) for the Nonlinear Burgers' Equation
(1981) Proc 4th GAMM Conf. Numer. Methods in Fluid Mechanics ed Viviand H., Vieweg Braunschweig

[22] DEVILLE M., KLEISER L. et MONTIGNY-RANNON F.,
Pressure and time treatment of Chebychev spectral solution of a Stokes problem
(1984) J. of Numerical Methods in Fluids pp 1149-1163

[23] DUBOIS Th.
Simulation numérique d'écoulements homogènes et non-homogènes par des méthodes spectrales multi-résolution
(1993), Thèse Université Orsay - Paris Sud

[24] FLECK J.A., MORRIS J.R. et FEIT M.D.,
Time Dependent Propagation Of High Energy Laser Beams through the Atmosphere
(1976), Applied Physics, vol **10**, pp 129-160.

[25] FUNARO D.,
Polynomial Approximation of Differential Equations
(1992), Spriger Verlag, Berlin-Heildelberg

[26] GOTTLIEB D. et ORSZAG S.A,
Numerical Analysis of Spectral Methods : Theory and Applications
(1977), SIAM-CBMS **26**, Philadelphia

[27] GOTTLIEB D.,
Communication privée.

[28] HERBST B.M., MORRIS J.Ll. et MITCHELL A.R.,
Numerical experience with the nonlinear Schroedinger Equation
(1985), J. of Computational Physics, vol **60**, pp 282-305.

[29] HERBST B.M., et WEIDEMAN J.A.C.,
Split Step Methods for the solution of the Nonlinear Schroedinger Equation
(1986), SIAM J. of Numerical Analysis, vol **23** num **3**, pp 485-507.

[30] HOPF E.,
The Partial Differential Equation $u_t + uu_x = \mu u_{xx}$,
(1950) Communications in Pure and Applied Mathematics, vol **3**, pp 201-230.

[31] JAUBERTEAU F.,
Résolution numérique des équations de Navier-Stokes instationnaires par méthodes spectrales. Méthode de Galerkin non linéaire.
(1990), Thèse Université Orsay - Paris Sud

[32] KIM J., MOIN P., MOSER R.,
Turbulence statistics in fully developed channel flow at low Reynolds number
(1987) J. of Fluids Mechanics, vol **77**, pp 133-166.

[33] KREISS H.O. et LORENTZ J.,
Initial-Boundary Value Problems and the Navier-Stokes Equations
(1974), Ed. Academic Press.

[34] KREISS H.O. et OLIGER J.,
Stability of the Fourier Method
(1979) SIAM J. Numerical Analysis, vol **16**, pp 421-433.

[35] LABROSSE G.,
Communication privée.

[36] LASCAUX P. et THEODOR R.,
Analyse numérique matricielle appliquée à l'art de l'ingénieur, tomes 1,2
(1986,1987), Masson, Paris

[37] LAX M., BATTEH J.H. et AGRAWAL G.P.,
Channeling of intense electromagnetic beams
(1981) J. of Applied Physics, vol **52**, pp 109-125.

[38] LIGHTHILL,
Viscosity Effects in Sound Waves of Finite amplitude
(1956), Cambridge Univ. Press, Cambridge

[39] PEYRET R.,
Introduction to Spectral Methods
(1986) Van Karman Institute Lecture Series 1986-04, Rhode-St Genese, Belgique

[40] SANZ-SERNA I.M. et VERWER J.G.,
Conservative and nonconservative schemes for the solution of the nonlinear Schroedinger equation
(1986) IMA, J. of Numerical Analysis, vol **6**, pp 25-42

[41] SHEN J.,
Efficient spectral-Galrkin method I. Direct solvers for second- and fourth-order equations by using Legendre polynomials. (1994) SIAM, J. of Scientific Computing, vol **15**, pp 1489-1505.

[42] SHEN J.,
Efficient Chebyshev-Legendre Galerkin Methods for Elliptic Problems
(1996) Proceedings of the third ICOSAHOM, Houston Journal of Mathematics, University of Houston

[43] STRANG G.,
On the construction and comparison of difference schemes
(1968) SIAM, J. of Numerical Analysis, vol **5**, pp 506-517

[44] STRAUSS W.,
nonlinear Wave Equation
(1989), SIAM-CBMS **73**, Philadelphia

[45] SU C.H. GARDNER C.S.,
Korteweg-de-Vries Equation and Generalizations :
III Derivation of the Korteweg-de-Vries Equation and Burgers Equation
(1969) J. of Mathematical Physics, vol **10**, pp 536-539.

[46] TAHAR T.R. et ABLOWITZ M.J.,
Analytical and Numerical Aspects of certain Nonlinear Evolution Equations. II Numerical, Nonlinear Schroedinger Equation
(1984) J. of Computational Physics, vol **55**, pp 203-230

[47] TEMAM R.,
Navier-Stokes Equations, Theory and Numerical Analysis
(1977), Ed North Holland, Amsterdam

[48] TEMAM R.,
Navier-Stokes Equations and Nonlinear Functional Analysis
(1983) SIAM-CBMS **41**, Philadelphia

[49] TEMAM R.,
Infinite Dimensional Systems in Mechanics and Physics, (1988) Applied Mathematica Sciences **68**, Springer Verlag

[50] WILLIAMSON J.H.,
Low storage Runge-Kutta schemes (1980), J. of Computational Physics, vol **35**, pp 48-56

[51] YANG B., GOTTLIEB D. et HESTHAVEN J.S.,
Spectral simulations of Electromagnetic Wave Scattering (1997) J. of Computational Physics, vol **134**, pp 216-230

[52] YEE K.S.,
Numerical simulation of initial boundary value problems involving Maxwell's equations in isotropic media (1966) IEEE Antennas and Propagation, vol **14**, pp 302-307

[53] ZAKARIA A.,
Etude de divers schémas pseudo-spectraux de type collocation pour la résolution des équations aux dérivées partielles. Applications aux équations de Navier-Stokes (1985), Thèse Université de Nice

n^0 d'impression : 2012
$3^{ème}$ trimestre 1998

Guy Moebs :
Applications de méthodes spectrales multi-niveaux à différents problèmes de la physique mathématique.

Résumé:

L'objet de ce travail est l'étude de certains aspects des méthodes spectrales multi-niveaux appliquées à differents problèmes de la physique mathématique.

Dans un premier temps, nous considérons l'équation de Schroedinger non linéaire faiblement amortie comme modèle de la transmission de signaux binaires dans une fibre optique. Nous employons la méthode Fourier-Collocation associée au schéma en temps Split-Step Agrawal. Une étude du spectre de Fourier nous permet de mettre en évidence le rôle prépondérant de l'opérateur linéaire dans l'équation régissant les hautes fréquences. La méthode qui en résulte est validée pour différents niveaux de séparation entre basses et hautes fréquences et démontre sa capacité à calculer ces dernières de manière plus économique.

Ensuite nous étudions le traitement de conditions limites non périodiques pour une méthode spectrale à deux niveaux : la méthode Tau-Legendre appliquée au problème d'électromagnétisme 2D de la cavité résonnante. Nous considérons différentes discrétisations temporelles conservatives, explicites ou implicites. La non-commutativité des opérateurs spectraux de dérivation spatiale et de projection orthogonale sur les basses et les hautes fréquences, entraine un couplage des équations projetées, qui est renforcé par le caractère global des conditions limites. Un algorithme itératif de type Gauss-Seidel par blocs, sous-relaxé, nous permet de résoudre ce système. Nous menons alors une étude comparative des différents schémas.

Enfin, nous présentons des résultats obtenus avec l'équation de Burgers stochastique prise comme modèle de turbulence. Pour des conditions limites périodiques nous étudions l'évolution en temps des quantités moyennées issues des projections sur les grandes et les petites structures. Pour des conditions limites de non glissement, nous comparons les résultats obtenus par discrétisation spatiale spectrale avec ceux obtenus à l'aide d'une discrétisation par différences finies.

Mots clés :

méthodes spectrales – méthodes multi-niveaux – séparation des échelles – equation de Schroedinger – equation de Maxwell – equation de Burgers – projections orthogonales