HAL
open science

Étude théorique de méthodes numériques pour les systèmes de réaction-diffusion; application à des équations paraboliques non linéaires et non locales

Magali Ribot

To cite this version:

Magali Ribot. Étude théorique de méthodes numériques pour les systèmes de réaction-diffusion; application à des équations paraboliques non linéaires et non locales. Mathématiques [math]. Université Claude Bernard - Lyon I, 2003. Français. NNT : . tel-00004563

HAL Id: tel-00004563

https://theses.hal.science/tel-00004563

Submitted on 6 Feb 2004

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre : 216-2003 Année 2003

THÈSE

présentée

devant l'UNIVERSITÉ CLAUDE BERNARD-LYON 1

pour l'obtention

du DIPLÔME DE DOCTORAT

(arrêté du 25 avril 2002)

Spécialité : Mathématiques Appliquées

présentée et soutenue publiquement le 11 Décembre 2003

par

Magali RIBOT

Étude théorique de méthodes numériques pour les systèmes de réaction – diffusion ; application à des équations paraboliques non linéaires et non locales.

Directrice de thèse : Michelle Schatzman

JURY :	Mme	Sylvie Benzoni-Gavage,	Examinatrice
	M	Albert Cohen,	Rapporteur
	M	Frédéric Poupaud,	Rapporteur
	Mme	Michelle Schatzman,	Directrice de thèse
	M	Denis Serre,	Président
	M	Joël Sommeria,	Examinateur

Remerciements

Je tiens tout d'abord à exprimer ma gratitude à Michelle Schatzman pour avoir encadré cette thèse ; elle m'a guidée vers des sujets passionnants et grâce à ses connaissances étendues, elle m'a fait découvrir de nombreux liens insoupçonnés entre les diverses branches des mathématiques. Elle a toujours su me laisser une grande liberté, tout en répondant à mes nombreuses questions quand il le fallait. Enfin, c'est grâce à elle que j'ai pu découvrir New York, ses musées, ses casernes de pompiers, ses cheesecakes et ses mathématiques... Pour tout cela, merci.

Je remercie également Albert Cohen, David Gottlieb et Frédéric Poupaud d'avoir accepté d'être rapporteurs de cette thèse. Je suis très honorée de l'intérêt qu'ils ont porté à ce travail. Leur lecture très attentive de ce manuscrit a permis la correction de nombreuses imprécisions et une amélioration sensible de la rédaction.

Sylvie Benzoni a toujours su faire preuve de patience, de bonne humeur et de disponibilité et j'ai toujours eu plaisir à discuter avec elle, de mathématiques ou d'autres choses. Je suis très heureuse qu'elle ait accepté de faire partie de ce jury et l'en remercie vivement.

Je suis également très reconnaissante à Denis Serre d'avoir présidé le jury de cette thèse. J'ai toujours eu grand plaisir à suivre ses cours et ses exposés, tous d'une remarquable clarté.

Une des parties de cette thèse puise ses racines dans les travaux de Joël Sommeria ; je suis très honorée par sa participation au jury et l'en remercie.

Je tiens aussi à exprimer ma grande gratitude à Stéphane Descombes ; il a guidé mes premiers pas dans la recherche et s'est toujours montré très disponible. L'attention qu'il a portée à mon travail et ses conseils avisés m'ont aidée de nombreuses fois. Travailler avec lui fut toujours un plaisir, qui j'espère se renouvellera dans l'avenir.

L'enthousiasme communicatif de Carole Rosier a toujours été très stimulant. Je suis très heureuse d'avoir pu réfléchir à de beaux problèmes avec elle et la remercie sincèrement pour son aide précieuse.

Pierre-Henri Chavanis a également grandement participé à l'élaboration de cette thèse. Grâce à ses vastes connaissances, il a réussi à me faire comprendre de nombreux points de physique et je lui en suis extrêmement reconnaissante.

Je tiens également à remercier chaleureusement tous les membres du laboratoire MAPLY ; les discussions, mathématiques ou non, furent nombreuses et variées. Un immense merci à Christiane Battoue et Christiane Ussel pour leur aide administrative, ainsi qu'à Thierry Dumont et Violaine Louvet pour leur aide informatique. Je voudrais également remercier Marc Massot pour son énergie qui a permis la création d'une ACI "Nouvelles Interfaces des Mathématiques".

L'enseignement effectué au cours de ces trois années a été un véritable plaisir et m'a permis de clarifier ma vision de nombreux points d'analyse numérique. Je tiens en particulier à remercier Stéphane Génieys qui a été un tuteur pédagogique hors pair, ainsi que toute l'équipe enseignante avec laquelle j'ai travaillé.

Je voudrais exprimer ma reconnaissance à ceux qui ont relu certaines de ces pages et à ceux qui m'ont aidée à préparer la soutenance d'une manière ou d'une autre.

J'aimerais aussi remercier tous mes amis-thésards, en particulier Adrien, Clair, David, Michaël et Pascal.

Je voudrais également adresser mes plus chaleureux remerciements à toute ma famille et mes amis ; qu'ils aient accepté de partir en randonnée, au ski, à l'escalade, au cinéma, au bout du monde, qu'ils m'aient accueillie chez eux pour un goûter, pour un dîner, pour un week-end ou tout simplement qu'ils aient pris le temps d'un coup de fil ou d'un mail, ce furent à chaque fois d'excellents moments.

Julien mérite une mention particulière ; par sa gentillesse et ses encouragements de tous les instants, il a été d'une aide inestimable.

Je tiens finalement à remercier mes parents pour leur soutien sans faille. Je leur dois beaucoup et je n'oublie pas que mon intérêt pour les mathématiques a commencé rue Émile Biscara.

Table des matières

Introduction 1
1. Introduction 1
2. Étude du schéma par régularisation du résidu ("Residual Smoothing Scheme" ou "RSS") 4
3. Convergence du schéma de Peaceman – Rachford pour des systèmes de réaction – diffusion 15
4. Étude numérique de l'équation de réaction – diffusion d'Allen – Cahn et application à la croissance de grains 20
5. Mécanique statistique des systèmes auto – gravitants de fermions 25

Chapitre 1. Étude du schéma par régularisation du résidu 39
1. Rappels sur l'approximation des semi – groupes continus par des semi – groupes discrets 39
2. Rappels sur les extrapolations de Richardson 41
3. Stabilité du schéma par régularisation du résidu et de ses extrapolations 46
4. Stabilité, convergence et ordre du schéma par régularisation du résidu et de ses extrapolations en norme d'énergie 72
5. Application du schéma par régularisation du résidu au préconditionnement d'une méthode spectrale par une méthode d'éléments finis 95
6. Développement asymptotique des polynômes de Legendre et de leurs points critiques 127

Chapitre 2. Approximation de Peaceman – Rachford 157
1. Rappels sur les opérateurs sectoriels et leur lien avec les semi-groupes analytiques 157
2. Convergence du schéma de Peaceman-Rachford pour des systèmes de réaction – diffusion 160

Chapitre 3. Un système d'Allen – Cahn avec transition triple ; application numérique à la croissance de grains 181
1. Rappels sur les méthodes multipas explicites - implicites 181
2. Une modélisation de la croissance de grains par un système de réaction – diffusion 182

Chapitre 4. Mécanique statistique des systèmes auto – gravitants de fermions 191
1. Description de l'algorithme spectral de Gauss – Lobatto – Legendre pour le système de Smoluchowski – Poisson 191
2. Effondrements, explosions et hystérésis : étude numérique de systèmes auto-gravitants de fermions 196

Bibliographie 223

Introduction

1. Introduction

Dans cette thèse, nous étudions principalement différentes méthodes numériques pour résoudre les systèmes de réaction – diffusion et nous les appliquons à des exemples concrets.

Les systèmes de réaction – diffusion sont des problèmes d'un intérêt majeur, que l'on retrouve en dynamique des populations, en cinétique chimique et dans bien d'autres domaines. Ils donnent lieu à l'apparition de structures complexes, comme des ondes solitaires, des ondes spirales ou encore des structures de Turing, étudiées dans le livre de J. Murray [**86**].

1.1. Systèmes de réaction – diffusion. Nous considérons un système de réaction – diffusion de la forme

$$\begin{cases} \dfrac{\partial u}{\partial t} - M\Delta u + F(u) = 0, & x \in \mathbb{R}^d,\ t > 0, \\ u(0,x) = u_0(x), & x \in \mathbb{R}^d, \end{cases} \tag{1.1}$$

où d est la dimension d'espace, $1 \leq d \leq 3$, m est le nombre d'équations du système et le nombre de composantes du vecteur solution u, M est la matrice $m \times m$ de diffusion et $F : \mathbb{R}^m \to \mathbb{R}^m$ est la fonction de réaction.

La méthode la plus simple pour discrétiser ce système en temps est la méthode d'Euler explicite ; l'inconvénient de cette méthode est que, pour qu'elle soit stable, le pas de temps Δt et le pas d'espace Δx doivent vérifier une condition CFL (Courant – Friedrichs – Lewy) du type $\Delta t \leq C\Delta x^2$.

On peut s'affranchir de cette contrainte en utilisant un schéma d'Euler implicite, qui, lui, est inconditionnellement stable. En revanche, programmer ce schéma nécessite la résolution d'un grand système non – linéaire, approché par plusieurs grands systèmes linéaires et ceci est très coûteux. On a alors recours à un préconditionneur P, c'est – à – dire qu'au lieu de résoudre le système $Ax = b$, on résout le système $P^{-1}Ax = P^{-1}b$; P doit alors vérifier deux propriétés : il doit être facile à inverser et le conditionnement de la matrice $P^{-1}A$ doit être inférieur à celui de la matrice A.

Ces deux schémas ne sont pas satisfaisants ; M. Schatzman [**109**] a fait une revue complète d'autres méthodes existantes, de leurs avantages et de leurs limites. Dans un premier temps, nous considérons une classe de ces méthodes, les méthodes de splitting, appelées également méthodes de directions alternées.

1.2. Méthodes de splitting. Le principe des méthodes de splitting est de résoudre séparément les deux parties de l'équation que l'on cherche à approcher. Expliquons ce principe sur l'équation (1.1) en séparant la partie diffusive et la partie réactive.

Soit $X^t v_0 = v(t,.)$ le flot de la partie réactive

$$\begin{cases} \dfrac{\partial v}{\partial t} + F(v) = 0, \\ v(0,.) = v_0 \end{cases}$$

et $Y^t w_0 = w(t,.)$ le semi – flot de la partie diffusive

$$\begin{cases} \dfrac{\partial w}{\partial t} - M\Delta w = 0, \\ w(0,.) = w_0; \end{cases}$$

alors la formule de Lie définie par

$$L^t u_0 = X^t Y^t u_0 \tag{1.2}$$

est formellement une approximation d'ordre un en temps de la solution du système complet (1.1).

On peut également définir l'approximation de Strang [**117**] suivante

$$S^t u_0 = X^{t/2} Y^t X^{t/2} u_0, \tag{1.3}$$

qui, elle, est formellement une approximation d'ordre deux. Cette formule a été étudiée dans le cas de semi – groupes holomorphes par S. Descombes et M. Schatzman [**52**]. S. Descombes a également étudié son extrapolation pour des systèmes de réaction – diffusion [**49**].

Ces formules sont des premières approximations et il faut encore les discrétiser en temps et en espace pour pouvoir les programmer. Dans ce document, nous étudierons une approximation en temps d'ordre deux de la formule de Strang, le schéma de Peaceman – Rachford.

1.3. Méthodes de préconditionnement. Lorsqu'on est amené à considérer un schéma implicite, comme l'approximation de Peaceman – Rachford, on doit résoudre plusieurs grands systèmes, ce qui nécessite l'utilisation d'un préconditionneur.

Pour des problèmes elliptiques, la résolution du système linéaire provenant de la discrétisation spatiale a été beaucoup étudiée ; des références en la matière sont les livres de Y. Saad [**106**] et de G. Meurant [**84**]. Dans ce cas, les préconditionneurs sont utilisés pour augmenter le taux de convergence des méthodes itératives.

Les préconditionneurs eux – mêmes ont été aussi largement considérés, toujours dans le cadre de problèmes stationnaires, comme par exemple dans le livre de C. Canuto et al. [**23**] ; les méthodes d'éléments finis fournissent des préconditionneurs adaptés. Par exemple, S. Brenner [**20**] étudie le préconditionnement de méthodes d'éléments finis compliquées par des méthodes d'éléments finis plus simples, comme des méthodes P^1 ou P^2. Il existe également plusieurs méthodes pour préconditionner la discrétisation spectrale d'un problème elliptique par des discrétisations aux différences finies ou aux éléments finis ; nous les détaillerons dans ce document.

Dans le cas de problèmes d'évolution, comme les problèmes paraboliques, la plupart des auteurs cherchent un préconditionneur après avoir appliqué la méthode des lignes, c'est – à – dire après une semi – discrétisation en temps ; le préconditionneur considéré est alors un préconditionneur de la matrice $\mathbf{1} + \Delta t A$, où A est la matrice obtenue après discrétisation en espace de la partie diffusive et Δt est le pas de temps. C'est le cas, par exemple, de F. Bornemann [**17, 18**] qui préconditionne la matrice $\mathbf{1}+\Delta t A$ grâce à des méthodes multiniveaux ; dans [**21**], P. Brown et C. Woodward testent différentes stratégies de préconditionnement dans le cas d'un problème de radiation diffusion couplé à un transfert matériel – énergie et ils étudient notamment le préconditionnement par un opérateur "splitté". Enfin, L. Mulholland et D. Sloan [**85**] préconditionnent les matrices $\mathbf{1}+\Delta t A$ provenant de la discrétisation pseudospectrale en Fourier par des méthodes aux différences finies.

Dans ce document, nous allons étudier dans un premier temps un schéma pour les équations paraboliques où le préconditionneur en espace est directement inclus dans la discrétisation en temps. C'est le schéma par régularisation du résidu appelé en anglais "Residual Smoothing Scheme" ou "RSS" ; il s'inspire directement des itérations de Richardson [**102**] pour résoudre un système linéaire, définies comme suit. On cherche à résoudre le système linéaire $AU = F$ en considérant les itérations suivantes :

$$M\,(U_{k+1} - U_k) = \tau(F - AU_k), \tag{1.4}$$

où τ est un paramètre de relaxation à choisir convenablement et M est un préconditionneur de A. On connaît un paramètre τ qui minimise le taux de convergence de la méthode en fonction de la plus petite et de la plus grande valeur propre de la matrice $M^{-1}A$. La régularisation du résidu est un moyen classique d'accélération de la convergence mais aussi d'amélioration de la stabilité des méthodes d'intégration en temps [**124**].

1.4. Extrapolations de Richardson. Dans ce document, nous considérons également les extrapolations de Richardson du schéma par régularisation du résidu.

Une extrapolation de Richardson d'un schéma est une combinaison linéaire d'itérations de ce schéma appliqué pour différents pas de temps choisis au préalable ; elle est donc extrêmement simple à calculer, une fois le schéma de départ connu.

Les coefficients de la combinaison linéaire sont tels que l'extrapolation de Richardson soit d'ordre plus élevé que le schéma de départ. Ce qui rend ce procédé intéressant et efficace est le fait que ces coefficients sont des rationnels connus explicitement et ne dépendant que des pas de temps choisis.

En conclusion, considérer les extrapolations de Richardson est un moyen simple d'augmenter l'ordre d'un schéma, sans amener de difficultés de programmation supplémentaires. Nous expliquons plus précisément le principe de ces extrapolations à la section 2 du chapitre 1, en s'inspirant d'une version algébrique du résultat donnée dans l'article de B.O. Dia et de M. Schatzman [**55**].

1.5. Organisation de la thèse. La thèse est divisée en quatre chapitres : dans le premier chapitre, on s'intéresse aux propriétés du schéma par régularisation du résidu. Le deuxième chapitre est consacré à la convergence d'un schéma de splitting semi – discrétisé en temps, l'approximation de Peaceman – Rachford. Dans le troisième chapitre, nous étudions la croissance de grains et essayons de la simuler grâce à un système de réaction – diffusion. Enfin dans le quatrième chapitre, nous programmons un système d'équations paraboliques non – locales venant de la mécanique statistique.

Dans cette introduction, nous allons expliquer en détails ces différents résultats.

2. Étude du schéma par régularisation du résidu ("Residual Smoothing Scheme" ou "RSS")

Dans le premier chapitre, après un rappel sur les semi – groupes et un rappel sur les extrapolations de Richardson, nous nous intéressons au schéma par régularisation du résidu et à ses extrapolations. Tout d'abord, nous montrons dans un cadre général que le schéma RSS et ses extrapolations sont stables pour la norme usuelle ; nous prouvons également que ses extrapolations sont stables, convergentes et d'ordre voulu en norme d'énergie. Enfin, nous appliquons ce schéma au cas d'une méthode spectrale préconditionnée par une méthode d'éléments finis. Pour vérifier une hypothèse d'équivalence des matrices considérées, nous devons établir des développements asymptotiques des polynômes de Legendre. Ces résultats sont présentés respectivement dans les quatre articles suivants : [**100**], [**99**], [**97**] et [**98**].

Le schéma par régularisation du résidu a été introduit par A. Averbuch, A. Cohen et M. Israeli [**4**] pour des équations paraboliques linéaires. C'est un schéma de discrétisation en temps, utilisant un préconditionneur pour la discrétisation en espace. Écrivons donc ce schéma inspiré des itérations de Richardson que nous avons définies à l'équation (1.4).

Soient A un opérateur auto – adjoint et le système

$$\left\{\begin{array}{l} \dfrac{du}{dt} + Au = f(t), \\ u(0) = u_0. \end{array}\right. \tag{2.1}$$

Soient Δt le pas de temps et U_n une approximation de la solution au temps $t_n = n\Delta t$. Si B est un préconditionneur de A, le schéma RSS s'écrit

$$U_{n+1} - U_n + \Delta t A U_n + \tau \Delta t B(U_{n+1} - U_n) = \Delta t F_n, \tag{2.2}$$

où τ est un paramètre à choisir ultérieurement.

A. Averbuch, A. Cohen et M. Israeli [**4**] ont étudié et programmé ce schéma dans le cas où A est un opérateur issu de la décomposition multiéchelle en ondelettes et B son préconditionneur diagonal ; ils ont appliqué ce schéma pour le traitement d'image avec des résultats satisfaisants. Quant à nous, nous l'appliquerons dans le cas du préconditionnement d'une méthode spectrale par une méthode d'éléments finis.

L'intérêt de ce schéma est qu'il est implicite, ce qui garantit une bonne stabilité numérique ; cependant, on traite implicitement le préconditionneur B facile à inverser et explicitement l'opérateur initial A qui peut être mal – conditionné. En effet, en prenant $B = 0$, on retrouve le schéma d'Euler explicite pour l'opérateur A. Ce schéma est de plus le plus simple des schémas de Runge – Kutta préconditionnés mis en évidence par M. Schatzman [**110**].

2.1. Stabilité de RSS et de ses extrapolations. Nous faisons alors plusieurs hypothèses :

- A est un opérateur auto – adjoint, borné inférieurement ;
- B est un opérateur auto – adjoint, de même domaine que A ;
- A et B sont équivalents au sens des formes quadratiques, c'est – à – dire il existe $c > 0$ tel que

$$c^{-1}(Bu, u) \leq (Au, u) \leq c(Bu, u). \tag{2.3}$$

On notera cette relation $c^{-1}B \prec A \prec cB$ ou encore $A \sim B$.

L'opérateur $P(t)$ définissant le schéma est donné par la formule

$$\begin{aligned} P(t) &= (\mathbf{1} + t\tau B)^{-1} (\mathbf{1} + t\tau B - tA) \\ &= \mathbf{1} - t\,(\mathbf{1} + t\tau B)^{-1} A. \end{aligned}$$

Nous allons considérer les extrapolations de Richardson de ce schéma, évoquées au paragraphe 1.4. Soient des entiers $1 \leq n_1 < n_2 < \cdots < n_k$

et les polynômes d'interpolation de Lagrange ℓ_j^k associés aux nœuds $1/n_j$, c'est – à – dire que le polynôme ℓ_j^k est le polynôme de degré $k-1$ qui vaut 1 en $1/n_j$ et 0 en $1/n_l$, $l \neq j$;on définit la k-ième extrapolation de Richardson de P par

$$P_k(t) = \sum_{j=1}^{k} \ell_j^k(0) P(t/n_j)^{n_j}. \tag{2.4}$$

Nous prouvons alors la stabilité du schéma (2.2) et de ses extrapolations (2.4), c'est – à – dire nous montrons que pour τ assez grand, il existe $C > 0$ tel que

$$-(1+Ct)\mathbf{1} \prec P_k(t) \prec (1+Ct)\mathbf{1};$$

l'une des principales difficultés est que les opérateurs P et donc P_k ne sont pas auto – adjoints.

Voici les grandes lignes de la preuve.

Soient

$$\widetilde{P}(t) = \mathbf{1} - t\sqrt{A}\,(\mathbf{1} + t\tau B)^{-1}\sqrt{A}$$

et

$$\widetilde{P}_k(t) = \sum_{j=1}^{k} \ell_j^k(0) \widetilde{P}(t/n_j)^{n_j},$$

les formes symétrisées des opérateurs P et P_k.

La première étape consiste à prouver la stabilité de $\widetilde{P}_k$, c'est – à – dire à montrer que pour τ assez grand, pour tout $t > 0$, on a

$$-\mathbf{1} \prec \widetilde{P}_k(t) \prec \mathbf{1}. \tag{2.5}$$

L'étape suivante est de comparer la forme symétrique $\widetilde{P}_k$ et la forme asymétrique P_k afin d'étendre l'estimation (2.5) à P_k. Pour ce faire, nous avons besoin d'hypothèses algébriques supplémentaires.

Soit une algèbre $\mathcal{M}$ d'opérateurs bornés sur H telle que pour tout opérateur m de $\mathcal{M}$, le commutateur $[\sqrt{B}, m]$ soit aussi dans $\mathcal{M}$; par ailleurs, on suppose qu'il existe deux éléments m_1 et m_2 de $\mathcal{M}$ tels que

$$\sqrt{A} = \sqrt{B} m_1 + m_2$$

avec m_1 tel que

$$\sqrt{B}\left[m_1, [\sqrt{B}, m_1]\right] \in \mathcal{M}. \tag{2.6}$$

On peut voir l'algèbre $\mathcal{M}$ comme l'ensemble des opérateurs pseudo-différentiels de degré 0.

Posons $T_k = P_k - \widetilde{P}_k$. Sous les hypothèses précédentes, on prouve séparément que pour $t \in [0,1]$,

$$\|T_k\| \leq C\sqrt{t} \tag{2.7}$$

et que

$$\|\widetilde{P}_k^* T_k + T_k^* \widetilde{P}_k\| \leq Ct. \tag{2.8}$$

Nous remarquons ici l'importance des adjoints des opérateurs T_k et $\widetilde{P_k}$ qui nous permettent de considérer l'opérateur auto – adjoint $\widetilde{P}_k^* T_k + T_k^* \widetilde{P}_k$.

Nous utilisons la relation

$$P_k^* P_k = (\widetilde{P}_k + T_k)^*(\widetilde{P}_k + T_k) = \widetilde{P}_k^* \widetilde{P}_k + T_k^* T_k + \widetilde{P}_k^* T_k + T_k^* \widetilde{P}_k,$$

ainsi que les estimations (2.5), (2.7) et (2.8) pour prouver la stabilité de P_k, c'est – à – dire nous établissons le théorème suivant :

THÉORÈME 2.1. *Pour tout τ assez grand, il existe $C > 0$ tel que pour tout $t \in [0, 1]$ on ait*

$$-(1 + Ct)\mathbf{1} \prec P_k(t) \prec (1 + Ct)\mathbf{1}.$$

Plusieurs lemmes de majoration des normes de $\mathbf{1} - P(t)$, de $\mathbf{1} - \widetilde{P}(t)$, de la différence de leurs puissances et de divers commutateurs de leurs puissances sont nécessaires pour établir (2.7) et (2.8). Afin de faciliter les calculs, nous prouvons un lemme général, appelé lemme de sandwich, qui permet d'estimer un produit d'opérateurs grâce au nombre d'occurrences de l'opérateur b dans le produit. La preuve de ce lemme consiste à commuter des termes et à "négliger" les commutateurs grâce à l'hypothèse

$$\forall m \in \mathcal{M},\ [\sqrt{B}, m] \in \mathcal{M},$$

déjà énoncée plus haut. Nous donnons deux preuves du lemme de sandwich, une avec des manipulations algébriques et l'autre à l'aide de graphiques.

Nous remarquons que la non – commutativité de l'opérateur A et de son préconditionneur B rend l'étude plus difficile. Une revue de la non – commutativité en analyse numérique et des problèmes qu'elle engendre est faite par M. Schatzman [**110**].

Cependant, dans le cas d'un problème de Laplace avec conditions aux limites de Dirichlet homogènes, les tentatives faites pour vérifier numériquement l'hypothèse (2.6) ont échoué. L'opérateur A correspondait alors à la matrice de discrétisation par une méthode spectrale de Gauss – Lobatto – Legendre et l'opérateur B à un préconditionneur aux éléments finis calculé aux points de Gauss – Lobatto. L'hypothèse (2.6) était très bien vérifiée à l'intérieur du domaine, mais ne l'était pas aux bords.

Organisation de la section 3 du chapitre 1. Dans l'article [**100**], qui correspond à la section 3, nous commençons par établir des résultats algébriques liés aux polynômes de Lagrange et par prouver la stabilité

de l'opérateur symétrisé à la sous – section 3.3. Nous comparons ensuite les formes symétriques et asymétriques à la sous – section 3.4, en prouvant le lemme de sandwich.

2.2. Stabilité, convergence et ordre de RSS et de ses extrapolations en norme d'énergie. Nous cherchons à simplifier les démonstrations précédentes et pour cela, nous introduisons la norme d'énergie définie par

$$|x|_A = (x^* A x)^{1/2}.$$

L'avantage de cette norme est qu'elle rend l'opérateur P auto – adjoint. En effet, l'adjoint P^* de P est donné par la formule

$$(Px, y)_A = (x, P^* y)_A, \text{ soit } x^T P^T A y = x^T A P^* y.$$

Donc l'adjoint de P pour la norme d'énergie vaut $P^* = A^{-1} P^T A$, ce qui donne

$$P^* = A^{-1} \left(\mathbf{1} - tA \left(\mathbf{1} + t\tau B \right)^{-1} \right) A = \mathbf{1} - t \left(\mathbf{1} + t\tau B \right)^{-1} A = P$$

et donc P est bien auto – adjoint pour cette norme.

Ainsi, en calculant la norme d'opérateur de P associée à la norme d'énergie, nous obtenons

$$\|P(t)\|_A = \|A^{-1/2} P(t)^* A P(t) A^{-1/2}\|^{1/2}$$

et il suffit donc d'étudier l'opérateur $Q(t) = A^{-1/2} P(t)^* A P(t) A^{-1/2}$, qui est, lui aussi, auto – adjoint. Nous prouvons alors plus facilement que pour τ assez grand, la norme d'énergie de $P(t)$ est inférieure ou égale à 1. La preuve de la stabilité de l'extrapolation $P_k(t)$ en norme d'énergie pour τ assez grand est un petit peu plus compliquée et le théorème s'énonce ainsi

Théorème 2.2. *Soient A et B des opérateurs positifs auto – adjoints de H tels que $A \prec cB$ pour $c > 0$. Alors, il existe $\tau_0 > 0$ tel que pour tout $\tau \geq \tau_0$, on ait l'estimation suivante :*

$$\forall t > 0, \quad \|P_k(t)\|_A \leq 1.$$

Le fait que l'opérateur soit auto – adjoint simplifie considérablement les choses par rapport au paragraphe précédent.

Nous prouvons alors la convergence des méthodes issues de P et P_k pour τ assez grand en utilisant la théorie des approximations des semi – groupes continus par des semi – groupes discrets de T. Kato [**72**], théorie qui est rappelée à la section 1 du chapitre 1. Plus précisément, nous montrons que pour tout f assez régulier

$$\left(\mathbf{1} + \frac{\mathbf{1} - P_k(t)}{t} \right)^{-1} f \xrightarrow[t \to 0]{s} (\mathbf{1} + A)^{-1} f$$

grâce au Théorème de Lax – Milgram et à des estimations en norme d'énergie. Nous avons alors le résultat de convergence suivant :

THÉORÈME 2.3. *Si $A \sim B$, il existe $\tau_0 > 0$ tel que pour tout $\tau \geq \tau_0$,*

$$\left\| P_k(t_n)^n - e^{-tA} \right\|_A \xrightarrow[\substack{n\to\infty \\ nt_n \to t}]{s} 0.$$

Enfin, nous montrons que le schéma (2.2) est d'ordre un et les extrapolations (2.4) d'ordre k. Pour cela, nous utilisons un résultat général de B. O. Dia et M. Schatzman [**55**] sur les extrapolations de Richardson et nous montrons le théorème suivant, où $\mathcal{H}_A^p$ est le domaine de $A^{p/2}$:

THÉORÈME 2.4. *Soit $\mathcal{H}_A^p = D(A^{p/2})$ et supposons que $\mathcal{H}_B^{2k+2}$ et $\mathcal{H}_A^{2k+2}$ sont isomorphes. Alors, il existe $p \in \mathbb{N}$, $\tau_0 > 0$ et $t_0 > 0$ tels que $\forall u \in \mathcal{H}_A^p$, $\forall \tau \geq \tau_0$ et $\forall T > 0$, il existe une constante $C(|u|_{\mathcal{H}_A^p}, T, \tau)$ dépendant de $|u|_{\mathcal{H}_A^p}$, T et τ telle que $\forall t \in (0, t_0]$ et $\forall n$ tels que $nt \leq T$,*

$$\left| P_k(t)^n u - e^{-ntA} u \right|_A \leq C(|u|_{\mathcal{H}_A^p}, T, \tau) t^k.$$

Organisation de la section 4 du chapitre 1. Dans l'article [**99**] présenté à la section 4, nous donnons des résultats préliminaires aux sous – sections 4.2 et 4.3 avec notamment la définition et les propriétés de l'équivalence de deux opérateurs. La stabilité du schéma est montrée à la sous – section 4.4 et celle des extrapolations à la sous – section 4.5 . Puis, la stabilité conditionnelle est étudiée à la sous – section 4.6, la convergence à la sous – section 4.7 et enfin l'ordre à la sous – section 4.8.

2.3. Application au préconditionnement d'une méthode spectrale par une méthode d'éléments finis. Les méthodes spectrales sont d'une grande efficacité numérique mais les matrices engendrées sont pleines et mal conditionnées ; en effet, le conditionnement de la matrice spectrale de taille N est généralement en $O(N^3)$. C'est pourquoi, il est nécessaire de considérer des préconditionneurs appropriés pour rendre ces méthodes efficaces.

Le préconditionnement des méthodes spectrales par des méthodes aux différences finies ou d'éléments finis a déjà été largement étudié dans le cas d'un problème elliptique : dès 1980, S. Orszag [**87**] propose un préconditionnement par une méthode aux différences finies ; il prouve l'équivalence des matrices de discrétisation correspondant à ces deux méthodes dans le cas de conditions aux limites périodiques et d'une base de Fourier. Cette équivalence a aussi été étudiée par P. Haldenwang et al. [**63**] dans le cas d'une méthode spectrale de Chebychev. C. Canuto et A. Quarteroni [**24**] ont testé différents préconditionnements d'une méthode de Chebychev par une méthode des éléments finis en donnant des estimations numériques des rayons spectraux. Finalement, M. Deville et E. Mund [**53**] ont comparé numériquement différentes méthodes d'éléments finis pour le même type de problème et ont étendu dans [**54**] cette étude pour des polynômes de Jacobi, plus généraux que les polynômes de Chebychev.

Un point important est que les méthodes aux différences finies ou d'éléments finis considérées sont calculées avec comme points de discrétisation les points des méthodes spectrales. Nous nous plaçons dans le cas de conditions aux bords de Dirichlet et nous considérons ici une méthode spectrale de collocation de Gauss – Lobatto – Legendre. Expliquons comment on la définit.

En dimension 1, les nœuds sont donnés par la formule de Gauss – Lobatto, qui nous sert de formule de quadrature et qui est rappelée ici ([**9**], par exemple). Soit un entier N ; on pose $\xi_0 = -1$ et $\xi_N = 1$. Il existe alors un unique ensemble de $N-1$ nœuds ξ_j de $[-1,1]$ et un unique ensemble de $N+1$ poids ρ_j tels que pour tout polynôme Φ de degré inférieur ou égal à $2N-1$, on ait l'égalité suivante

$$\int_{-1}^{1} \Phi(x)dx = \sum_{j=0}^{N} \Phi(\xi_j)\rho_j. \tag{2.9}$$

Les ξ_j sont les zéros des dérivées des polynômes de Legendre L_N et nous pouvons exprimer explicitement les nœuds ρ_k, qui sont des réels strictement positifs :

$$\begin{aligned} \rho_0 = \rho_N &= \frac{2}{N(N+1)}, \\ \rho_k &= \frac{2}{N(N+1)L_N^2(\xi_k)},\ 1 \leq k \leq N-1. \end{aligned}$$

La formule (2.9) nous donne donc à la fois les nœuds de discrétisation et la formule de quadrature que l'on exprime sous la forme du produit scalaire discret suivant :

$$(u,v)_N = \sum_{0 \leq k \leq N} u(\xi_k)v(\xi_k)\rho_k. \tag{2.10}$$

Il ne nous reste plus qu'à choisir une base pour décomposer la fonction solution ; la base des polynômes interpolateurs de Lagrange aux points ξ_k est une base adaptée qui simplifie beaucoup les expressions rencontrées, et notamment la matrice de masse associée à cette méthode spectrale.

Expliquons brièvement ce que sont les matrices de masse et de rigidité sur l'exemple de l'équation de la chaleur

$$\frac{\partial u}{\partial t} - \Delta u = 0, \tag{2.11}$$

avec conditions aux bords de Dirichlet.

En ayant décomposé la solution u sur une base Φ_k de fonctions qui s'annulent aux bords, fonctions que l'on prend également comme

fonctions test, la formulation variationnelle discrétisée de (2.11) s'écrit :

$$\forall k \in \{1, \cdots, N-1\},$$

$$\frac{\partial}{\partial t} \sum_{j=1}^{N-1} u_j(t)(\Phi_j, \Phi_k)_N + \sum_{j=1}^{N-1} u_j(t)(\nabla\Phi_j, \nabla\Phi_k)_N = 0, \tag{2.12}$$

c'est – à – dire sous forme matricielle

$$M\frac{\partial U}{\partial t} + KU = 0,$$

où $U(t)$ est le vecteur $(u_1(t), \cdots, u_{N-1}(t))$, M est ce qu'on appelle la matrice de masse avec $M_{i,j} = (\Phi_i, \Phi_j)_N$ et K est la matrice de rigidité avec $K_{i,j} = (\nabla\Phi_i, \nabla\Phi_j)_N$. Nous remarquons qu'il y a une matrice de masse et une matrice de rigidité par entier N et pour plus de clarté, nous omettrons l'indice N dans la suite.

Dans le cas de la méthode spectrale de Gauss – Lobatto – Legendre et du produit scalaire (2.10), le choix des polynômes de Lagrange l_j associés aux nœuds ξ_j comme base est particulièrement bien adapté ; en effet, on obtient alors

$$M(i,j) = (l_i, l_j)_N = \delta_{i,j}\rho_i \tag{2.13}$$

et

$$K(i,j) = (l_i', l_j')_N = \sum_{0 \le k \le N} l_i'(\xi_k) l_j'(\xi_k)\rho_k. \tag{2.14}$$

La matrice de masse est donc extrêmement simple, puisqu'elle est diagonale.

Si M_S (resp. M_F) et K_S (resp. K_F) désignent les matrices de masse et de rigidité de la méthode spectrale (resp. des éléments finis), on définit $A = M_S^{-1/2} K_S M_S^{-1/2}$ et $B = M_F^{-1/2} K_F M_F^{-1/2}$. Pour appliquer les résultats précédents sur le schéma par régularisation du résidu aux matrices spectrales et d'éléments finis, il faut prouver que l'hypothèse d'équivalence (2.3) est vérifiée pour A et B.

Pour pouvoir calculer facilement l'inverse de la racine carrée de la matrice tridiagonale M_F, nous la diagonalisons en utilisant le procédé de "mass – lumping". Ce procédé consiste à approcher les intégrales définissant les coefficients de M_F par une méthode des trapèzes. Le "mass – lumping" est un procédé largement utilisé en analyse numérique. Nous obtenons alors la matrice de masse diagonale définie par les coefficients suivants :

$$(M_F)_{i,j} = \delta_{i,j}\frac{\xi_{i+1} - \xi_{i-1}}{2}. \tag{2.15}$$

Grâce à l'équivalence entre K_S et K_F montrée par S. Parter et E. Rothman [**92**], il suffit de prouver que $M_S^{1/2} M_F^{-1/2} K_F M_F^{-1/2} M_S^{1/2}$ et K_F

sont équivalentes. Soit la norme H^1 discrète définie par

$$\|U\|_{\mathbb{H}_N^1} = (U^* K_F U)^{1/2} = \left(\sum_{k=0}^{N-1} \frac{|U_{k+1} - U_k|^2}{\xi_{k+1} - \xi_k} \right)^{1/2};$$

il suffit alors de montrer que la matrice $M_F^{-1/2} M_S^{1/2}$ est uniformément bornée par rapport à N en norme H^1 discrète. Pour ce faire, nous avons besoin de deux estimations.

D'une part, soit σ_k les coefficients de la matrice diagonale $M_F^{-1} M_S$ donnés par

$$\sigma_k = \frac{2\rho_k}{\xi_{k+1} - \xi_{k-1}}; \tag{2.16}$$

nous montrons que ces coefficients sont bornés indépendamment de k et de N, en utilisant des estimations de C. Bernardi et Y. Maday [**9**] sur les points de discrétisation et sur les poids de la méthode de Gauss – Lobatto – Legendre.

D'autre part, nous rappelons que le carré de la norme H^1 discrète de la matrice $M_F^{-1/2} M_S^{1/2}$ est donné par

$$\max \left\{ \sum_{k=0}^{N-1} \frac{\left| \sqrt{\sigma_{k+1}} U_{k+1} - \sqrt{\sigma_k} U_k \right|^2}{\xi_{k+1} - \xi_k} : \|U\|_{\mathbb{H}_N^1} = 1 \right\}. \tag{2.17}$$

Soit $\lfloor x \rfloor$ le plus grand entier inférieur ou égal à x. Pour établir une estimation uniforme de cette norme, il nous faut majorer indépendamment de N la somme

$$\Sigma_N = \sum_{k=0}^{\lfloor (N-1)/2 \rfloor} \mu_k, \tag{2.18}$$

où

$$\mu_k = \frac{(2 - |\xi_{k+1}| - |\xi_k|)}{\xi_{k+1} - \xi_k} \left| \frac{1}{\sigma_{k+1}} - \frac{1}{\sigma_k} \right|^2, \tag{2.19}$$

avec la convention $1/\sigma_0 = 1/\sigma_N = 0$.

La preuve de ce résultat repose sur des développements précis des polynômes de Legendre L_N et des zéros de leurs dérivées, développements qui seront détaillés à la section suivante.

Dans [**89**], S. Parter obtient des asymptotiques des polynômes de Legendre et de leurs extrema et les utilise dans [**90, 91**] pour prouver l'équivalence considérée. Nous utilisons, quant à nous, des asymptotiques différentes, qui sont par ailleurs obtenues d'une autre façon.

Le résultat de cette équivalence peut être étendu à la dimension 2 en tensorisant les points de discrétisation pour obtenir une grille de discrétisation bidimensionnelle. Puis, nous coupons les rectangles ainsi obtenus selon une de leur diagonale pour obtenir deux triangles

et considérer des éléments finis P^1. Les matrices s'expriment en fonction de produits tensoriels des matrices calculées en dimension 1. La généralisation à la dimension 2 se fait alors de manière tout à fait naturelle.

Des simulations numériques effectuées dans le cas de conditions aux limites de Dirichlet, vérifient l'équivalence des matrices prouvée précédemment. D'autres simulations mettent en évidence le rôle du paramètre τ dans la stabilité et confirment numériquement l'ordre des extrapolations. Enfin, nous calculons l'erreur de la méthode dans le cas où les solutions de l'équation sont de régularités différentes comme C^∞, C^1 non C^2 ou C^2 non C^3.

Des simulations dans un autre cas de conditions aux bords, comme des conditions de type Robin, ont montré qu'il était nécessaire de traiter plus finement les conditions aux limites.

Organisation de la section 5 du chapitre 1. Dans l'article [**97**], qui fait l'objet de la section 5, nous prouvons tout d'abord l'équivalence des matrices de masse en norme L^∞ à la sous – section 5.3. Ensuite, à la sous – section 5.4, nous prouvons l'équivalence de ces mêmes matrices de masse en norme H^1 discrète. Nous généralisons les résultats précédents à la dimension 2 à la sous – section 5.5 et finalement nous présentons des résultats numériques de RSS et de ses extrapolations à la sous – section 5.6.

2.4. Développement asymptotique des polynômes de Legendre et de leurs points critiques. Nous avons donc vu à la sous – section précédente que pour montrer l'équivalence entre les matrices de masse en norme H^1 discrète, nous avions besoin de développements précis des extrema des polynômes de Legendre L_N. La littérature classique sur les polynômes orthogonaux, et notamment le livre de G. Szegő [**118**], nous fournit des asymptotiques qui ne couvrent pas toute la zone qui nous intéresse. Alors que L. Gatteschi [**61**] utilise une méthode de Sturm pour trouver des asymptotiques des extrema, nous adaptons une méthode de la phase stationnaire à une formule de représentation des polynômes de Legendre.

Soit $N' = \lfloor (N-1)/2 \rfloor$. Expliquons tout d'abord pourquoi nous découpons l'intervalle $\{0, \cdots, N'\}$ en trois régions $\{0, \ldots, K\}$, $\{K, \Lambda N\}$ et $\{\Lambda N, N'\}$ et donnons les résultats disponibles dans le livre de G. Szegő [**118**]. Un premier résultat de G. Szegő nous dit que si z appartient à un intervalle borné et si N tend vers l'infini,

$$L_N\left(\cos\frac{z}{N}\right) \sim J_0(z),$$

où J_0 est la fonction de Bessel d'ordre 0. Un résultat analogue sur L'_N nous donne le développement suivant

$$\pi - \operatorname{Arccos}\xi_k = \frac{z_k}{N} + o(1/N),$$

où z_k est le k-ième zéro de la fonction de Bessel J_1 d'ordre 1 (Théorème 8.1.2 du livre de G. Szegő [**118**]). Par ailleurs, ces développements sont suffisants ; en effet, nous ne considérons d'abord qu'un nombre fini de μ_k pour $k \in \{0, \cdots, K\}$ et il suffit donc de montrer que μ_k admet une limite quand N tend vers l'infini .

Dans la troisième zone $\{\Lambda N, \ldots, N'\}$, nous utilisons le développement des polynômes ultrasphériques $P_N^{(\lambda)}(\cos\theta)$ donné à la formule (8.21.14) de G. Szegő [**118**] ; les polynômes de Legendre sont reliés aux polynômes ultrasphériques par la relation $L_N = P_N^{(1/2)}$ et la dérivée k-ième de L_N est proportionnelle à $P_{N-k}^{(k+1/2)}$. Plus précisément, étant donné

$$\omega_{N,\lambda} = \binom{N+\lambda-1}{N} = \frac{\Gamma(N+\lambda)}{\Gamma(N+1)\Gamma(\lambda)};$$

le développement avec reste uniforme en N et en $\theta \in [\Lambda/2, \pi/2]$ est

$$\begin{aligned} P_N^{(\lambda)}(\cos\theta) = \frac{2\omega_{N,\lambda}}{(2\sin\theta)^\lambda} \sum_{\nu=0}^{p-1} \omega_{\nu,\lambda} \frac{(1-\lambda)\ldots(\nu-\lambda)}{(N+\lambda-1)\ldots(N+\lambda-\nu)} \\ \times \frac{\cos\big((N-\nu+\lambda)\theta - (\nu+\lambda)\pi/2\big)}{(2\sin\theta)^\nu} + O(N^{\lambda-p-1}). \end{aligned} \tag{2.20}$$

Nous devons calculer le développement des zéros des dérivées des polynômes de Legendre pour $k \in \{\lfloor \Lambda N \rfloor, \cdots, N'\}$ et nous utilisons pour cela un théorème quantitatif des fonctions implicites de P. de Mottoni et M. Schatzman [**47**] et nous obtenons alors le lemme suivant :

Lemme 2.5. *Soit*

$$\theta_{0,k} = \frac{\pi/4 + k\pi}{N + 3/2}.$$

Alors $\forall \Lambda \in (0, 1/2)$, $\exists C$, C' *tels que* $\forall N \geq 2$ *et* $\forall k \in \{\lfloor \Lambda N \rfloor, \ldots, \lceil (1-\Lambda)N \rceil\}$, *il existe un unique zéro* θ_k *de* $P_N^{(3/2)}(\cos\theta)$ *dans une boule de rayon* C'/N^2 *autour de* $\theta_{0,k}$ *; de plus, on a*

$$\left| \theta_k - \theta_{0,k} + \frac{3}{8N^2 \tan\theta_{0,k}} - \frac{9}{8N^3 \tan\theta_{0,k}} \right| \leq CN^{-4}.$$

Compte tenu de la relation $L'_N = P_{N-1}^{(3/2)}$, il nous suffit d'appliquer ce lemme pour trouver les développements uniformes des zéros de L'_N.

Il reste donc à étudier la deuxième zone $\{K, \cdots, \lfloor \Lambda N \rfloor\}$, pour laquelle nous devons calculer un développement des polynômes de Legendre et de leurs dérivées. Nous partons d'une représentation intégrale des polynômes ultrasphériques (4.10.3) du livre de G. Szegő [**118**] donnée par

$$P_N^{(\lambda)}(x) = \frac{2^{1-2\lambda}}{(\Gamma(\lambda))^2} \frac{\Gamma(N+2\lambda)}{N!} \int_0^\pi \left(x + i\sqrt{1-x^2}\cos\varphi\right)^N \sin^{2\lambda-1}\varphi \, d\varphi.$$

Nous appliquons alors une méthode de la phase stationnaire comme celle de L. Hörmander [**66**] ; cependant, contrairement à la phase de L. Hörmander qui était du type $Nf(\varphi)$, elle est dans notre cas de la forme $f(N,\varphi)$, ce qui induit une difficulté supplémentaire.

Nous obtenons alors le résultat suivant :

THÉORÈME 2.6. *Étant donné $\lambda = p + 1/2$, où $p \in \mathbb{N}$, il existe des polynômes réels $Q_{\nu,\lambda}$ de degré ν tels que $\forall k \in \mathbb{N}$, $\forall K \in \mathbb{N}$ et $\forall \Lambda \in (0, 1/2)$, on ait $\forall N \geq 2$ et $\forall z \in [\pi K, \pi\Lambda N]$:*

$$\left| P_N^{(\lambda)}(\cos(z/N)) - 2\sqrt{\pi} Z(\lambda, N) \mathrm{Re}\left(ie^{iz} \sum_{\nu=\lambda-1/2}^{k-1} \chi_N^{-(\nu+1/2)} Q_{\nu,\lambda}(\chi_N/N) \right) \right|$$
$$\leq C(K, \Lambda, k, \lambda) \left(N^{-1} + z^{-1} \right)^{k-2\lambda+1},$$

où $\chi_N = -iN \sin(z/N) e^{-iz/N}$ et où

$$Z(\lambda, N) = \frac{2^{1-2\lambda}}{(\Gamma(\lambda))^2} \frac{\Gamma(N + 2\lambda)}{N!}.$$

L'application de ce théorème permet d'obtenir des asymptotiques de L_N et de ses trois premières dérivées en calculant explicitement les coefficients des polynômes $Q_{\nu,\lambda}$. Enfin, nous calculons le développement des extrema des polynômes de Legendre, toujours grâce au théorème des fonctions implicites.

Organisation de la section 6 du chapitre 1. Dans l'article [**98**], correspondant à la section 6, nous expliquons à la sous – section 6.2 la preuve des asymptotiques des extrema des polynômes de Legendre dans la première région. Nous donnons ensuite à la sous – section 6.3 les asymptotiques des polynômes de Legendre et de leurs extrema pour la deuxième région.

3. Convergence du schéma de Peaceman – Rachford pour des systèmes de réaction – diffusion

Dans le chapitre 2, nous nous intéressons à un schéma de splitting semi – discrétisé en temps, le schéma de Peaceman – Rachford pour des systèmes de réaction – diffusion. Nous montrons que le schéma converge et est d'ordre deux en temps et nous appliquons ce schéma à deux exemples de systèmes d'équations au comportement complexe.

Le chapitre 2 se décompose en deux sections, un rappel sur les opérateurs sectoriels et leur application aux semi – groupes analytiques et l'article [**51**].

Nous avons vu dans le paragraphe 1.2 que les formules de Lie et de Strang sont des premières approximations et qu'il faut encore les discrétiser en temps et en espace pour pouvoir les programmer.

Dans le cadre des systèmes de réaction – diffusion, nous nous intéressons ici à une discrétisation en temps de la formule de Strang (1.3) ;

cette discrétisation a été introduite par D. Peaceman et H. Rachford [**94**] pour résoudre l'équation de la chaleur et est définie par

(3.1)
$$P^t u_0 = (\mathbf{1} + tF/2)^{-1} (\mathbf{1} + tM\Delta/2) (\mathbf{1} - tM\Delta/2)^{-1} (\mathbf{1} - tF/2) u_0.$$

Il est à noter que le terme d'erreur entre la formule de Lie (1.2) et la solution exacte s'exprime en fonction du commutateur de $-M\Delta$ et F et celui entre la formule de Strang et la solution exacte en fonction des doubles commutateurs. La non – commutativité engendre donc une perte de l'ordre, mais aussi une perte de stabilité, comme on peut le voir dans l'article de M. Schatzman [**107**].

Cependant, dans le cas linéaire, M. Schatzman [**107**] a aussi montré la stabilité de

$$P(t) = (\mathbf{1} + tA/2)^{-1} (\mathbf{1} - tB/2) (\mathbf{1} + tB/2)^{-1} (\mathbf{1} - tA/2) \tag{3.2}$$

pour des opérateurs A et B positifs auto – adjoints dont le commutateur des racines carrées $\sqrt{A}$ et $\sqrt{B}$ est dominé par $\sqrt{A} + \sqrt{B}$; on peut, par exemple, voir A et B comme des opérateurs pseudodifférentiels d'ordre deux.

L'intérêt de l'approximation (3.1) est qu'elle est implicite aussi bien sur la partie linéaire que sur la partie non – linéaire, ce qui donne une bonne stabilité numérique. De plus, ce schéma résout séparément les parties linéaires et non – linéaires, qui ne se traitent pas de la même manière lorsqu'elles sont implicites. Enfin, c'est une approximation d'ordre deux en temps et la convergence est prouvée sans l'hypothèse M auto – adjointe, mais avec l'hypothèse $\operatorname{Re}\sigma(M) > 0$.

3.1. Convergence et ordre du schéma de Peaceman – Rachford. Nous prouvons ici la stabilité, la convergence et l'ordre de ce schéma dans le cadre des systèmes de réaction – diffusion.

Pour cela, nous commençons par faire quelques hypothèses nécessaires dans la suite ; en particulier, nous supposons que F est une fonction de classe $\mathrm{C}^{\mathbf{7}}$ dont les dérivées jusqu'à l'ordre 7 sont bornées et telle que $F(0) = 0$. Nous supposons également que M est une matrice carrée dont le spectre est inclus dans le demi – plan complexe $\{\operatorname{Re} z > 0\}$. Enfin nous supposons que la condition initiale u_0 est dans l'espace $L^2(\mathbb{R}^d)^m \cap L^\infty(\mathbb{R}^d)^m$.

La méthode employée pour prouver la convergence suit une idée de B. O. Dia et M. Schatzman [**57**]. Tout d'abord, nous calculons algébriquement une formule d'erreur dans le cas linéaire entre l'approximation (3.2) et la solution exacte $\exp(-t(A + B))$. Pour cela, nous utilisons la formule de Duhamel suivante : si V vérifie

$$\frac{\partial V}{\partial t} + AV = f,$$

la formule de Duhamel pour la solution V s'écrit alors

$$V(t) = e^{-tA}\,V(0) + \int_0^t e^{-(t-s)A}\,f(s)\,ds. \tag{3.3}$$

Ainsi, nous calculons formellement $\dot{P}(t) + (A+B)P(t)$ que l'on peut mettre sous la forme

$$\frac{\partial P}{\partial t}(t) + (A+B)P(t) = \frac{t^2}{4}S(t/2)$$

où S fait intervenir le commutateur de A et B, $[A,B]$ et le double commutateur $[B,[A,B]]$; l'expression du terme S est donnée explicitement à l'équation (2.12) du chapitre 2. En utilisant alors la formule de Duhamel (3.3), nous trouvons l'égalité suivante

$$P(2t) - \exp(-2t(A+B)) = 2\int_0^t s^2 e^{-2(t-s)(A+B)}S(s)\,ds,$$

qui est le principal résultat linéaire.

Ensuite, nous validons ces calculs algébriques par des estimations analytiques dans le cas où $A = V(x)\times\mathbf{1}$, où V est un potentiel borné et $B = -M\Delta$. Nous utilisons, entre autres, le fait que l'opérateur B est sectoriel et nous obtenons que, pour $s\in[0,6]$,

$$\|P(t) - \exp(-t(A+B))\|_{\mathcal{L}(H^s,L^2)} = O\left(t^{s/2}\right). \tag{3.4}$$

Finalement, nous passons au cas non – linéaire. Le premier théorème est le résultat de stabilité suivant :

Théorème 3.1. *Il existe une constante $C_0 > 0$ telle que pour t assez petit et $\forall (u_0,v_0)\in L^2(\mathbb{R}^d)^m\times L^2(\mathbb{R}^d)^m$, on ait*

$$\left|P^t u_0 - P^t v_0\right|_{L^2} \le (1+C_0 t)\,|u_0 - v_0|_{L^2}. \tag{3.5}$$

Il se montre en utilisant le fait que l'opérateur

$$L(t) = (\mathbf{1} + tM\Delta/2)(\mathbf{1} - tM\Delta/2)^{-1}$$

est de norme inférieure ou égale à 1 et le fait que F est une fonction lipschitzienne.

La preuve de l'estimation de la norme de $L(t)$ est assez astucieuse, on en donne une idée ici. Soit la matrice symétrique définie positive donnée par

$$S = \int_0^{+\infty} e^{-sM^*}e^{-sM}ds;$$

elle vérifie la relation suivante

$$M^*S + SM = \mathbf{1}. \tag{3.6}$$

On utilise alors le produit scalaire $\langle x,y\rangle = (Sx,y)$ dans $\mathbb{R}^m$, grâce auquel on écrit les normes L^2 de $L(t)u$ et de u en passant en Fourier.

On obtient alors, grâce aux symétries de l'opérateur $L(t)$, l'expression suivante

$$|L(t)\,u|_{L^2}^2 - |u|_{L^2}^2 = -2t \int_{\mathbb{R}^d} v^*(\xi)(M^*S + SM)|\xi|^2 v(\xi)\, d\xi$$

où v est défini à l'équation (2.29) du chapitre 2 ; cette relation nous donne le résultat voulu en utilisant la relation (3.6).

Le second théorème important est le résultat de convergence suivant où T^t est la solution exacte du système (1.1) :

THÉORÈME 3.2. *Pour u_0 dans $H^6(\mathbb{R}^d)^m \cap C^6(\mathbb{R}^d)^m$ dont les six premières dérivées sont bornées et pour t assez petit, on a l'estimation suivante :*

$$\left|P^t u_0 - T^t u_0\right|_{L^2} \leq C(|u_0|_{6,\infty}) t^3 \left|u_0\right|_{H^6}. \tag{3.7}$$

Ce théorème se prouve en utilisant le résultat (3.4) du cas linéaire et grâce à une formule de Taylor dont on connaît le reste.

Du Théorème 3.2 valable pour un pas de temps, nous déduisons le résultat général pour n pas de temps ; ceci prouve l'ordre deux de l'approximation :

THÉORÈME 3.3. *Pour tout u_0 dans $H^6(\mathbb{R}^d)^m \cap C^6(\mathbb{R}^d)^m$ dont les six premières dérivées sont bornées et $\forall \tau > 0$, il existe $C(|u_0|_{6,\infty})$ et h_0 tels que $\forall h \in (0, h_0]$, $\forall n$ tel que $nh \leq \tau$*

$$\left|\left(P^h\right)^n u_0 - T^{nh} u_0\right|_{L^2} \leq C(|u_0|_{6,\infty}) h^2 \left|u_0\right|_{H^6}.$$

3.2. Applications numériques au système de taches de léopard et à l'équation de Ginzburg – Landau. Nous avons tout d'abord testé notre méthode pour le système suivant avec les paramètres $a > 0$, $b > 0$, $\alpha > 0$, $\gamma > 0$, $\rho > 0$, $K > 0$ et $d > 1$:

$$\begin{cases} \dfrac{\partial u}{\partial t} = \Delta u + \gamma \left(a - u - \dfrac{\rho u v}{1 + u + K u^2} \right), \\ \dfrac{\partial v}{\partial t} = d \Delta v + \gamma \left(\alpha(b - v) - \dfrac{\rho u v}{1 + u + K u^2} \right). \end{cases} \tag{3.8}$$

Ce système a été introduit par J. Murray [**86**] pour modéliser la formation de motifs et notamment la formation de taches de léopard. C'est un modèle de système de Turing, où une solution stationnaire stable du système sans diffusion se déstabilise sous l'effet de la diffusion.

Nous avons représenté à la figure 1 l'évolution des isovaleurs du module de la solution avec comme condition initiale quatre pics. Nous constatons l'apparition de motifs similaires à des taches de léopard.

Nous avons également programmé l'équation quintique de Ginzburg – Landau ; cette équation a des solutions intéressantes appelées pulses, mises en évidence par O. Thual et S. Fauve [**121**] ; ces solutions ont été étudiées théoriquement par P. de Mottoni et M. Schatzman [**47**].

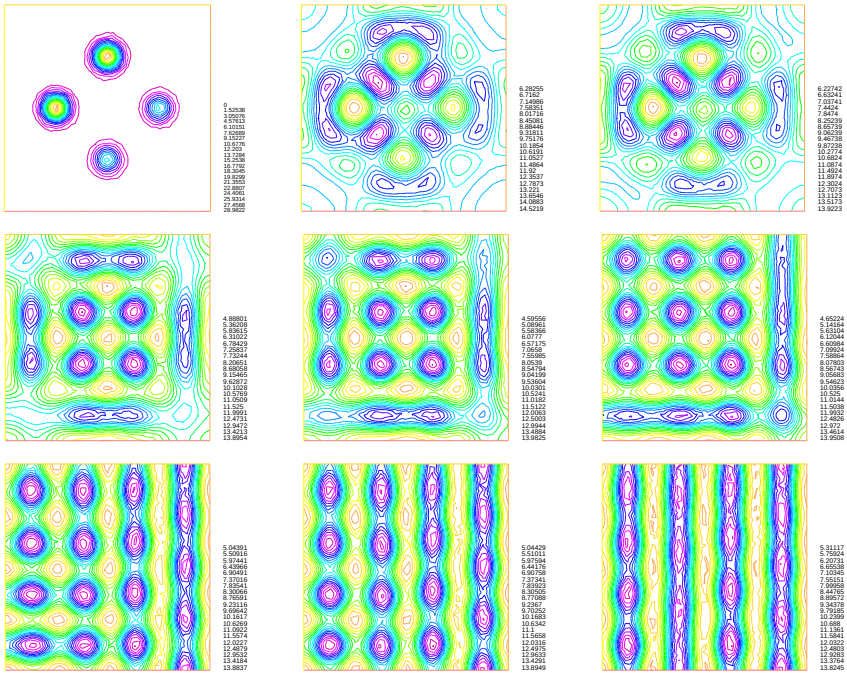


FIG. 1. Évolution de la solution du système (3.8), avec quatre pics comme condition initiale.

C'est une équation dans le plan complexe qui se met sous la forme d'un système (1.1) en séparant la partie réelle de la partie imaginaire.

À la figure 2, nous avons simulé l'évolution de la rencontre de deux pulses. Nous voyons qu'après une phase transitoire, la solution se stabilise en un pulse.

3.3. Organisation de la section 2 du chapitre 2. À la sous – section 2.2, nous établissons tout d'abord une expression algébrique de l'erreur entre l'approximation de Peaceman – Rachford (3.2) et la solution exacte $\exp(-t(A+B))$, dans le cas où les deux opérateurs sont linéaires.

Puis, à la sous – section 2.3, nous validons analytiquement le résultat de la sous – section précédente 2.2 quand $A = V(x) \times \mathbf{1}$ et $B = -M\Delta$.

À la sous – section 2.4, nous prouvons la stabilité du schéma (3.1) de Peaceman – Rachford puis nous nous servons du résultat obtenu à la sous – section précédente 2.3 pour prouver un résultat similaire dans le cas non – linéaire. Enfin, nous prouvons la convergence du schéma .

À la dernière sous – section 2.5, nous présentons les résultats numériques obtenus pour le système conduisant à la formation de taches de léopard et pour l'équation de Ginzburg – Landau.

FIG. 2. Évolution de la solution de l'équation quintique de Ginzburg – Landau, avec deux pulses comme condition initiale.

4. Étude numérique de l'équation de réaction – diffusion d'Allen – Cahn et application à la croissance de grains

La croissance de grains est un problème important de la science des matériaux pour comprendre en particulier les structures polycristallines. C'est pourquoi, de nombreux modèles de cette croissance ont été étudiés numériquement.

Ce chapitre se décompose en deux sections ; la première est consacrée à des rappels sur les méthodes multipas et la seconde à l'article [**101**].

Dans la section 2 du chapitre 3, nous passons en revue les modélisations déjà utilisées pour la croissance de grains. Ensuite, nous expliquons le phénomène de transition propre aux équations de réaction – diffusion et nous l'utilisons pour simuler la croissance de grains assimilés à des surfaces. Nous expliquons enfin quelles sont les méthodes numériques employées et nous donnons les principaux résultats numériques.

D. Kinderlehrer et C. Liu [**73**] ont étudié théoriquement la croissance de grains grâce au principe de croissance par courbure contrôlée. Dans cette méthode, les grains sont des polygones et on fait évoluer leurs frontières et leurs points de jonction ; seules les jonctions triples

sont considérées car ce sont les plus stables, les autres jonctions se transformant rapidement en un certain nombre de nouvelles jonctions triples. Pour faire évoluer les frontières elles – mêmes, on utilise l'équation de Mullins (2.1) du chapitre 3 qui relie la vitesse normale de la frontière à la tension de surface et à la courbure ; la condition de Herring (2.2) du chapitre 3 permet de trouver la position d'équilibre des jonctions triples. D'un point de vue théorique, D. Kinderlehrer et C. Liu [**73**] ont montré que ces équations possédaient une solution à condition que la donnée initiale soit assez proche d'un état stationnaire. Par ailleurs, de toute suite d'états ainsi obtenus, on peut extraire une sous – suite qui converge vers un état d'équilibre, qui n'est pas nécessairement celui proche de la condition initiale.

Du principe de croissance par courbure contrôlée, on peut déduire une méthode numérique basée sur les discrétisations de l'équation de Mullins et de la condition de Herring [**74, 75**]. Mais l'algorithme utilisé doit aussi tenir compte d'événements critiques tels que la disparition d'un côté de la frontière ou la disparition d'un grain. Par ailleurs, ce schéma est explicite compte tenu de sa complexité et le pas de temps utilisé doit donc être très petit, ce qui induit des temps de calcul assez longs.

Les autres méthodes habituellement considérées sont des méthodes statistiques [**119**] ; on peut utiliser une méthode de Monte – Carlo sur un réseau de points auxquels on associe un spin par point et que l'on modifie au hasard selon une probabilité de transition. La frontière de grains passe ensuite entre les spins d'orientations différentes. Cet algorithme est cependant très coûteux car il nécessite de discrétiser l'intérieur des grains. Une dernière méthode consiste à programmer une évolution déterministe basée sur la règle du $n-6$, où n est le nombre de côtés du grain ; cette règle dit que l'évolution temporelle de l'aire d'un grain en dehors des événements critiques est linéaire en $n-6$. On fait disparaître aléatoirement des côtés ou des grains selon des lois de probabilité calculées à l'avance ; l'inconvénient est que pour obtenir ces lois de probabilité, il faut tout d'abord programmer la croissance de grains grâce à une autre méthode, par exemple celle basée sur la croissance par courbure contrôlée.

4.1. Système d'Allen – Cahn et évolution d'interface. Compte tenu de ces difficultés, il serait donc intéressant d'avoir une équation d'évolution simple qui modéliserait la croissance de grains et que l'on pourrait programmer en utilisant une méthode implicite ou semi – implicite.

Soit Φ un potentiel pair avec deux puits en ± 1 d'égales profondeurs et notons $\phi = \Phi'$ sa dérivée. Soit $h > 0$ un petit paramètre fixé, alors

l'équation d'Allen – Cahn s'écrit

$$\frac{\partial u}{\partial t} - h^2 \Delta u + \phi(u) = 0. \tag{4.1}$$

La solution u de cette équation est proche de 1 en certains endroits, de -1 en d'autres et développe des interfaces entre les deux.

P. de Mottoni et M. Schatzman [**46, 45**] ont montré que cette équation admettait une solution dont on pouvait avoir un développement asymptotique à tout ordre à condition que la condition initiale soit assez proche d'une "transition" entre -1 et 1. On peut également obtenir un développement asymptotique d'une hypersurface caractérisant cette transition au temps t, pour t assez petit et il a été montré que la vitesse de déplacement de cette hypersurface est normale et proportionnelle à la courbure moyenne. Enfin entre les temps $O(\log(1/h))$ et $O(1/\sqrt{h})$, la solution de (4.1) est arbitrairement proche d'une transition sur un voisinage de l'interface Γ et arbitrairement proche de ± 1 hors de ce voisinage. Ce résultat est particulièrement important, car il précise l'intervalle de temps sur lequel l'évolution a lieu.

Nous nous proposons donc d'utiliser une version bidimensionnelle de cette équation avec un potentiel complexe de la forme

$$W(u) = |u-a|^2 |u-b|^2 |u-c|^2,$$

où a, b et c sont les sommets d'un triangle équilatéral. Soit ε un petit paramètre ; nous écrivons l'équation sous la forme

$$\frac{\partial u}{\partial t} - \varepsilon \Delta u + \frac{1}{\varepsilon} DW(u)^T = 0, \tag{4.2}$$

où DW désigne la différentielle du potentiel W et nous utilisons des conditions de Neumann aux bords. Nous nous attendons à ce que la solution u de cette équation développe des interfaces entre des régions $u \sim a$, $u \sim b$ et $u \sim c$, régions qui seront considérées comme des grains.

4.2. Méthodes de programmation et résultats numériques. Nous programmons donc l'équation (4.2) grâce à des méthodes multipas ou de splitting adaptées aux équations de réaction – diffusion en temps et grâce à des schémas aux différences finies en espace. Pour la discrétisation temporelle par méthodes de splitting, nous utilisons une méthode de Strang d'ordre deux et nous avons deux possibilités : $S_1^t = D^{t/2} R^t D^{t/2}$ ou $S_2^t = R^{t/2} D^t R^{t/2}$ où R^t est le flot de la partie réactive et D^t le flot de la partie diffusive. D'après les calculs de S. Descombes et M. Massot [**50**], dans le cas d'un problème raide, c'est – à – dire ε petit dans l'équation (4.2), il est préférable d'utiliser la formule S_2^t plutôt que la formule S_1^t.

À la figure 3, nous présentons les résultats obtenus à partir d'une condition initiale aléatoire pour $\varepsilon = 1$ jusqu'au temps d'évolution $T = 10$. La figure 4 correspond à $\varepsilon = 0.1$ et $T = 30$.

Tout d'abord, les simulations pour $\varepsilon = 1$ et $\varepsilon = 0.1$ donnent des résultats assez similaires, si ce n'est le temps de formation des interfaces, bien plus grand pour $\varepsilon = 0.1$. On remarque également que l'on n'obtient pas tout à fait les grains espérés, car ils ne sont pas polygonaux ; le problème est peut être que la condition de Herring qui relie les forces de ligne des trois frontières n'est pas prise en compte, alors que c'est une condition importante dans le processus de croissance de grains. La condition initiale aléatoire pose aussi sûrement un problème car jusque là, que ce soit dans les preuves théoriques ou dans les simulations numériques, elle a toujours été choisie proche d'un état d'équilibre. Cependant, on peut être satisfait du changement d'échelle qui apparaît clairement dans nos simulations.

Enfin, à la figure 5, nous représentons les résultats obtenus pour $\varepsilon = 0.01$ et $T = 900$. Nous remarquons alors une nette amélioration car quelques grains deviennent polygonaux. Nous espérons qu'un ε plus petit permettra des résultats encore meilleurs. Il est à noter que pour les ε petits, l'équation (4.2) devient raide et il faut donc assurer une bonne stabilité numérique. De plus, le temps de formation des interfaces augmente quand ε est petit et il est donc préférable de prendre un schéma d'ordre élevé qui permet des pas de temps plus grands.

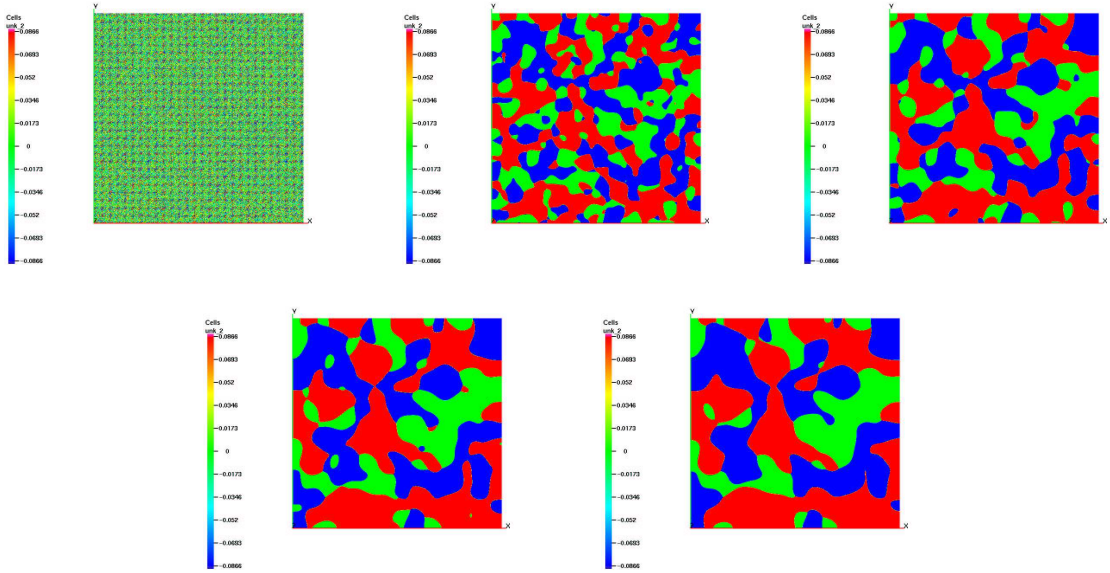


FIG. 3. Cas $\varepsilon = 1$. Évolution du système aux temps $t = 0.1$, $t = 2.5$, $t = 5$, $t = 7.5$ et $t = 10$.

4.3. Organisation de la section 2 du chapitre 3. Dans la section 2, nous présentons les différentes méthodes déjà utilisées pour programmer la croissance de grains à la sous – section 2.1. À la sous – section 2.2 , nous expliquons les principaux résultats de P. de Mottoni

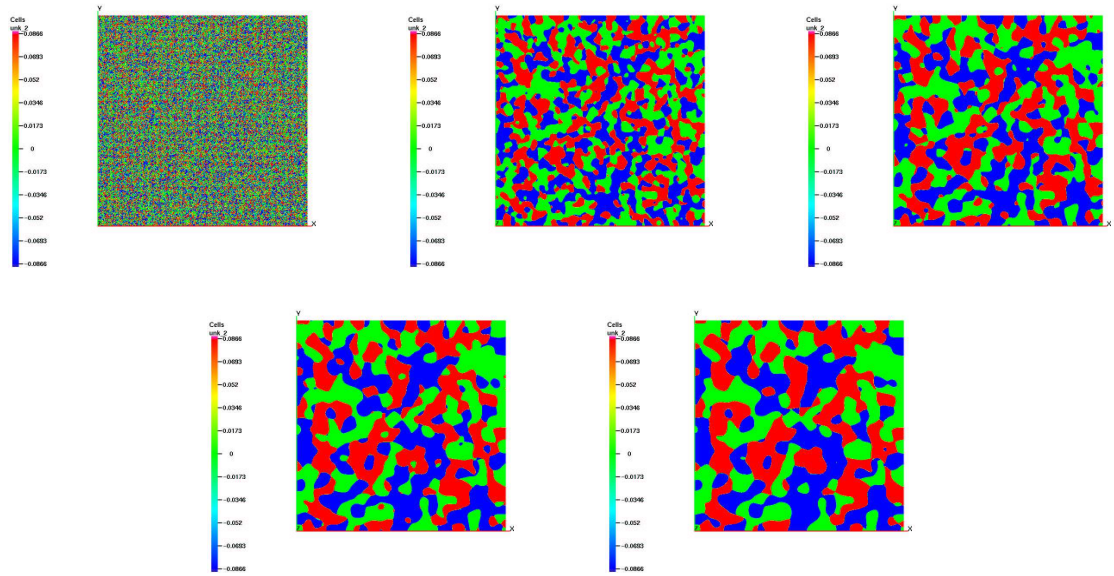


FIG. 4. Cas $\varepsilon = 0.1$. Évolution du système aux temps $t = 0.1$, $t = 7.5$, $t = 15$, $t = 22$ et $t = 30$.

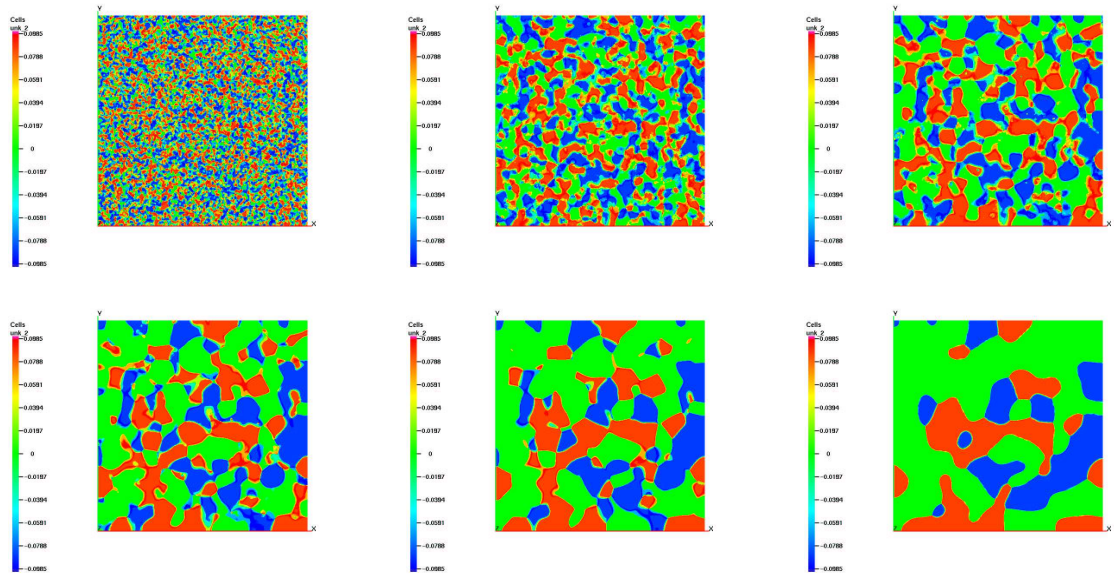


FIG. 5. Cas $\varepsilon = 0.01$. Évolution du système aux temps $t = 10$, $t = 90$, $t = 200$, $t = 300$, $t = 400$ et $t = 900$.

et M.Schatzman [**46, 45**] sur l'évolution d'interface. Finalement, à la sous – section 2.3, nous expliquons les méthodes numériques sous – jacentes au code que nous avons utilisé et nous présentons nos résultats numériques à la sous – section 2.4.

5. Mécanique statistique des systèmes auto – gravitants de fermions

Nous nous intéressons également dans cette thèse à la programmation d' un système venant de la mécanique statistique, un système de fermions auto – gravitants où l'on tient compte des interactions à longue portée. La mécanique statistique permet d'étudier les états asymptotiques de ce système mais donne peu de renseignements dynamiques. C'est pourquoi, nous cherchons à modéliser l'évolution de ce système par des équations non – stationnaires. Les modèles considérés ont également des applications directes en biologie, en chimiotactisme des populations bactériologiques [**86**]. Par exemple, les bactéries *E. Coli* sécrètent une substance qui a un effet d'attraction à longue portée sur les autres organismes et on peut observer des phénomènes semblables à ceux de la mécanique statistique.

Dans ce chapitre, nous passons tout d'abord en revue les différentes équations modélisant les systèmes auto – gravitants de particules ponctuelles et leurs liens et nous expliquons les phénomènes prédits par des méthodes statistiques théoriques. Ensuite, nous généralisons les modèles évoqués au cas de particules quantiques, les fermions, et nous détaillons les différences de comportements entre le cas classique et le cas fermionique. Puis nous choisissons parmi tous ces modèles le modèle le plus simple à programmer. Nous expliquons les méthodes numériques employées et les résultats obtenus. En particulier, nous vérifions que les solutions asymptotiques de l'équation considérée sont en accord avec les prédictions des états d'équilibre statistique, ce qui nous permet de valider le choix de cette équation. Enfin, les simulations numériques de l'équation nous donnent également des renseignements dynamiques qui sont difficiles à obtenir théoriquement. Ces résultats sont expliqués en détails dans [**34**].

Ces équations appartiennent à une nouvelle classe d'équations de dérive – diffusion généralisée introduite par Chavanis [**31**]. Elles s'écrivent

$$\frac{\partial \rho}{\partial t} = \vec{\nabla}\left(\frac{1}{\xi}\left(p'(\rho)\vec{\nabla}\rho + \rho\vec{\nabla}\int u\left(|\vec{r}-\vec{r'}|\right)\rho(\vec{r'},t)d^3\vec{r'}\right)\right),$$

où $p(\rho)$ est une équation d'état quelconque et u un potentiel d'interaction binaire. L'importance et la richesse de ces équations sont considérables et elles apparaissent dans divers domaines de la physique, l'astrophysique, l'hydrodynamique ou de la biologie. Elles sont associées à un formalisme de thermodynamique généralisée [**31, 39, 41**] étendant celui de Boltzmann, Fermi et Tsallis.

5.1. Systèmes auto – gravitants de particules classiques. Soit un système de N particules en interaction gravitationnelle. Nous décrivons ce système de façon statistique et en utilisant l'approximation

champ moyen quand N tend vers l'infini, nous trouvons des modèles continus et non discrets. Les systèmes d'équations différentielles ordinaires modélisant les problèmes discrets à N corps seraient a priori plus faciles à programmer que les équations aux dérivées partielles des modèles continus ; cependant, pour avoir une modélisation correcte, il faudrait un très grand nombre d'équations différentielles ordinaires dans le système et c'est pourquoi il paraît plus simple de programmer les modèles continus.

Soit $f(\vec{r}, \vec{v}, t)$ la fonction de distribution des particules à la position $\vec{r}$ avec la vitesse $\vec{v}$ et au temps t.

La densité de particules, ou d'étoiles ici, est alors donnée par l'intégrale

$$\rho(\vec{r}, t) = \int f(\vec{r}, \vec{v}, t) d^3\vec{v} \tag{5.1}$$

et la masse totale du système par

$$M = \int \rho(\vec{r}, t) d^3\vec{r}. \tag{5.2}$$

Soit Φ le potentiel gravitationnel vérifiant l'équation de Poisson qui le relie à la densité ρ

$$\Delta\Phi = 4\pi G\rho; \tag{5.3}$$

l'énergie s'écrit alors

$$E = \frac{1}{2}\int f v^2 d^3\vec{r} d^3\vec{v} + \frac{1}{2}\int \rho\Phi d^3\vec{r}. \tag{5.4}$$

Nous nous plaçons tout d'abord dans le cas microcanonique, c'est – à – dire que le système évolue à masse M et énergie E fixées. Pour trouver l'état d'équilibre statistique, on maximise l'entropie de Boltzmann

$$S = -\int f \ln f d^3\vec{r} d^3\vec{v} \tag{5.5}$$

sous les contraintes M et E constantes. En introduisant des multiplicateurs de Lagrange, on peut écrire

$$\delta S - \beta\delta E - \alpha\delta M = 0, \tag{5.6}$$

dont la solution est la distribution de Boltzmann :

$$f = Ae^{-\beta(v^2/2+\Phi)}, \tag{5.7}$$

où β est l'inverse de la température T à la constante de Boltzmann près.

On retrouve la même distribution dans le cas canonique, c'est – à – dire lorsque la masse M et la température T sont fixées en minimisant cette fois – ci l'énergie libre $F = E - TS$.

Nous nous intéressons à présent à une étude dynamique des systèmes auto – gravitants et nous nous attendons à ce que les états asymptotiques des solutions des équations considérées soient des états d'équilibre statistique.

Un premier modèle continu du problème à N corps est le modèle de Landau donné à l'équation (2.11) du chapitre 4 ; il prend en compte les rencontres proches entre les étoiles. En revanche, sa simplification aux temps courts, l'équation de Vlasov, néglige ces rencontres ; $\vec{F}$ désigne la force gravitationnelle, c'est – à – dire $-\vec{\nabla}\Phi$, l'équation de Vlasov s'écrit alors

$$\frac{\partial f}{\partial t} + \vec{v}\frac{\partial f}{\partial \vec{r}} + \vec{F}\frac{\partial f}{\partial \vec{v}} = 0. \tag{5.8}$$

L'état asymptotique de l'équation de Vlasov pour une condition initiale donnée peut être différent de l'état asymptotique de l'équation de Landau pour la même condition initiale. Mathématiquement parlant, la différence entre les états asymptotiques de ces deux équations vient de la non commutativité des limites quand $t \to +\infty$ et $N \to +\infty$. Le modèle de Landau correspond au cas où l'on passe d'abord à la limite sur le temps t, alors que le modèle de Vlasov apparaît lorsque l'on passe d'abord à la limite sur le nombre de particules N. L'équation de Landau est adaptée pour les amas globulaires et l'équation de Vlasov pour l'étude des galaxies. Cependant, ces équations sont compliquées et nous considérons ici d'autres modèles plus simples présentant le même type de comportements.

Un autre modèle possible est le modèle de Kramers – Chandrasekhar donné par des équations stochastiques en milieu homogène qui modélisent les collisions par un mouvement brownien des étoiles. Ce modèle peut se déduire de l'équation de Landau sous l'hypothèse du bain thermique, c'est – à – dire l'hypothèse que toutes les particules sont à l'équilibre sauf une qui évolue dans ce bain thermique ; cependant, il ne conserve pas l'énergie.

Dans le cas d'un milieu non – homogène, l'équation à considérer est un "mélange" des équations de Landau et de Kramers – Chandrasekhar, où la partie collisionnelle de Landau est simplifiée grâce au modèle stochastique. On obtient ainsi l'équation de Kramers – Poisson :

$$\frac{\partial f}{\partial t} + \vec{v}\frac{\partial f}{\partial \vec{r}} + \vec{F}\frac{\partial f}{\partial \vec{v}} = \frac{\partial}{\partial \vec{v}}\left(D\frac{\partial f}{\partial \vec{v}} + \xi f \vec{v}\right). \tag{5.9}$$

Cette équation décrit un système de particules auto – gravitantes (P. H. Chavanis, C. Rosier et C. Sire [**40**]). Lorsqu'on passe à la limite haute friction $\xi \to +\infty$, on obtient l'équation de Smoluchowski – Poisson obtenue dans les articles de P. H. Chavanis, J. Sommeria, R. Robert [**37**]

et de P. H. Chavanis [**31**] :

$$\frac{\partial \rho}{\partial t} = \nabla \left\{ \frac{1}{\xi} \left(T \nabla \rho + \rho \nabla \Phi \right) \right\}, \tag{5.10}$$

que l'on considérera dans la suite.

La courbe calorique ou diagramme d'équilibre, c'est – à – dire l'inverse de la température en fonction de l'opposé de l'énergie adimensionnée pour les états d'équilibre statistique, est représentée à la figure 6.

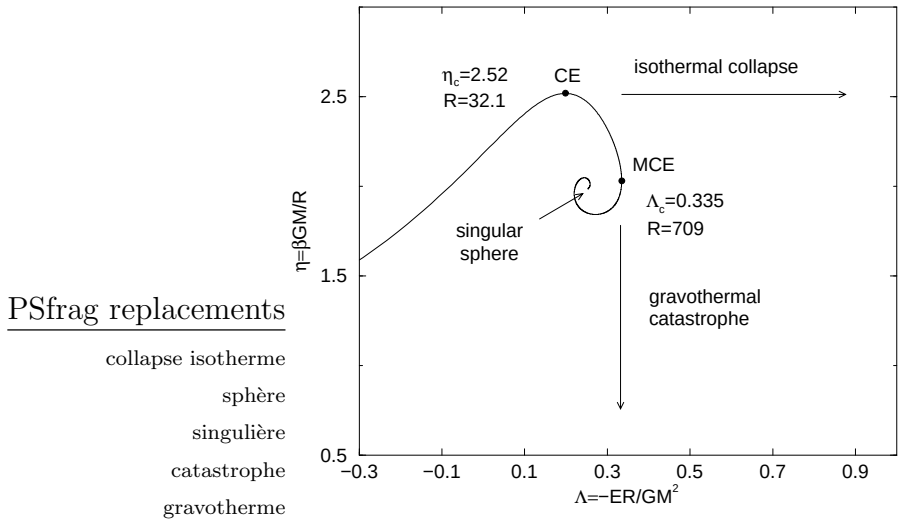


FIG. 6. Courbe calorique pour un gaz auto – gravitant de particules classiques.

Tout d'abord dans le cas microcanonique (M et E fixées), il n'existe pas d'état d'équilibre stable en – dessous d'une énergie critique E_c [**2**]. Dans ce cas, le système subit une catastrophe gravotherme ; le cœur du système devient de plus en plus dense et chaud jusqu'à l'effondrement (D. Lynden – Bell et R. Wood [**83**]).

Ce problème de maximisation d'entropie de Boltzmann à masse et énergie fixées n'admet pas de maxima globaux ; en revanche, il existe des maxima locaux stables qui sont situés sur la branche supérieure de la courbe de la figure 6 jusqu'au point MCE.

Dans le cas canonique (M et T fixées), W. Bonnor [**16**] a noté qu'il n'existait pas d'état d'équilibre stable en – dessous d'une température critique T_c ; ceci est un résultat statistique visible à la figure 6. Ces états statistiques sont généralement les états asymptotiques des modèles dynamiques et dans le cas où l'équilibre statistique n'existe pas, nous observons donc dynamiquement un "collapse isotherme", c'est – à – dire mathématiquement une explosion de la solution en temps fini. L'évolution du collapse est visible à la figure 7 et on remarque qu'il est autosimilaire (P. H. Chavanis, C. Rosier et C. Sire [**40**]).

Les points de la figure 6 sur la branche supérieure jusqu'au point CE sont des minima locaux stables de l'énergie libre ; les points de la branche inférieure sont des points critiques instables de l'énergie libre ; au dessus de la valeur critique β_c, il n'y a plus de points critiques de l'énergie libre.

On remarque que sur la courbe de la figure 6 certains points sont stables microcanoniquement mais instables canoniquement. Les deux ensembles canoniques et microcanoniques ne sont donc pas équivalents.

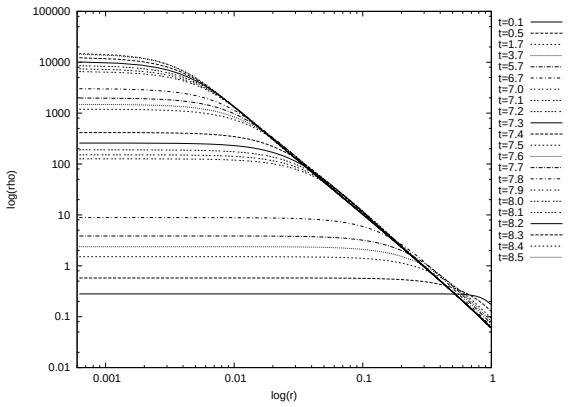


FIG. 7. Collapse isotherme à la température T=0.39 dans le cas classique. Ces courbes représentent la densité en fonction du rayon à différents temps ; ici, la densité centrale croît avec le temps.

Mathématiquement, les systèmes auto – gravitants ont été étudiés dans le cas canonique par P. Biler [**11**] et P. Biler et T. Nadzieja [**13**] sous la forme d'une unique équation aux dérivées partielles (5.26) donnée plus loin et équivalente au système de Smoluchowski – Poisson ; dans le premier de ces deux articles, l'auteur prouve l'existence de solutions autosimilaires et de solutions stationnaires et donnent des conditions sur la donnée initiale pour obtenir l'explosion, au sens mathématique du terme. Dans le deuxième article, les auteurs montrent sous certaines conditions l'existence et la non – existence de solutions globales en temps. Dans le cas microcanonique, C. Rosier [**105**] a étudié le système de Smoluchowski – Poisson ; elle a prouvé l'existence locale en temps et l'unicité de solutions faibles et, sous l'hypothèse d'une condition initiale radiale, la positivité de la densité et la croissance de l'entropie au cours du temps. Par ailleurs, toujours dans le cas microcanonique, P. Biler et T. Nadzieja [**14**] ont prouvé la non – existence globale en temps des solutions pour des énergies négatives. Enfin, P. H. Chavanis, C. Rosier et C. Sire [**40, 112, 113, 41**] fournissent des développements analytiques des solutions autosimilaires, confirmés par des simulations numériques.

5.2. Cas des systèmes auto – gravitants de fermions. Pour définir une phase condensée non singulière, c'est – à – dire une densité très concentrée près de 0 en une structure cœur – halo et qui n'explose pas, nous devons modifier un peu les modèles précédemment étudiés. Il y a plusieurs manières de le faire : nous pouvons considérer le cas des sphères dures, qui, contrairement aux particules, ont un volume ; alors, il est impossible d'agglomérer une infinité de particules en un point, ce qui stoppe le collapse isotherme. Nous pouvons également utiliser un potentiel gravitationnel régularisé comme dans l'article de P. H. Chavanis et I. Ispolatov [**33**] ou bien nous pouvons étudier un système de fermions, particules quantiques qui suivent le principe d'exclusion de Pauli.

Ces fermions ont potentiellement des applications importantes en astrophysique, notamment pour la compréhension des naines blanches et des étoiles à neutrons. Ils ont déjà été étudiés d'un point de vue physique notamment par P. H. Chavanis [**29**], qui a obtenu les diagrammes de phase présentés aux figures 8 et 9. Cette étude a été généralisée au cas avec rotation par P. H. Chavanis et M. Rieutord [**35**].

Introduisons donc l'entropie de Fermi – Dirac

$$S = -\int \left(f \ln f + (\eta_0 - f)\ln(\eta_0 - f)\right) d^3\vec{r} d^3\vec{v}, \tag{5.11}$$

où η_0 est la valeur maximale que peut prendre la fonction de distribution f et est donnée par le principe d'exclusion de Pauli.

En maximisant l'entropie S donnée par l'équation (5.11) à masse et énergie fixées, nous obtenons la distribution de Fermi – Dirac

$$f = \frac{\eta_0}{1 + \lambda e^{\beta(v^2/2+\Phi)}}. \tag{5.12}$$

Elle peut aussi être obtenue dans le cas canonique en minimisant l'énergie libre $F = E - TS$ à masse et température fixées.

On peut alors étendre les modèles présentés à la sous – section précédente 5.1 au cas fermionique. On définit ainsi des équivalents fermioniques, mis en évidence par P. H. Chavanis [**31**], de l'équation de Landau et de l'équation (5.9) de Kramers qui devient

$$\frac{\partial f}{\partial t} + \vec{v}\frac{\partial f}{\partial \vec{r}} + \vec{F}\frac{\partial f}{\partial \vec{v}} = \frac{\partial}{\partial \vec{v}}\left(D\left(\frac{1}{\eta_0 - f}\frac{\partial f}{\partial \vec{v}} + \beta f \vec{v}\right)\right). \tag{5.13}$$

Soit p la pression donnée par

$$p(\vec{r}, t) = \frac{1}{3}\int f(\vec{r}, \vec{v}, t) v^2 d^3\vec{v}. \tag{5.14}$$

À la limite haute friction, c'est – à – dire en faisant tendre ξ vers $+\infty$, dans l'équation de Kramers (5.13), P. H. Chavanis [**31**] montre

que l'on trouve l'équation de Smoluchowski fermionique

$$\frac{\partial \rho}{\partial t} = \nabla \left\{ \frac{1}{\xi} \left(\nabla p + \rho \nabla \Phi \right) \right\}, \tag{5.15}$$

que l'on étudie couplée avec l'équation de Poisson (5.3).

On remarque que l'équation d'état classique d'un gaz parfait $p = T\rho$ couplée à cette équation redonne l'équation (5.10) du cas classique.

Cette équation peut aussi être obtenue par le principe de production maximale d'entropie énoncé par P. H. Chavanis, R. Robert et J. Sommeria [**37**] en minimisant dans le cas canonique la variation temporelle de l'énergie libre F.

Pour obtenir les équations d'état, on fait l'hypothèse d'équilibre thermodynamique local, c'est – à – dire on suppose que la fonction de distribution est donnée localement par

$$f(\vec{r}, \vec{v}, t) = \frac{\eta_0}{1 + \lambda(\vec{r}, t) e^{\beta v^2/2}}. \tag{5.16}$$

On introduit les intégrales de Fermi

$$I_n(t) = \int_0^{+\infty} \frac{x^n}{1 + t e^x} dx. \tag{5.17}$$

Les équations (5.1) et (5.14) donnent alors l'équation d'état $p = p(\rho)$ sous forme implicite en fonction de ces intégrales de Fermi :

$$\rho = 4\sqrt{2}\pi \eta_0 T^{3/2} I_{1/2}(\lambda), \tag{5.18a}$$

$$p = \frac{8}{3}\sqrt{2}\pi \eta_0 T^{5/2} I_{3/2}(\lambda). \tag{5.18b}$$

Dans la suite et pour l'étude dynamique, nous nous intéresserons donc encore au système de Smoluchowski – Poisson mais fermionique cette fois et nous nous placerons dans le cas canonique. Donnons tout d'abord des résultats sur les états d'équilibre statistique.

Le comportement du système auto – gravitant de fermions est bien différent du comportement d'un système de particules classiques. Ceci est notamment visible à la figure 8 qui provient de l'article de P. H. Chavanis [**29**] où sont représentées les courbes caloriques pour différentes valeurs du paramètre de dégénérescence μ, qui est le paramètre η_0 adimensionné, et la courbe pour $\mu = +\infty$, c'est – à – dire la courbe pour le cas classique déjà représentée à la figure 6.

Nous nous intéressons plus particulièrement ici au cas $\mu = 10^3$ et nous présentons à la figure 9 la courbe calorique pour un gaz auto – gravitant de fermions.

Nous remarquons que dans ce cas, il existe toujours une solution stable contrairement au cas classique. En dessous de la température T_c, seule une phase condensée existe ; entre les deux températures critiques T_c et T_*, une phase condensée et une phase uniforme, dite gazeuse, coexistent ; enfin, au – dessus de la température T_*, il ne reste que la

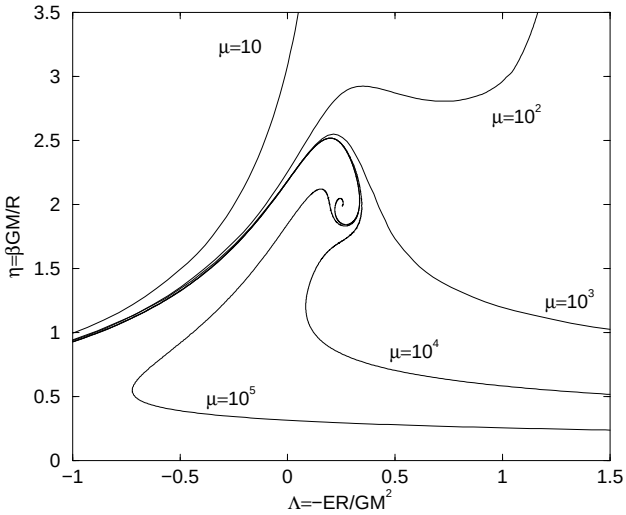


FIG. 8. Courbes caloriques pour un gaz auto – gravitant de fermions pour différentes valeurs de μ.

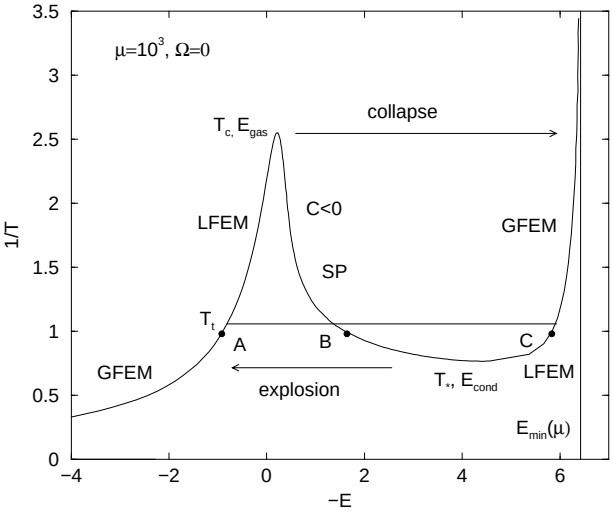


FIG. 9. Courbe calorique pour un gaz auto – gravitant de fermions avec $\mu = 10^3$. L'explosion signalée ici est une explosion au sens physique du terme, c'est – à – dire le contraire du collapse.

phase gazeuse. Le diagramme prédit un phénomène d'hystérésis qui sera étudié numériquement à la sous – section suivante 5.4. Il est également à noter que ce que l'on appelle collapse ici n'est pas le phénomène observé dans le cas classique et décrit à la figure 7. En effet, ici, la

densité n'explose pas avec le temps et se stabilise en une fonction très condensée à l'origine avec une structure cœur – halo.

Du point de vue mathématique, l'étude du système auto – gravitant de fermions de Smoluchowski – Poisson s'avère difficile, notamment à cause de la non unicité de la valeur asymptotique de la densité mise en évidence numériquement. P. Biler, P. Laurençot et T. Nadzieja [**12**] ont prouvé cependant l'existence d'une solution globale en temps et ont étudié les solutions stationnaires des équations (5.15) et (5.3) dans le cas de conditions de Dirichlet aux bords et de conditions de frontières libres. Les états d'équilibre ont été étudiés, quant à eux, par R. Stańczy [**115**].

À partir de maintenant, nous considérons des solutions à symétrie sphérique dans une sphère de rayon R.

5.3. Méthodes numériques employées. Nous considérons le système de Smoluchowski – Poisson pour $x \in [0,1]$ avec les variables adimensionnées n pour la densité, P pour la pression, ψ pour le potentiel gravitationnel et θ pour la température, c'est – à – dire

$$\left\{\begin{array}{l} \dfrac{\partial n}{\partial t} = \dfrac{1}{x^2}\dfrac{\partial}{\partial x}\left\{x^2\left(\dfrac{\partial P}{\partial x} + n\dfrac{\partial \psi}{\partial x}\right)\right\}, \\ \Delta\psi = 4\pi n, \end{array}\right. \tag{5.19}$$

avec les équations d'état

$$\left\{\begin{array}{l} n = \dfrac{\mu}{4\pi}\theta^{3/2} I_{1/2}(\lambda), \\ P = \dfrac{\mu}{6\pi}\theta^{5/2} I_{3/2}(\lambda), \end{array}\right. \tag{5.20}$$

et les conditions aux limites

$$\left\{\begin{array}{l} \dfrac{\partial \psi}{\partial x}(0) = 0,\ \dfrac{\partial \psi}{\partial x}(1) = 1, \\ \psi(1) = -1, \\ \dfrac{\partial P}{\partial x}(1) + \dfrac{\partial \psi}{\partial x}(1) n(1) = \dfrac{\partial P}{\partial x}(1) + n(1) = 0. \end{array}\right. \tag{5.21}$$

Cette dernière condition assure que la masse

$$\int_0^1 n(x) 4\pi x^2 dx$$

est conservée au cours du temps. Nous choisissons la densité initiale n_0 de telle sorte que la masse adimensionnée soit égale à 1.

Nous utilisons alors une méthode spectrale de collocation de Gauss – Lobatto – Legendre pour discrétiser les équations précédentes (5.19) en espace et un schéma semi – implicite pour la discrétisation en temps. Le système non – linéaire des équations d'état est résolu par une méthode de Newton. Les détails de l'algorithme sont donnés à la section préliminaire 1 du chapitre 4.

Une fois la densité et le potentiel gravitationnel connus à l'instant t, nous pouvons calculer l'énergie et l'énergie libre du système. L'énergie adimensionnée vaut

$$-\Lambda = -6\pi \int_0^1 P(x)x^2 dx + 2\pi \int_0^1 n(x)\psi(x)x^2 dx \tag{5.22}$$

et l'énergie libre adimensionnée

$$\begin{aligned} F = &-4\pi\theta \int_0^1 n(x)\ln(\lambda(x))x^2 dx - \frac{2}{3}\mu\theta^{5/2}\int_0^1 I_{3/2}(\lambda(x))x^2 dx \\ &+ 2\pi \int_0^1 n(x)\psi(x)x^2 dx. \end{aligned} \tag{5.23}$$

Les résultats obtenus grâce à cette méthode sont satisfaisants jusqu'à $\mu = 100$ au moins mais pour $\mu = 10^3$, le calcul des états condensés pose quelques problèmes ; les états gazeux sont, quant à eux, calculés correctement. Notamment lors du collapse, le changement de concavité de la densité au cours du temps provoque des oscillations numériques dues au choix de la méthode spectrale ; ces oscillations entraînent alors une valeur de la densité en 0 trop faible et l'énergie calculée est inférieure à la valeur prédite par la théorie. Par ailleurs, le nombre limité des points de Gauss – Lobatto – Legendre ne nous permet pas d'avoir une précision suffisante. Nous présentons à la figure 14 les résultats obtenus avec la méthode spectrale pour $\mu = 100$; ils sont tout à fait en accord avec la théorie et en particulier avec la courbe correspondant à $\mu = 100$ sur la figure 8. Notamment, la valeur théorique de la limite $\Lambda_{\max}$ est

$$\Lambda_{\max} = \lim_{\theta\to 0} \Lambda(\theta) \simeq 0.06418\mu^{2/3}, \tag{5.24}$$

ce qui donne pour $\mu = 100$, $\Lambda_{\max} \simeq 1.38$; ceci est bien vérifié numériquement à la figure 14. Il est également à noter que les états gazeux pour $\mu = 10^3$ sont calculés par la même méthode.

Pour avoir une méthode plus satisfaisante, on utilise un schéma aux différences finies en espace avec un maillage adapté et surtout plus fin. À cet effet, on récrit le système (5.19)- (5.21) sous la forme d'une unique équation aux dérivées partielles ; en introduisant la masse locale $M(x,t)$ contenue à l'instant t dans la boule de rayon x et définie par

$$M(x,t) = \int_0^x n(r)4\pi r^2 dr, \tag{5.25}$$

le système d'équations (5.19)- (5.21) est équivalent à l'équation suivante

$$\frac{\partial M}{\partial t} = 4\pi x^2 \frac{\partial P}{\partial x} + \frac{1}{x^2} M \frac{\partial M}{\partial x}, \tag{5.26}$$

avec pour équations d'état reliant implicitement P et M

$$\begin{cases} P = \dfrac{\mu}{6\pi}\theta^{5/2} I_{3/2}(\lambda), \\ \dfrac{1}{x^2}\dfrac{\partial M}{\partial x} = \mu\theta^{3/2} I_{1/2}(\lambda), \end{cases} \tag{5.27}$$

et pour conditions aux limites

$$M(0,t) = M(1,t) = 0. \tag{5.28}$$

Cette méthode donne effectivement dans le cas de l'état condensé une densité plus satisfaisante mais nécessite un pas de temps très petit, les calculs sont de ce fait beaucoup plus longs.

5.4. Résultats numériques obtenus. Grâce aux méthodes numériques expliquées au paragraphe précédent, nous pouvons suivre un cycle d'hystérésis pour $\mu = 10^3$: nous nous plaçons à la température $\theta = 1$ et nous prenons comme condition initiale une densité uniforme sur l'intervalle $[0, 1]$. Nous obtenons un état gazeux à l'équilibre, c'est – à – dire une densité non homogène mais avec un faible contraste.

Nous diminuons la température petit à petit et nous obtenons des états d'équilibre similaires au précédent, comme le montrent les différents profils de la figure 13. Les énergies d'équilibre ainsi obtenues forment la branche la plus à gauche de la figure 10.

À la température $\theta_c = 0.39$, nous observons un phénomène de collapse isotherme illustré à la figure 11. Contrairement au cas classique représenté à la figure 7, la densité n'explose pas : grâce aux effets quantiques, elle augmente avec le temps et se stabilise en un état très concentré près de l'origine ; c'est une structure cœur – halo. Pour des températures inférieures à θ_c, l'état gazeux a disparu et seul l'état condensé subsiste.

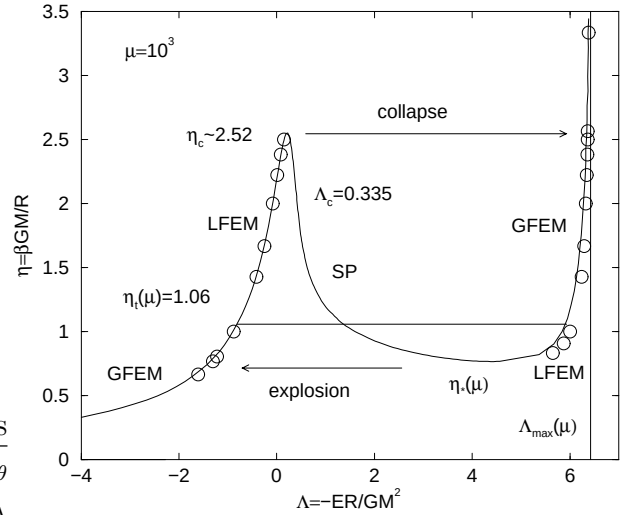
Puis nous augmentons la température à partir de θ_c et nous obtenons une suite d'états condensés, dont les profils sont représentés à la figure 13. On obtient ainsi la branche droite de la figure 10.

À la température $\theta_* = 1.24$, nous observons une explosion isotherme comme le montre la figure 12. Cette explosion, qui est le contraire du collapse, nous fait passer d'un état condensé à un état gazeux. Pour des températures supérieures à θ_*, seul existe l'état gazeux.

On remarque que les points de la figure 10 coïncident avec la courbe théorique de la figure 9.

Nous présentons à la figure 14 des résultats similaires pour le cas $\mu = 100$.

5.5. Organisation du chapitre 4. Dans la première section 1 du chapitre 4, nous expliquons comment programmer la méthode spectrale de collocation de Gauss – Lobatto – Legendre appliquée au système fermionique de Smoluchowski – Poisson.

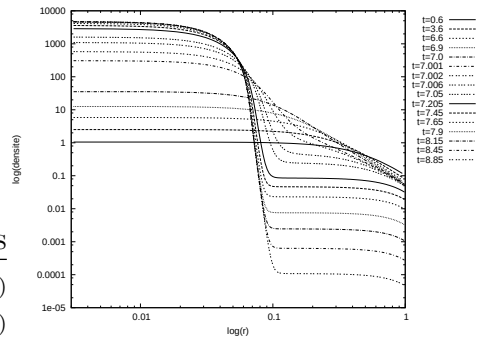


PSfrag replacements

$1/\theta$

Λ

FIG. 10. Courbe calorique pour $\mu = 10^3$; les cercles correspondent aux résultats numériques. On observe que pour une température donnée, il peut y avoir deux énergies d'équilibre, l'une correspondant à un état gazeux, l'autre à un état condensé.



PSfrag replacements

$\log(r)$

$\log(\rho)$

FIG. 11. Collapse isotherme pour la température $\theta = 0.39$ avec $\mu = 10^3$. Ces courbes représentent la densité en fonction du rayon pour différents temps et la densité centrale croît avec le temps.

Ensuite, dans la section 2, nous présentons dans la sous – section 2.2 les différentes équations modélisant les systèmes auto – gravitants de particules classiques.

Dans la sous – section 2.3 de l'article, nous étudions les équivalents fermioniques des précédents modèles et nous montrons que les équations de Kramers et de Smoluchowski peuvent s'obtenir à partir du principe de production maximale d'entropie.

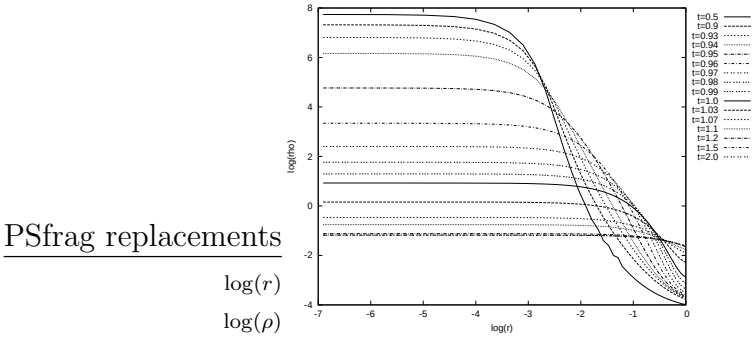


FIG. 12. Explosion isotherme pour la température $\theta = 1.24$ avec $\mu = 10^3$. Ces courbes représentent la densité en fonction du rayon et la densité centrale décroît avec le temps.

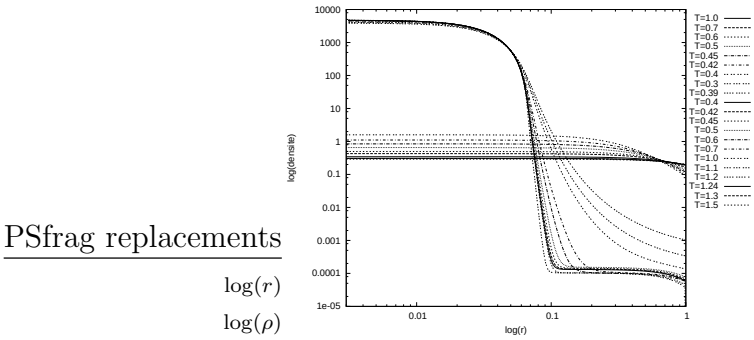


FIG. 13. Profils à l'équilibre pour différentes températures. On remarque nettement la différence entre les états gazeux et les états condensés avec une structure cœur – halo.

Finalement, dans la dernière sous – section 2.4 de l'article, nous nous intéressons aux comportements asymptotiques et à l'évolution dynamique de ces systèmes et nous mettons en évidence numériquement les phénomènes de collapse, d'explosion et d'hystérésis présentés à la sous – section 5.4 précédente.

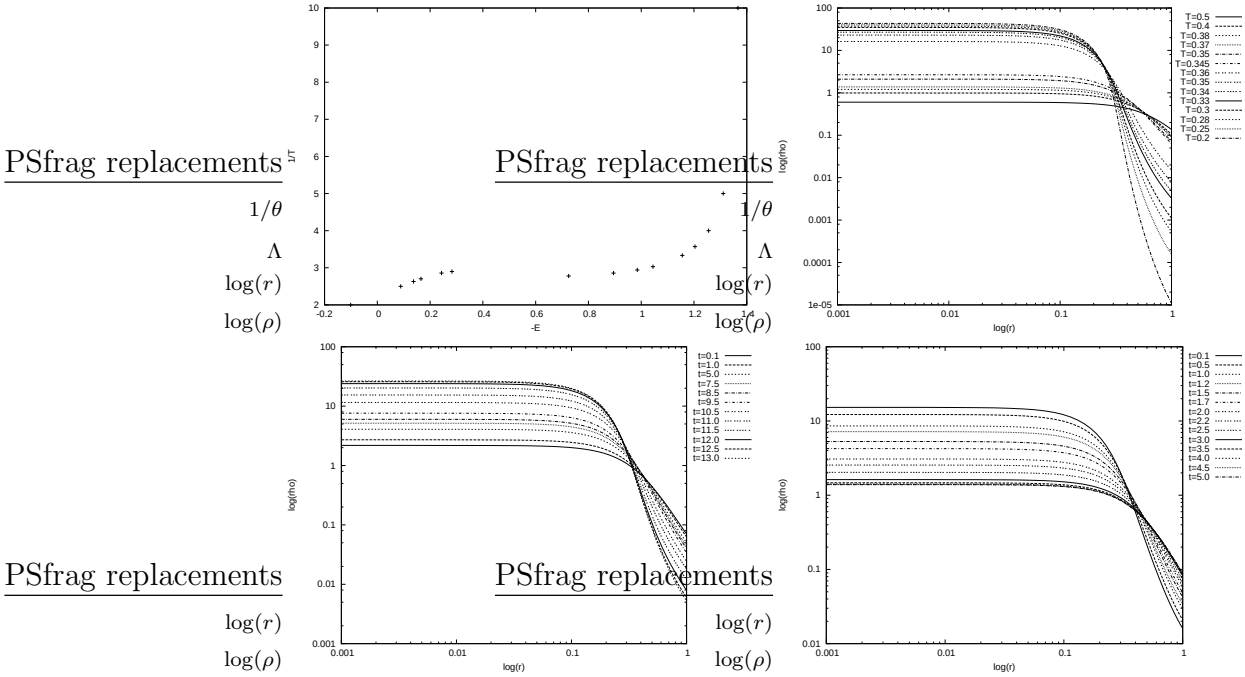


FIG. 14. Cas où $\mu = 100$: courbe calorique, profils pour différentes températures, évolution du collapse isotherme (la densité centrale croît avec le temps) et évolution de l'explosion isotherme (la densité centrale décroît avec le temps). Les simulations ont été effectuées grâce à une méthode spectrale.

CHAPITRE 1

Étude du schéma par régularisation du résidu

1. Rappels sur l'approximation des semi – groupes continus par des semi – groupes discrets

Dans la démonstration du théorème 4.12 de l'article [**99**] qui prouve la convergence du schéma par régularisation du résidu, nous utilisons un théorème de T. Kato [**72**] qui décrit l'approximation des semi – groupes continus par des semi – groupes discrets. Nous rappelons ici ce qu'est un semi – groupe discret et quel est ce théorème.

1.1. Semi – groupes discrets. Pour reprendre les notations de T. Kato, nous considérons ici l'approximation d'un semi – groupe $U(t)$ par une suite de semi – groupes discrets $\{U_n\}$.

Un semi – groupe discret est une famille $\{U^k\}_{k\in\mathbb{N}}$ de puissances d'un opérateur U. Nous lui associons un "pas de temps" τ et nous écrivons ce semi – groupe sous la forme $U^k = U(k;\tau)$, $k \in \mathbb{N}$.

Définition 1.1. *Soit*

$$T = \tau^{-1}\left(\mathbf{1} - U(1;\tau)\right);$$

on appelle générateur de $\{U(k;\tau)\}$ *l'opérateur* $-T$ *et on peut ainsi écrire*

$$U(k;\tau) = (\mathbf{1} - \tau T)^k.$$

Définition 1.2. *On dit qu'un semi – groupe discret* $\{U(k;\tau)\}$ *est borné s'il existe une constante* $M \geq 1$ *telle que* $\|U(k;\tau)\| \leq M$ *pour tout* k *dans* $\mathbb{N}$.

1.2. Approximation d'un semi – groupe continu par des semi – groupes discrets. Dans cette section, nous considérons une suite $\{U_n\}$ de semi – groupes discrets de pas de temps τ_n et de générateurs $-T_n$, où $\tau_n \to 0$ quand $n \to \infty$.

Définition 1.3. *Une suite* $\{U_n\}$ *de semi – groupes discrets approche un semi – groupe continu* $U = \{U(t)\}$ *en* $t = t_0$ *si*

$$U_n(k_n;\tau_n) \xrightarrow{s} U(t_0) \text{ quand } n \to \infty,$$

pour toute suite $\{k_n\}$ *d'entiers telle que* $k_n\tau_n \to t_0$.

On dit alors que $\{U_n\}$ approche U, ce que l'on note $U_n \to U$ si $\{U_n\}$ approche U pour tout t dans $[0, +\infty[$.

On a alors le lemme suivant :

LEMME 1.4. *Si $U_n \to U$, la suite U_n est uniformément quasi – bornée, c'est-à-dire*

$$\|U_n(t)\| \leq M e^{\beta t}, t \geq 0,$$

où M et β sont indépendants de n et de t et où $U_n(t)$ est défini pour tout $t \geq 0$ par

$$U_n(t) = U_n\left(\lfloor t/\tau \rfloor \tau\right).$$

Il est donc nécessaire pour que $\{U_n\}$ approche U que $\{U_n\}$ soit uniformément quasi – bornée ; on énonce donc le théorème suivant :

THÉORÈME 1.5. *Soit $\{U_n\}$ une suite uniformément quasi – bornée et U un semi – groupe de générateur $-T$. $\{U_n\}$ approche U ssi $T_n \xrightarrow{s} T$; il est donc nécessaire que*

$$(T_n + \zeta)^{-1} \xrightarrow{s} (T + \zeta)^{-1} \tag{1.1}$$

pour tout ζ tel que $\mathrm{Re}\,(\zeta) > \beta$ et il est suffisant que (1.1) soit vérifié pour un ζ tel que $\mathrm{Re}\,(\zeta) > \beta$.

Au Théorème 4.12, où on applique ce théorème, la suite de semi – groupes est uniformément quasi – bornée grâce à la propriété de stabilité du schéma par régularisation du résidu prouvée précédemment.

2. Rappels sur les extrapolations de Richardson

Dans cette partie, nous étudions le schéma par régularisation du résidu défini à l'équation (2.2) ; nous prouvons qu'il est stable, convergent et d'ordre un. À partir de ce schéma, nous pouvons définir ses extrapolations de Richardson.

L'intérêt de ces extrapolations est que, d'une part, si le schéma initial est d'ordre p, la k-ième extrapolation de Richardson est d'ordre $p+k-1$. D'autre part, les extrapolations de Richardson se définissent de façon très simple à partir du schéma initial ; pratiquement, programmer l'extrapolation de Richardson d'un schéma revient à itérer un certain nombre de fois le schéma avec différents pas de temps (plus précisément, on itère n_j fois le schéma avec le pas de temps t/n_j, où les n_j sont des entiers définis à l'avance) et à effectuer une combinaison linéaire des résultats ainsi obtenus.

On a donc ainsi un schéma aussi facile à programmer que le schéma initial et d'ordre plus élevé.

Nous donnons dans la section suivante 2.1 une idée de la preuve de ce résultat dans le cadre de séries formelles et nous énonçons dans la section 2.2 l'application du Théorème 2.1 au cas de l'approximation d'un semi – groupe. Ces résultats se trouvent dans l'article de B.O. Dia et de M. Schatzman [**55**]. Les extrapolations de Richardson avaient déjà été étudiées pour les équations différentielles dans l'article de Bulirsch et Stoer [**22**] ou dans le livre de Romberg [**104**], ainsi que dans le livre de Hairer, Nørsett et Wanner [**62**] ; cependant, Dia et Schatzman en donnent une version algébrique, qui permet d'obtenir des résultats sur l'extrapolation de Richardson en dimension infinie.

2.1. Une version formelle des extrapolations de Richardson. Nous allons d'abord donner une version de la convergence des extrapolations de Richardson formulée grâce à des séries formelles.

Soit U une algèbre unitaire sur un corps qui est $\mathbb{R}$ ou $\mathbb{C}$ et soit $U[[X]]$ l'anneau des séries formelles à une indéterminée.

THÉORÈME 2.1. *Soit f dans $U[[X]]$ telle que*

$$f(X) = \mathbf{1} + aX + \sum_{j\geq 2} f_j X^j. \tag{2.1}$$

On suppose que la différence entre f et l'exponentielle de Xa est de valuation $p+1$ avec $p \geq 1$, soit

$$d(X) = f(X) - \exp(Xa) = \sum_{j\geq p+1} d_j X^j. \tag{2.2}$$

Alors pour tout entier k et pour tout choix d'entiers $1 \leq n_1 < \cdots < n_k$, il existe k coefficients rationnels α_j, $1 \leq j \leq k$, tels que la série formelle

$$\sum_{j=1}^{k} \alpha_j f(X/n_j)^{n_j} - \exp(Xa)$$

soit de valuation $p + k$.

Expliquons la preuve de ce théorème.

Démonstration. Pour étudier le terme $f(X/n_j)^{n_j}$, nous allons passer au logarithme et donc la première étape est le calcul du logarithme de f.

Notons

$$g = \ln f = \sum_{j \geq 1} (-1)^{j-1} \frac{(f - \mathbf{1})^j}{j}.$$

D'après l'hypothèse (2.1), le terme devant X dans g vaut a. Étudions les termes suivants. Pour cela, on écrit f sous la forme

$$f(X) = (\mathbf{1} + d(X) \exp(-aX)) \exp(aX),$$

où d est défini à l'équation (2.2) et on note

$$\gamma(X) = \ln (\mathbf{1} + d(X) \exp(-aX)) \, ;$$

d'après l'hypothèse (2.2), d et donc γ sont alors de valuation $p + 1$. Pour calculer g, on utilise alors le Lemme de Campbell – Hausdorff, qui donne le développement du logarithme d'un produit d'exponentielles. On obtient alors l'égalité suivante

$$\begin{aligned} g(X) = \ln f(X) &= \ln (\exp(\gamma(X)) \exp(aX)) \\ &= \gamma(X) + aX + \frac{1}{2}[\gamma(X), aX] \\ &+ \frac{1}{12}[[\gamma(X), aX], aX] + \frac{1}{12}[[\gamma(X), aX], \gamma(X)] \\ &+ \cdots + L_k(\gamma(X), aX) + \cdots , \end{aligned}$$

où L_k est un polynôme de Lie de degré k, c'est – à – dire une somme de termes composés de $k-1$ commutateurs emboîtés faisant intervenir γ et aX.

La valuation de $L_2(\gamma(X), aX) = [\gamma(X), aX]$ est au moins égale à la somme des valuations de γ et aX, soit $p+2$. Plus généralement, chaque monôme de L_k contient k termes dont au moins une fois $\gamma(X)$, qui est de valuation $p + 1$; L_k est donc au moins de valuation $p + k$.

Finalement, g peut donc s'écrire sous la forme suivante :

$$g(X) = aX + \sum_{i \geq p+1} g_i X^i = X \bar{g}(X),$$

où

$$\bar{g}(X) = a + \sum_{i \geq p} g_{i+1} X^i. \tag{2.3}$$

La deuxième étape consiste à calculer $f(X/n)^n$, qui s'écrit :

$$f(X/n)^n = \exp(ng(X/n)) = \exp\left(X\bar{g}(X/n)\right) = h(X, 1/n),$$

avec

$$h(X,Y) = \exp\left(X\bar{g}(XY)\right) = \mathbf{1} + \sum_{j \geq 1} \frac{X^j \bar{g}(XY)^j}{j!}.$$

Donc $h - \mathbf{1}$ est une somme de termes du type

$$\frac{1}{j!} X^{j+k_1+\cdots+k_j} Y^{k_1+\cdots+k_j} g_{k_1+1} \cdots g_{k_j+1}, \tag{2.4}$$

pour $j \geq 1$ et d'après l'équation (2.3), les coefficients k_i sont soit nuls, soit supérieurs à p ; en particulier, $k_1 + \cdots + k_j < p$ implique que les k_i sont tous nuls. En écrivant

$$h(X,Y) = \sum_{i \geq 0} Y^i h_i(X),$$

on en déduit que pour tout $i = 1, \cdots, p-1$, la série h_i est nulle. On s'aperçoit enfin, grâce à la forme des termes (2.4) et au fait que $g_1 = a$, que $h_0 = \exp(aX)$ et que h_i, pour $i \geq p$, est de valuation au moins égale à $i+1$.

Ainsi, on peut écrire h sous la forme

$$h(X,Y) = \exp(aX) + \sum_{i \geq p} Y^i h_i(X),$$

où h_i est de valuation au moins $i+1$; $f(X/n)^n$ s'écrit alors

$$f(X/n)^n = \exp(aX) + \sum_{i \geq p} \frac{h_i(X)}{n^i}.$$

Nous pouvons alors conclure la preuve ainsi : la différence entre $\sum \alpha_j f(X/n_j)^{n_j}$ et l'exponentielle de aX s'écrit

$$\begin{aligned} &\sum_{j=1}^{k} \alpha_j f(X/n_j)^{n_j} - \exp(aX) \\ &\quad = \left(\sum_{j=1}^{k} \alpha_j - 1\right) \exp(aX) + \sum_{i \geq p} h_i(X) \sum_{1 \leq j \leq k} \frac{\alpha_j}{n_j^i}. \end{aligned} \tag{2.5}$$

Pour que cette différence soit de valuation la plus élevée possible, on choisit les α_j tels que

$$\left\{\begin{array}{l} \displaystyle\sum_{j=1}^{k} \alpha_j = 1, \\ \displaystyle\sum_{1\leq j\leq k} \frac{\alpha_j}{n_j^i} = 0,\ \forall i = p, \cdots, p+k-2; \end{array}\right. \tag{2.6}$$

le système (2.6) est un système de Vandermonde, qui admet donc une unique solution. Alors, le terme de valuation la plus basse dans la somme de l'équation (2.5) est

$$h_{p+k-1}(X) \sum_{1\leq j\leq k} \frac{\alpha_j}{n_j^{p+k-1}},$$

de valuation $p+k$, ce qui conclut la preuve. □

Remarque 2.2. *Dans le cas où $p = 1$, on connaît la valeur exacte de la solution du système* (2.6) *: soit ℓ_j^k, $1 \leq j \leq k$, les k polynômes interpolateurs de Lagrange de degré $k-1$ associés aux nœuds $1/n_j$, $1 \leq j \leq k$, on a alors $\alpha_j = \ell_j^k(0)$, $1 \leq j \leq k$.*

En effet, pour tout entier m, le polynôme $P_m(x) = \sum_{1\leq j\leq k} \ell_j^k(x)/n_j^m$ est un polynôme de degré $k-1$ tel que pour tout $1 \leq l \leq k$, $P_m(1/n_l) = 1/n_l^m$; on en déduit donc que pour $0 \leq m \leq k-1$, $P_m(x) = x^m$ et en particulier $P_m(0) = \delta_{0,m}$ où δ est le symbole de Kronecker. Ceci prouve bien que les $\ell_j^k(0)$ sont solutions du système (2.6).

2.2. Application à l'approximation des semi – groupes fortement continus. Nous appliquons maintenant le théorème de la section précédente à l'approximation des semi – groupes fortement continus. Nous énonçons ici les hypothèses et le résultat principal de la section 3 de [**55**].

Soit X un Banach et A un opérateur sur X. Soit Y un sous – espace dense de X tel que pour tout entier m de $\mathbb{Z}$, $Y \subset D(A^m)$. Nous notons alors par $\mathcal{Z}_k$ l'espace suivant :

$$\mathcal{Z}_k = \{L : Y \to Y \text{ tels que } \forall m \in \mathbb{Z} \sup_{u\in Y\setminus\{0\}} \frac{|A^{m-k}Lu|}{|A^m u|} \leq C_m < \infty\}$$

et nous définissons les normes suivantes :

$$|L|_{m,k} = \sup_{u\in Y\setminus\{0\}} \frac{|A^{m-k}Lu|}{|A^m u|}.$$

Nous supposons alors que A engendre un semi – groupe fortement continu.

Nous supposons également que l'approximation $f : [0, \tau] \to \mathcal{B}(X)$, où $\mathcal{B}(X)$ désigne les opérateurs bornés de X, peut se développer ainsi :

$$f(t) = \mathbf{1} + \sum_{j=1}^{r} f_j t^j + t^r \varepsilon_r(t), \tag{2.7}$$

où $f_1 = -A$, $f_1, \cdots, f_r \in \mathcal{Z} = \cup_{k \in \mathbb{Z}} \mathcal{Z}_k$ et où $\varepsilon_r : [0, \tau] \to \mathcal{Z}_l$ vérifie pour tout entier m de $\mathbb{Z}$,

$$\lim_{t \to 0} |\varepsilon_r(t)|_{m,l} = 0. \tag{2.8}$$

Nous faisons maintenant une dernière hypothèse sur l'approximation par f du semi – groupe engendré par A : il existe deux entiers $q \leq r$ et h tels que

$$\lim_{t \to 0} t^{-q} \left| f(t) - \exp(-tA) \right|_{m,h} = 0. \tag{2.9}$$

Nous pouvons alors énoncer le théorème suivant qui prouve que la k-ième extrapolation de Richardson d'un schéma, pour $k \geq 2$, est d'ordre supérieur au schéma initial :

Théorème 2.3. *Soit $k \geq 1$ et k entiers naturels $n_1, \cdots, n_k$. Sous les hypothèses* (2.7), (2.8) *et* (2.9), *on a pour tout entier $p \leq \min(r + 1 - q, k)$, l'existence de k rationnels $\alpha_1, \cdots, \alpha_k$ et d'un entier i tels que pour tout $m \in \mathbb{Z}$,*

$$\lim_{t \to 0} t^{-(p+q-1)} \left| \sum_{j=1}^{k} \alpha_j f(t/n_j)^{n_j} - \exp(-tA) \right|_{m,i} = 0.$$

Ce théorème avec $q = 2$ et $r = k + 1$ nous permet de prouver le Théorème 4.15 de la section 4.8, c'est – à – dire la convergence des extrapolations de Richardson du schéma par régularisation du résidu.

3. Stabilité du schéma par régularisation du résidu et de ses extrapolations

Stability of the extrapolations of the Residual Smoothing Scheme

Magali Ribot and Michelle Schatzman

Abstract : *We show that the extrapolation of the RSS scheme which amounts to including a preconditioning in the integration of a stiff differential system or the discretization of a partial differential equation is stable, and provides therefore a scheme whose computational cost is independent of the order of precision.*

3.1. Introduction. The time integration of a parabolic problem can be done by two means: either an explicit scheme, subject to the CFL (Courant–Friedrichs–Lewy) condition, a drastic condition which bounds from above the time step by the square of the space step or by an implicit scheme which requires an efficient solver, and in particular the use of preconditioners.

The usual method to precondition a parabolic problem is to be reduced to precondition an elliptic problem: we semidiscretize in time and then we look for a preconditioner for the matrix $I+\Delta t A$ instead of A for an elliptic problem, like in F.Bornemann [**17, 18**] , L.Mulholland and D.Sloan [**85**] or P.Brown and C.Woodward [**21**]. Here we study a scheme, the Residual Smoothing Scheme, which includes the preconditioner in the time discretization.

Let A be a self-adjoint operator in a Hilbert space; we assume that A is bounded from below and we consider the problem

$$\left\{\begin{array}{r} \dfrac{du}{dt}+Au=0, \\ u(0)=u_0. \end{array}\right. \tag{3.1}$$

Without loss of generality, we may assume that for all x in the domain $D(A)$ of A, we have

$$x^*Ax \geq |x|^2. \tag{3.2}$$

Indeed, if A is bounded from below, there exists C in $\mathbb{R}$ such that

$$x^*Ax \geq C|x|^2,$$

we set $v=ue^{-\lambda t}$ in (3.1) and we obtain

$$\left\{\begin{array}{r} \dfrac{dv}{dt}+(A+\lambda)v=0, \\ v(0)=u_0, \end{array}\right.$$

that is to say a system analogous to (3.1) with $\widetilde{A} = A + \lambda$ instead of A. We choose λ such that $C + \lambda \geq 1$ and therefore inequality (3.2) holds for $\widetilde{A}$.

Assume that B is a self – adjoint unbounded operator which has the same domain as A and which satisfies

$$x^* A x \leq c x^* B x \tag{3.3}$$

for some strictly positive constant c for all x in $D(A)$.

The residual smoothing scheme has been considered in Averbuch et al. [**4**] as an alternative to the backward Euler scheme; it is given by

$$\frac{U_{n+1} - U_n}{\Delta t} + \tau B(U_{n+1} - U_n) + AU_n = 0, \tag{3.4}$$

where τ is a parameter which can be chosen to enforce stability.

In [**4**], the authors prove that the scheme (3.4) is unconditionally stable for τ large enough if (3.3) holds.

Let us define the operator

$$P(t) = (\mathbf{1} + t\tau B)^{-1}(\mathbf{1} + t(\tau B - A)).$$

Another form of P is given by

$$P(t) = \mathbf{1} - R(t), \quad R(t) = t(\mathbf{1} + t\tau B)^{-1} A. \tag{3.5}$$

Therefore, equation (3.4) can be rewritten as:

$$U_{n+1} = P(t) U_n.$$

Let us define the Richardson's extrapolations of the scheme.

Given any choice of integers $1 \leq n_1 < n_2 < \cdots < n_k$, the Richardson's extrapolation of $P(t)$ is

$$P_k(t) = \sum_{j=1}^{k} \ell_j^k(0) P(t/n_j)^{n_j}$$

where ℓ_j^k are the elements of the Lagrange interpolation basis with knots $1/n_j$.

We notice here that once the scheme is computed, it is very easy to compute its extrapolations; moreover, if P is of order 1, its Richardson extrapolation P_k is of order k. Therefore, computing the Richardson extrapolations is an easy way to obtain high order schemes.

In this article, we prove that the Richardson's extrapolations P_k are stable. We notice that $P(t)$ and $R(t)$ are not self-adjoint, so we symmetrize them by introducing

$$\widetilde{R}(t) = t\sqrt{A}\,(\mathbf{1} + t\tau B)^{-1}\sqrt{A}$$

and

$$\widetilde{P}(t) = \mathbf{1} - \widetilde{R}(t).$$

We first prove using assumption (3.3) that the Richarson's extrapolations $\widetilde{P}_k$ of $\widetilde{P}$ are stable. The rest of the article is devoted to compare P_k and $\widetilde{P}_k$ in order to prove the stability of P_k.

For this purpose, we need the following assumptions: we define an algebra $\mathcal{M}$ and we assume that the commutator of $\sqrt{B}$ with any element of $\mathcal{M}$ belongs to $\mathcal{M}$. We also assume that

$$\sqrt{A} = \sqrt{B}m_1 + m_2,$$

where m_1 and m_2 belong to $\mathcal{M}$ and that

$$\sqrt{B}\left[m_1, [\sqrt{B}, m_1]\right] \in \mathcal{M}. \tag{3.6}$$

We also need some estimates on $R(t)$, $\widetilde{R}(t)$, the difference of their powers and the commutator of their powers. For this purpose, we show the so-called Sandwich Lemma which makes the proofs of these estimates easier. From these estimates, we can control the difference between symmetrized and unsymmetrized forms and we can complete the proof of the stability of the Richardson's extrapolations $P_k(t)$.

However, we tried to check numerically assumption (3.6) using a spectral method for A and a finite elements method for B and we failed to check it on the boundaries, although this assumption holds perfectly inside the domain. Therefore, in [**99**], we prove the stability, the convergence and the order of the Residual Smoothing Scheme and its extrapolations in energy norm $|x|_A = (x^*Ax)^{1/2}$. For this purpose, the only assumption we make is the equivalence between A and B and we do not need assumption (3.6).

Then, we show that the equivalence between A and B holds when A is a spectral method and B a finite elements method in [**97**]. We prove this result not only for the one-dimensional case but also for higher dimensions. We also present in this article some numerical simulations performed with the Residual Smoothing Scheme and its extrapolations. Eventually, to prove the equivalence between operators A and B, we need some precise asymptotics on Legendre polynomials and their extrema. This is done in [**98**] using a stationary phase method.

The article is organized as follows: in section 3.2, we first define an order relation between operators and we study its properties; then, in section 3.3, we prove some algebraic results dealing with Lagrange interpolation polynomials and then we show the stability of the symmetrized extrapolations of RSS. Section 3.4 is devoted to the comparison between the unsymmetrized and the symmetrized forms of RSS and is split in three subsections: in subsection 3.4.1, we prove the so-called Sandwich Lemma; in subsection 3.4.2, we use this Sandwich Lemma to show some estimates on $R(t)$, $\widetilde{R}(t)$, on some differences of their powers and on some commutators of their powers; eventually, in

subsection 3.4.3, we prove the stability of the unsymmetrized form of RSS using the previous results.

3.2. An order on self-adjoint operators. For the reader's convenience, we recall here the definition of the order on two operators coming from assumption (3.3) and the properties of this relation, which are already given in [**99**].

In this article, we denote by $\mathbf{1}$ the identity operator in any vector space. We recall that every self-adjoint operator T in a Hilbert space possesses a spectral decomposition

$$T = \int_{\mathbb{R}} \lambda dP(\lambda),$$

where $dP(\lambda)$ is the spectral measure associated to T. We will say that a self-adjoint operator T is positive semi-definite if for all $x \in D(T)$, $x^*Tx \geq 0$. If T is positive semi-definite, the square root of T is defined by

$$\sqrt{T} = \int_{\mathbb{R}} \sqrt{\lambda} dP(\lambda).$$

We define as follows a partial order relation between self-adjoint and bounded from below operators in a Hilbert space H:

(3.7)
$$T_1 \prec T_2 \Longrightarrow D(T_2) \subset D(T_1) \text{ and } \forall x \in D(T_2), \quad x^*T_1x \leq x^*T_2x.$$

We define the relations $\succ$ to be the opposite relation to $\prec$.

We may relate this equivalence relation to algebraic operations; in particular, if S is a self-adjoint operator which is bounded from below, it is plain that

$$T_1 \prec T_2 \Longrightarrow T_1 + S \prec T_2 + S.$$

If S is any bounded operator from a Hilbert space H_1 to H, and if the domain of S^*T_jS for $j = 1, 2$ is defined as $S^{-1}D(T_j)$, we have also:

$$T_1 \prec T_2 \Longrightarrow S^*T_1S \prec S^*T_2S. \tag{3.8}$$

The proof is performed through the change of variable $x = Sy$.

Another important fact is the following:

Lemma 3.1. *If T_1 and T_2 are positive self-adjoint and injective, then*

$$T_1 \prec T_2 \Longrightarrow T_2^{-1} \prec T_1^{-1}.$$

Proof. This can be deduced from the proof of Theorem VI.2.21 in Kato's book [**72**]. □

Observe that if $T_1 \prec T_2$ then for any powers $\alpha \in]0, 1[$, $T_1^\alpha \prec T_2^\alpha$. Indeed a formula of Balakrishnan in [**5**] which is given in Yosida's book [**125**] gives the representation of T^α:

$$x \in D(T) \Rightarrow T^\alpha x = \frac{\sin(\alpha\pi)}{\pi} \int_0^\infty \lambda^{\alpha-1} (\lambda\mathbf{1} + T)^{-1} Tx d\lambda.$$

The relation

$$(\lambda\mathbf{1}+T)^{-1}\,T=\mathbf{1}-\lambda\,(\lambda\mathbf{1}+T)^{-1}$$

is classical; we infer from Lemma 3.1 that

$$(\lambda\mathbf{1}+T_1)^{-1}\,T_1\prec(\lambda\mathbf{1}+T_2)^{-1}\,T_2;$$

therefore, it is plain that $T_1^\alpha\prec T_2^\alpha$.

However $T_1\prec T_2$ does not imply $T_1^n\prec T_2^n$ for all n in $\mathbb{N}$; a counter-example is, for instance

$$T_1=\begin{pmatrix}2\varepsilon & 0\\ 0 & 2/\varepsilon\end{pmatrix},\quad T_2=\begin{pmatrix}\varepsilon & 1/2\\ 1/2 & 1/\varepsilon\end{pmatrix}.$$

The reader will check that for all positive ε, $T_1\succ T_2$, while for all small enough ε, it is not true that $T_1^2\succ T_2^2$. However, if the self-adjoint, positive operators T_1 and T_2 commute, and in particular if one of them is scalar, the conclusion is true and this can be checked simply with the help of the spectral theorem.

3.3. The algebra of extrapolation and the stability of the symmetrized extrapolation. We define the coefficients of Richardson's extrapolation as follows: let ℓ_j^k be the Lagrange basis relative to the nodes $1/n_j$, $1\le j\le k$, with $1\le n_1<\cdots<n_j<\cdots<n_k$:

$$\ell_j^k(t)=\prod_{\{i:i\neq j\}}\frac{t-1/n_i}{(1/n_j)-(1/n_i)}. \tag{3.9}$$

Some well-known choices for these nodes are, for example, the harmonic ones, i.e. $n_j=j$, the Romberg ones with $n_j=2^j$ or the Bulirsch ones, defined as follows:

$$1,2,3,4,6,8,12,16,\cdots,2^j,\frac{3}{2}2^j,2^{j+1},\frac{3}{2}2^{j+1},\cdots.$$

By definition of the Lagrange basis, the following equalities hold:

$$\sum_{j=1}^k\ell_j^k(t)=1, \tag{3.10}$$

$$\forall p=1,\cdots,k-1,\quad \sum_{j=1}^k\ell_j^k(t)\frac{1}{n_j^p}=t^p;$$

therefore the following identities hold

$$\forall p=0,\cdots,k-1,\quad \sum_{j=1}^k\frac{\ell_j^k(0)}{n_j^p}=\delta_{0p}. \tag{3.11}$$

The Richardson's extrapolation is given by

$$P_k(t)=\sum_{j=1}^k\ell_j^k(0)P(t/n_j)^{n_j},$$

which can also be written

$$
\begin{aligned}
P_k(t) = \mathbf{1}\sum_{j=1}^{k} \ell_j^k(0) - \sum_{j=1}^{k} \ell_j^k(0) n_j R(t/n_j) \\
+ \sum_{j=1}^{k}\sum_{i=2}^{n_j} (-1)^i \ell_j^k(0) C_{n_j}^i R(t/n_j)^i.
\end{aligned} \tag{3.12}
$$

Formally, we have the expansion

$$P(t) = \mathbf{1} - tA + O(t^2) = \exp(-tA) + O(t^2),$$

which implies the expansion

$$P_k(t) = \exp(-tA) + O(t^{k+1}),$$

as recalled in section 2.

We define a symmetrized form of R, P and P_k:

$$
\begin{aligned}
\widetilde{R}(t) &= t\sqrt{A}(\mathbf{1} + t\tau B)^{-1}\sqrt{A}, \\
\widetilde{P}(t) &= \mathbf{1} - \widetilde{R}(t), \text{ and}
\end{aligned} \tag{3.13}
$$

$$\widetilde{P}_k(t) = \sum_{j=1}^{k} \ell_j^k(0)\big(\mathbf{1} - \widetilde{R}(t/n_j)\big)^{n_j}. \tag{3.14}$$

A more explicit form of $\widetilde{P}_k$ is given by

$$
\begin{aligned}
\widetilde{P}_k(t) = \mathbf{1}\sum_{j=1}^{k} \ell_j^k(0) - \sum_{j=1}^{k} \ell_j^k(0) n_j \widetilde{R}(t/n_j) \\
+ \sum_{j=1}^{k}\sum_{i=2}^{n_j} (-1)^i \ell_j^k(0) C_{n_j}^i \widetilde{R}(t/n_j)^i.
\end{aligned} \tag{3.15}
$$

From equation (3.10), the first term $\mathbf{1}\sum_{j=1}^{k} \ell_j^k(0)$ is equal to $\mathbf{1}$; the second term $\sum_{j=1}^{k} \ell_j^k(0) n_j \widetilde{R}(t/n_j)$ can be expressed as $t\sqrt{A}\phi_k(t\tau B, 0)\sqrt{A}$, where ϕ_k is the following function

$$\phi_k(s,t) = \sum_{j=1}^{k} \frac{\ell_j^k(t)}{1 + s/n_j}.$$

We search for an equivalent for ϕ_k at infinity and for this purpose, we first notice that the function $\phi_k(s,\cdot)$ interpolates the function $f : t \mapsto 1/(1+st)$ at the points $1/n_j$, $1 \le j \le k$.

Lemma 3.2. *The function $s \mapsto \phi_k(s,0)$ is strictly positive over $\mathbb{R}^+$.*

Proof. For any function g, denote by $g[x_1, \cdots, x_n]$ the divided difference of the function g at the knots $x_1, \cdots, x_n$.

We use Newton's form of interpolation:

$$\phi_k(s,t) = f(1/n_1) + f[1/n_1, 1/n_2](t - 1/n_1) + \cdots \\ + f[1/n_1, 1/n_2, \cdots, 1/n_k](t - 1/n_1)(t - 1/n_2)\cdots(t - 1/n_{k-1});$$

therefore, for $t = 0$, we obtain

$$\begin{aligned} \phi_k(s,0) = f(1/n_1) - \frac{f[1/n_1, 1/n_2]}{n_1} + \frac{f[1/n_1, 1/n_2, 1/n_3]}{n_1 n_2} + \cdots \\ + \frac{(-1)^{k-1} f[1/n_1, 1/n_2, \cdots, 1/n_k]}{n_1 n_2 \cdots n_k}. \end{aligned} \tag{3.16}$$

For $j \geq 1$, let S^j be the simplex

$$S^j = \{x \in (\mathbb{R}^+)^j : x_1 + \cdots + x_j \leq 1\};$$

the divided differences are given by the integral representation

$$f[a_1, \cdots, a_{j+1}] = \int_{S^j} f^{(j)}(a_1 + t_1(a_2 - a_1) + \cdots + t_j(a_{j+1} - a_j)). \tag{3.17}$$

But in our particular case,

$$f^{(j)}(t) = \frac{j!(-s)^j}{(1+st)^{j+1}} \tag{3.18}$$

and therefore the divided difference containing j terms is positive for j odd and negative for j even.

Consequently, substituting (3.17) and (3.18) into (3.16), we see immediately that the assertion of the lemma is true. □

We need another algebraic fact:

Lemma 3.3. *For all $k = 1, 2, \cdots$ the following identity holds:*

$$\sum_{j=1}^{k} n_j \ell_j^k(0) = \sum_{j=1}^{k} n_j. \tag{3.19}$$

Proof. Write

$$T_k = \sum_{j=1}^{k} n_j \ell_j^k(0).$$

We infer from formula (3.9) that $\ell_j^k(0)$ is given as

$$\ell_j^k(0) = (-n_j)^{k-1} \prod_{i:i\neq j} \frac{1}{n_i - n_j}.$$

Therefore, we have the relation

$$T_k - T_{k-1} = n_k \ell_k^k(0) + \sum_{j=1}^{k-1} (-1)^{k-1} n_j \Big(\frac{n_j^{k-1}}{\prod_{i\neq j}^{k}(n_i - n_j)} \\ + \frac{n_j^{k-2}}{\prod_{i\neq j}^{k-1}(n_i - n_j)} \Big)$$

and by elementary manipulations,

$$T_k - T_{k-1} = n_k \ell_k^k(0) + n_k \sum_{j=1}^{k-1} \frac{(-n_j)^{k-1}}{\prod_{i\neq j}(n_i - n_j)} = n_k \sum_{j=1}^{k} \ell_j^k(0).$$

We conclude with the help of (3.11). □

In the following lemma, we show that the norm of $\widetilde{R}$ is bounded:

LEMMA 3.4. *There exists $c > 0$ such that for all $t > 0$,*

$$\left\|\widetilde{R}(t)\right\| \leq c/\tau.$$

PROOF. Using assumption (3.3), we find that

$$t\tau A \prec c(\mathbf{1} + t\tau B).$$

Thus, using Lemma 3.1, we obtain

$$(\mathbf{1} + t\tau B)^{-1} \prec \frac{cA^{-1}}{t\tau}.$$

and finally using assertion (3.8) with $S = \sqrt{A}$, we find that

$$t\sqrt{A}(\mathbf{1} + t\tau B)^{-1}\sqrt{A} \prec \frac{c}{\tau}\mathbf{1}.$$

□

We are now able to prove the stability of the symmetrized extrapolation:

THEOREM 3.5. *For all $k \in \mathbb{N}$, for any choice of integers $1 \leq n_1 < n_2 < \cdots < n_k$, there exists $\tau_0 > 0$ such that for $\tau \geq \tau_0$, $\widetilde{P}_k$ is stable, and more precisely, for all $t > 0$ the following inequality holds in the sense of quadratic forms:*

$$-\mathbf{1} \prec \widetilde{P}_k(t) \prec \mathbf{1}. \tag{3.20}$$

PROOF. The first two terms in (3.15) are

$$\mathbf{1}\sum_{j=1}^{k} \ell_j^k(0) = \mathbf{1}$$

and

$$-\sum_{j=1}^{k} \ell_j^k(0) n_j \widetilde{R}(t/n_j) = -t\sqrt{A}\sum_{j=1}^{k} \ell_j^k(0)(\mathbf{1} + t\tau B/n_j)^{-1}\sqrt{A}. \tag{3.21}$$

Lemma 3.2 says that the function $\psi_k(s) = \phi_k(s, 0)$ is strictly positive on $\mathbb{R}^+$; lemma 3.3 enables us to find an equivalent of ψ_k at infinity:

$$\psi_k(s) \sim \sum_{j=1}^{k} \frac{n_j \ell_j^k(0)}{s} = \frac{\sum_{j=1}^{k} n_j}{s},$$

and therefore, there exist numbers μ_k and μ'_k such that

$$\frac{\mu_k}{1+s} \leq \psi_k(s) \leq \frac{\mu'_k}{1+s}. \tag{3.22}$$

It is plain that the expression (3.21) can be rewritten

$$-t\sqrt{A}\psi_k(t\tau B)\sqrt{A}; \tag{3.23}$$

using (3.22) this expression is bounded from above by

$$-t\mu_k\sqrt{A}(\mathbf{1}+t\tau B)^{-1}\sqrt{A} = -\mu_k\widetilde{R}(t). \tag{3.24}$$

The other terms of $\widetilde{P}_k(t)$ are estimated as follows: for all $j = 1, \cdots, k$, we have the inequality

$$(\mathbf{1}+t\tau B) \prec n_j(\mathbf{1}+t\tau B/n_j),$$

from which we infer the following estimate with the help of Lemma 3.1 and assertion (3.8),

$$\widetilde{R}(t/n_j) \prec \widetilde{R}(t). \tag{3.25}$$

Since, for any positive integer i, the following inequality holds in virtue of Lemma 3.4:

$$\widetilde{R}(t/n_j)^i \prec \left(\frac{c}{\tau}\right)^i \mathbf{1},$$

we can deduce that

$$\widetilde{R}(t/n_j)^i \prec \left(\frac{c}{\tau}\right)^{i-1} \widetilde{R}(t/n_j)$$

and using equation (3.25), we obtain that

$$\widetilde{R}(t/n_j)^i \prec \left(\frac{c}{\tau}\right)^{i-1} \widetilde{R}(t). \tag{3.26}$$

Therefore, using the upper bounds (3.24) and (3.26), we can estimate $\widetilde{P}_k$ as follows

$$\widetilde{P}_k(t) \prec \mathbf{1} - \mu_k\widetilde{R}(t) + \sum_{j=1}^{k}\sum_{i=2}^{n_j}\left|\ell_j^k(0)\right|C_{n_j}^i\left(\frac{c}{\tau}\right)^{i-1}\widetilde{R}(t).$$

The inequality $\widetilde{P}_k(t) \prec \mathbf{1}$ will therefore be satisfied if, in the sense of quadratic forms,

$$\sum_{j=1}^{k}\sum_{i=2}^{n_j}\left|\ell_j^k(0)\right|C_{n_j}^i\left(\frac{c}{\tau}\right)^{i-1}\widetilde{R}(t) \prec \mu_k\widetilde{R}(t); \tag{3.27}$$

but (3.27) is satisfied if

$$\sum_{j=1}^{k}\sum_{i=2}^{n_j}\left|\ell_j^k(0)\right|C_{n_j}^i\left(\frac{c}{\tau}\right)^{i-1}\mathbf{1} \prec \mu_k\mathbf{1}$$

also in the sense of quadratic forms; we infer immediately the inequality $\widetilde{P}_k \prec \mathbf{1}$ for τ large enough; the other inequality is treated similarly: we will have $\widetilde{P}_k \succ -\mathbf{1}$ if

$$\mu'_k \frac{c}{\tau} + \sum_{j=1}^{k} \sum_{i=2}^{n_j} |\ell_j^k(0)| C_{n_j}^i \left(\frac{c}{\tau}\right)^i \leq 2$$

and the previous argument also shows the stability for large enough values of τ. □

3.4. Comparison of the symmetrized and of the unsymmetrized forms of RSS. Let us denote by b the square root of B and by a the square root of A.

In order to be able to compare P_k and $\widetilde{P}_k$, we make assumptions on the commutators of $b = \sqrt{B}$ with the elements of an appropriate algebra $\mathcal{M}$; these elements will behave essentially as pseudodifferential operators of degree 0. However, we will formulate our conditions independently of the pseudodifferential theory. Therefore we assume that $\mathcal{M}$ is an unitary algebra of bounded operators from H to itself such that:

$$\begin{cases} \text{for all integer } k, \text{ the elements of } \mathcal{M} \\ \quad \text{map the domain of } B^{k/2} \text{ to itself;} \end{cases} \tag{3.28}$$

and

(3.29) the commutator of b with any element of $\mathcal{M}$ belongs to $\mathcal{M}$.

We will write for simplicity,

$$[b, m_i] = m'_i.$$

We assume also that there exist m_1 and m_2 in $\mathcal{M}$ such that

$$a = \sqrt{A} = bm_1 + m_2; \tag{3.30}$$

and that the expression

$$b\,[m_1, [b, m_1]] = b\,[m_1, m'_1] \in \mathcal{M}. \tag{3.31}$$

Let us introduce a finite number of time steps t_i for $1 \leq i \leq N$. In the following, we will have in general $t_i = t/n_i$.

We introduce the following notation to make calculations more readable:

$$\beta(t) = (1 + t\tau B)^{-1/2} \text{ and } \beta_i = \beta(t_i).$$

We remark that here $\beta(t)$ is the square root of β defined in [**99**].

3.4.1. *The Sandwich Lemma.* Let us describe the sandwich lemma.

We denote by p a product of elements b, β_i and elements of $\mathcal{M}$, and we let $\deg(p, b)$ be the number of b factors in the product, $\deg(p, \beta)$ the number of β_i factors. The p will be omitted when there is no doubt on the expression considered. We call p a sandwich: it contains different sorts of bread which are β_i, meat which is b and salad made out of elements of $\mathcal{M}$. The sandwich is good if there are more slices of bread than of meat, and the quantity of salad is arbitrary. The sandwich lemma says that we can estimate p in operator norm if it is a good sandwich, and we prove it by moving around the ingredients and showing that we get equivalent sandwiches. The term "sandwich lemma" has already been used by Schatzman [**107**] where the sandwich contained two sorts of meat and two sorts of bread.

Now, we will introduce some algebraic tools in view of stating and proving that lemma.

By convention bold symbols denote letters and words, and ordinary symbols denote operators; concatenation and formal linear combination of words are permissible, but we do not perform operator multiplications or additions on them. Of course, we may perform additions and multiplications on operators. The classical definition of polynomials in non commutative variables makes this idea very precise.

Let indeed $\boldsymbol{\mathcal{Y}}$ be an alphabet, and let $\boldsymbol{\mathcal{Y}}^*$ be the monoid of finite words written with the letters of $\boldsymbol{\mathcal{Y}}$, and including as neutral element the empty word; the operation in $\boldsymbol{\mathcal{Y}}^*$ is the concatenation of words. The algebra of non commutative polynomials $\mathbb{Q}\langle\boldsymbol{\mathcal{Y}}\rangle$ is the space of mappings from $\boldsymbol{\mathcal{Y}}^*$ to $\mathbb{Q}$, which vanish except on a finite set of elements of $\boldsymbol{\mathcal{Y}}^*$. In particular, for every element of $\mathbb{Q}\langle\boldsymbol{\mathcal{Y}}\rangle$, only a finite number of letters of $\boldsymbol{\mathcal{Y}}$ is used. The degree of $\mathbf{p} \in \mathbb{Q}\langle\boldsymbol{\mathcal{Y}}\rangle$ with respect to an indeterminate $\mathbf{y} \in \boldsymbol{\mathcal{Y}}$ is the largest number of occurrences of $\mathbf{y}$ in any monomial of $\mathbf{p}$.

Denote by $\boldsymbol{\mathcal{M}}$ the list of elements of $\mathcal{M}$, seen as an alphabet; if m_1 and m_2 belong to $\mathcal{M}$ and $m_3 = m_1 m_2$, then $\mathbf{m_1}$ and $\mathbf{m_2}$ belong to $\boldsymbol{\mathcal{M}}$ but the word $\mathbf{m_1 m_2}$ is not equal to the element $\mathbf{m_3}$ of $\boldsymbol{\mathcal{M}}$. We denote by $\mathbf{w_i}$ the words, elements of $\boldsymbol{\mathcal{M}}^*$.

We consider the following alphabet, for $N \in \mathbb{N}$,

$$\boldsymbol{\mathcal{X}}_N = \{\mathbf{b}\} \cup \{\boldsymbol{\beta_i},\, 1 \leq i \leq N\} \cup \boldsymbol{\mathcal{M}},$$

and

$$\boldsymbol{\mathcal{X}} = \bigcup_{N \in \mathbb{N}} \boldsymbol{\mathcal{X}}_N.$$

We will use the algebras of non commutative polynomials

$$\mathbb{Q}\langle\boldsymbol{\mathcal{X}}_N\rangle \text{ and } \mathbb{Q}\langle\boldsymbol{\mathcal{X}}\rangle$$

and we observe that

$$\mathbb{Q}\langle\boldsymbol{\mathcal{X}}\rangle = \bigcup_{N \in \mathbb{N}} \mathbb{Q}\langle\boldsymbol{\mathcal{X}}_N\rangle.$$

Let us define now some operators algebras: $\mathcal{A}_N$ is the $\mathbb{Q}$-algebra generated by the operators b, $\{\beta_1 \cdots \beta_N\}$ and $\mathcal{M}$ and

$$\mathcal{A} = \bigcup_{N \in \mathbb{N}} \mathcal{A}_N;$$

for $N = 0$, $\mathcal{A}_0$ is generated by b and $\mathcal{M}$; it contains $a = bm_1 + m_2$.

To relate $\mathbb{Q}\langle \boldsymbol{\mathcal{X}} \rangle$ and $\mathcal{A}$, we define the substitution

$$\Phi : \mathbb{Q}\langle \boldsymbol{\mathcal{X}} \rangle \to \mathcal{A}$$

as the unique algebra morphism such that

$$\begin{cases} \Phi(\mathbf{b}) = b, \\ \Phi(\boldsymbol{\beta_i}) = \beta_i, \\ \Phi(\mathbf{m}) = m \text{ for all } \mathbf{m} \in \boldsymbol{\mathcal{M}}. \end{cases}$$

This substitution can be restricted to a mapping from $\mathbb{Q}\langle \boldsymbol{\mathcal{X}}_N \rangle$ to $\mathcal{A}_N$ and it enables us to deduce analytic results from algebraic ones.

We now define the degree of an element $\mathbf{p} \in \mathbb{Q}\langle \boldsymbol{\mathcal{X}} \rangle$. This definition cannot be analytic: is the analytic expression $(1 + t\tau b^2)\beta(t)^2 = 1$ of degree 0 in both b and β or of degree 2 ? For $\mathbf{p} \in \mathbb{Q}\langle \boldsymbol{\mathcal{X}} \rangle$, the degrees of the polynomial $\mathbf{p}$ with respect to $\mathbf{b}$ and $\boldsymbol{\beta_i}$,

$$\deg(\mathbf{p}, \mathbf{b}) \text{ and } \deg(\mathbf{p}, \boldsymbol{\beta_i})$$

are well defined and we let the global degree with respect to $\boldsymbol{\beta}$ be

$$\deg(\mathbf{p}, \boldsymbol{\beta}) = \sum_i \deg(\mathbf{p}, \boldsymbol{\beta_i}).$$

We also define $\boldsymbol{\mathcal{Q}}_k$ the algebraic set generated by the monomials $\mathbf{p}$ such that

$$\deg(\mathbf{p}, \mathbf{b}) \leq \deg(\mathbf{p}, \boldsymbol{\beta}) \leq k$$

and $\boldsymbol{\mathcal{Q}}_{k,l}$, for $l \leq k$, the set of monomials $\mathbf{p}$ such that

$$l = \deg(\mathbf{p}, \mathbf{b}) \leq \deg(\mathbf{p}, \boldsymbol{\beta}) = k.$$

Let $\boldsymbol{\mathcal{Q}} = \bigcup_{k \in \mathbb{N}} \boldsymbol{\mathcal{Q}}_k$.

We can prove now the following sandwich lemma:

Lemma 3.6 (Sandwich Lemma). *Let $k \in \mathbb{N}$ and $\mathbf{p} \in \boldsymbol{\mathcal{Q}}_k$. Define $t = \min_{i \in [1,k]}(t_i)$; then there exists $C > 0$ such that if $t \in [0, 1]$*

$$\|\Phi(\mathbf{p})\| \leq C t^{-\deg(\mathbf{p},\mathbf{b})/2}. \tag{3.32}$$

To prove this lemma, we identify the principal part of a monomial $\mathbf{p}$ such that $\deg(\mathbf{p}, \mathbf{b}) = d$ and $\deg(\mathbf{p}, \boldsymbol{\beta}) = 0$. Thus, $\mathbf{p}$ can be written

$$\mathbf{p} = \mathbf{w_0 b w_1 b} \cdots \mathbf{b w_d} \text{ with } \mathbf{w_i} \in \boldsymbol{\mathcal{M}}^*. \tag{3.33}$$

Lemma 3.7. *Let $\mathbf{p}$ be given by* (3.33) *and let $\mathbf{w} = \mathbf{w_0 w_1} \cdots \mathbf{w_d}$ and $\mathbf{r} = \mathbf{p} - \mathbf{w b}^d$. Then, there exists $\mathbf{r}'$ such that $\deg(\mathbf{r}', \mathbf{b}) = d - 1$, $\deg(\mathbf{r}', \boldsymbol{\beta}) = 0$ and $\Phi(\mathbf{r}) = \Phi(\mathbf{r}')$.*

Proof. The proof is left to the reader; it is sufficient to show it by induction on d, commuting the $\mathbf{w_i}$ factors with the $\mathbf{b}$ factors and using the hypothesis (3.29). $\square$

Now, let us prove the sandwich lemma:

Proof. The spectral theorem gives immediately the following estimate for all $t > 0$

$$\|b\beta(t)\| \leq 1/\sqrt{\tau t}. \tag{3.34}$$

We prove our lemma by induction on k; if $k = 0$, then $\Phi(\mathbf{p})$ is a sum of elements of $\mathcal{M}$, they are bounded and the result is clear. We give two proofs: in the first one we perform algebraic manipulations; the second one is graphic.

First proof: suppose that the result holds true for all $\mathbf{p} \in \boldsymbol{\mathcal{Q}}_k$; we prove it for monomials $\mathbf{p}$ of $\boldsymbol{\mathcal{Q}}_{k+1}$ by an embedded induction on the index $l \leq k+1$ for which $\mathbf{p} \in \boldsymbol{\mathcal{Q}}_{k+1,l}$; if $l = 0$, then $\Phi(\mathbf{p})$ is a product of β_i and of elements of $\mathcal{M}$; they are bounded and $\Phi(\mathbf{p})$ is consequently bounded.

Suppose now that the result holds true for all monomials $\mathbf{p}$ of $\boldsymbol{\mathcal{Q}}_{k+1,l}$ and let $\mathbf{p}$ be a monomial of $\boldsymbol{\mathcal{Q}}_{k+1,l+1}$. It can be written

$$\mathbf{p} = \mathbf{q_0}\boldsymbol{\beta_1}\mathbf{q_1}\boldsymbol{\beta_2}\cdots\mathbf{q_k}\boldsymbol{\beta_{k+1}}\mathbf{q_{k+1}},$$

where for all $j \in [0, k+1]$, $\deg(\mathbf{q_j}, \boldsymbol{\beta}) = 0$ and

$$\sum_j \deg(\mathbf{q_j}, \mathbf{b}) = l + 1.$$

Let $j_1 = \min\{j \in [0, k+1] : \deg(\mathbf{q_j}, \mathbf{b}) > 0\}$. The aim of the calculation is to write $\Phi(\mathbf{p})$ as the concatenation of two terms plus a sum of terms estimated by the embedded induction assumption: among those two concatenated terms, one contains only one b and one β, which is estimated using (3.34), and the other one is estimated using the induction assumption. For this purpose, let us commute some terms in $\mathbf{p}$. We consider two cases: the case when $j_1 = 0$ and the case when $j_1 > 0$.

Assume that $j_1 = 0$: this means that $\mathbf{q_0}$ contains factors $\mathbf{b}$; set $d = \deg(\mathbf{q_0}, \mathbf{b}) \geq 1$. Thanks to Lemma 3.7, $\mathbf{q_0}$ can be written

$$\mathbf{q_0} = \mathbf{w}\mathbf{b}^d + \mathbf{r},$$

and there exists $\mathbf{r}'$ such that $\deg(\mathbf{r}', \mathbf{b}) = d - 1$, $\deg(\mathbf{r}', \boldsymbol{\beta}) = 0$ and $\Phi(\mathbf{r}) = \Phi(\mathbf{r}')$. Thus, letting

$$\mathbf{p_1} = \mathbf{r}'\boldsymbol{\beta_1}\mathbf{q_1}\boldsymbol{\beta_2}\cdots\mathbf{q_k}\boldsymbol{\beta_{k+1}}\mathbf{q_{k+1}} \text{ and}$$
$$\mathbf{p_2} = \mathbf{w}\mathbf{b}^d\boldsymbol{\beta_1}\mathbf{q_1}\boldsymbol{\beta_2}\cdots\mathbf{q_k}\boldsymbol{\beta_{k+1}}\mathbf{q_{k+1}},$$

$$\Phi(\mathbf{p}) = \Phi(\mathbf{p_1}) + \Phi(\mathbf{p_2});$$

by induction, since $\mathbf{p_1} \in \boldsymbol{\mathcal{Q}}_{k+1,l}$, we have the estimate

$$\|\Phi(\mathbf{p_1})\| \leq Ct^{-l/2}.$$

To show that

$$\|\Phi(\mathbf{p})\| \leq Ct^{-(l+1)/2},$$

it is thus sufficient to prove that $\mathbf{p_2}$ verifies

$$\|\Phi(\mathbf{p_2})\| \leq Ct^{-(l+1)/2}.$$

In $\mathbf{p_2}$, we move $d-1$ factors $\mathbf{b}$ from the left of $\boldsymbol{\beta_1}$ to its right:

$$\begin{aligned}\mathbf{p_2} &= \mathbf{w}\mathbf{b}^d\boldsymbol{\beta_1}\mathbf{q_1}\boldsymbol{\beta_2}\cdots\mathbf{q_k}\boldsymbol{\beta_{k+1}}\mathbf{q_{k+1}}\\ &= \mathbf{w}\mathbf{b}\boldsymbol{\beta_1}\mathbf{b}^{d-1}\mathbf{q_1}\boldsymbol{\beta_2}\cdots\mathbf{q_k}\boldsymbol{\beta_{k+1}}\mathbf{q_{k+1}}\\ &\quad + \mathbf{w}\mathbf{b}[\mathbf{b}^{d-1},\boldsymbol{\beta_1}]\mathbf{q_1}\boldsymbol{\beta_2}\cdots\mathbf{q_k}\boldsymbol{\beta_{k+1}}\mathbf{q_{k+1}}.\end{aligned}$$

And substituting, we find that

(3.35)

$$\Phi(\mathbf{p_2}) = wb\beta_1 b^{d-1}q_1\beta_2\cdots q_k\beta_{k+1}q_{k+1} + \underbrace{wb[b^{d-1},\beta_1]q_1\beta_2\cdots q_k\beta_{k+1}q_{k+1}}_{=0},$$

as b and β commute. Define

$$\mathbf{p_3} = \mathbf{b}^{d-1}\mathbf{q_1}\boldsymbol{\beta_2}\cdots\mathbf{q_k}\boldsymbol{\beta_{k+1}}\mathbf{q_{k+1}} \in \boldsymbol{\mathcal{Q}}_{k,l};$$

then

$$\Phi(\mathbf{p_2}) = w\Phi(\mathbf{b}\boldsymbol{\beta_1})\Phi(\mathbf{p_3}). \tag{3.36}$$

Thus, using (3.34) and the induction assumption, we get the estimates

$$\|\Phi(\mathbf{b}\boldsymbol{\beta_1})\| \leq Ct^{-1/2} \text{ and } \|\Phi(\mathbf{p_3})\| \leq Ct^{-l/2}.$$

The estimate for $\Phi(\mathbf{p_2})$ and thus for $\Phi(\mathbf{p})$ is now clear.

Assume now that $j_1 > 0$. This means that $\mathbf{p_0}$ contains no $\mathbf{b}$ factor and for $i < j_1$, $\mathbf{q_i} = \mathbf{w_i} \in \boldsymbol{\mathcal{M}}^*$. Let $\mathbf{q_{j_1}} = \mathbf{w_{j_1,1}}\mathbf{b}\widehat{\mathbf{q_{j_1}}}$. We move one factor $\mathbf{b}$ of $\mathbf{q_{j_1}}$ to the right of $\boldsymbol{\beta_1}$. Let

$$\begin{aligned}\mathbf{p_1} &= \mathbf{w_0}\boldsymbol{\beta_1}\mathbf{b}\mathbf{w_1}\boldsymbol{\beta_2}\cdots\mathbf{w_{j_1,1}}\widehat{\mathbf{q_{j_1}}}\cdots\boldsymbol{\beta_{k+1}}\mathbf{q_{k+1}} \text{ and}\\ \mathbf{p_2} &= \mathbf{w_0}\boldsymbol{\beta_1}[\mathbf{w_1}\boldsymbol{\beta_2}\cdots\mathbf{w_{j_1,1}},\mathbf{b}]\widehat{\mathbf{q_{j_1}}}\cdots\boldsymbol{\beta_{k+1}}\mathbf{q_{k+1}};\end{aligned}$$

therefore $\mathbf{p} = \mathbf{p_1} + \mathbf{p_2}$.

We estimate $\Phi(\mathbf{p_1})$ by cutting it as we had cut $\Phi(\mathbf{p_2})$ in (3.36), and we find

$$\|\Phi(\mathbf{p_1})\| \leq Ct^{-(l+1)/2}.$$

The second term $\mathbf{p_2}$ can be developed:

$$\begin{aligned}\mathbf{p_2} &= \sum_{i=1}^{j_1-1}\mathbf{w_0}\cdots\boldsymbol{\beta_i}[\mathbf{w_i},\mathbf{b}]\boldsymbol{\beta_{i+1}}\cdots\mathbf{w_{j_1,1}}\widehat{\mathbf{q_{j_1}}}\cdots\mathbf{q_{k+1}}\\ &\quad + \sum_{i=2}^{j_1}\mathbf{w_0}\cdots\mathbf{w_{i-1}}[\boldsymbol{\beta_i},\mathbf{b}]\mathbf{w_i}\cdots\mathbf{w_{j_1,1}}\widehat{\mathbf{q_{j_1}}}\cdots\mathbf{q_{k+1}}\\ &\quad + \mathbf{w_0}\cdots\boldsymbol{\beta_{j_1}}[\mathbf{w_{j_1,1}},\mathbf{b}]\widehat{\mathbf{q_{j_1}}}\cdots\mathbf{q_{k+1}}.\end{aligned} \tag{3.37}$$

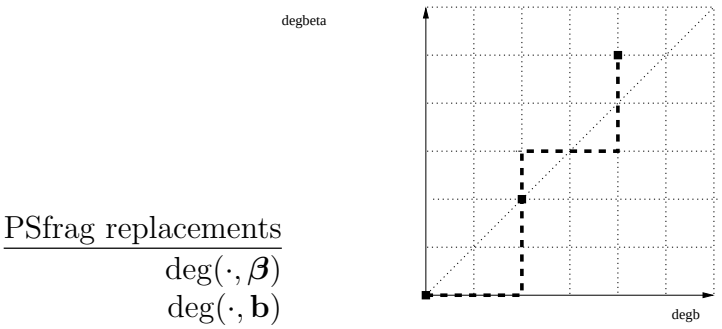


FIGURE 1. Graph of the expression (3.38)

Applying Φ, the second sum of the right hand side of (3.37) vanishes because b and β commute. Thus, there remains

$$\Phi(\mathbf{p_2}) = \sum_{i=1}^{j_1-1} w_0 \cdots \beta_i \underbrace{[w_i, b]}_{=w'_i \in \mathcal{M}} \beta_{i+1} \cdots w_{j_1,1} \widehat{q_{j_1}} \cdots q_{k+1}$$
$$+ w_0 \cdots \beta_{j_1} \underbrace{[w_{j_1,1}, b]}_{=w'_{j_1,1} \in \mathcal{M}} \widehat{q_{j_1}} \cdots q_{k+1}.$$

If we define

$$\mathbf{p_{3,i}} = \mathbf{w_0} \cdots \boldsymbol{\beta_i} \mathbf{w'_i} \boldsymbol{\beta_{i+1}} \cdots \mathbf{w_{j_1,1}} \widehat{\mathbf{q_{j_1}}} \cdots \mathbf{q_{k+1}} \in \boldsymbol{\mathcal{Q}}_{k+1,l} \text{ and}$$
$$\mathbf{p_4} = \mathbf{w_0} \cdots \boldsymbol{\beta_{j_1}} \mathbf{w'_{j_1,1}} \widehat{\mathbf{q_{j_1}}} \cdots \mathbf{q_{k+1}} \in \boldsymbol{\mathcal{Q}}_{k+1,l};$$

then,

$$\Phi(\mathbf{p_2}) = \sum_{i=1}^{j_1-1} \Phi(\mathbf{p_{3,i}}) + \Phi(\mathbf{p_4}).$$

It is now sufficient to apply the induction assumption to conclude that

$$\|\Phi(\mathbf{p_2})\| \leq C t^{-l/2}.$$

Second proof: we illustrate this proof with the help of figures 1 to 3. We read a monomial of $\boldsymbol{\mathcal{Q}}$ beginning at $(0,0)$; each time we find $\mathbf{b}$ we perform a step to the right, and each time we find $\boldsymbol{\beta}$, we perform an upward step. For example, the graph corresponding to the expression

$$\mathbf{p} = \mathbf{m_0 b m_1 b} \boldsymbol{\beta(t)}^2 \mathbf{m_2} \boldsymbol{\beta(t)} \mathbf{b}^2 \mathbf{m_3} \boldsymbol{\beta(t)}^2 \tag{3.38}$$

is shown in dashed line in figure 1.

Suppose now that the sandwich lemma has been proved for all $\mathbf{p} \in \boldsymbol{\mathcal{Q}}_k$; it is sufficient to prove the estimate for the monomials of $\boldsymbol{\mathcal{Q}}_{k+1}$ such that $\deg(\mathbf{p}, \mathbf{b}) = \deg(\mathbf{p}, \boldsymbol{\beta}) = k+1$; the monomials of $\boldsymbol{\mathcal{Q}}_{k+1}$ such that $\deg(\mathbf{p}, \boldsymbol{\beta}) < k+1$ have already been estimated through the induction assumption; the monomials satisfying $\deg(\mathbf{p}, \mathbf{b}) < \deg(\mathbf{p}, \boldsymbol{\beta}) = k+1$ will be estimated later.

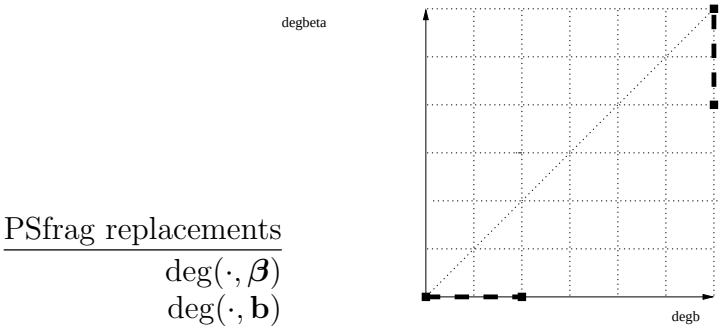


FIGURE 2. First case when a decomposition does not exist.

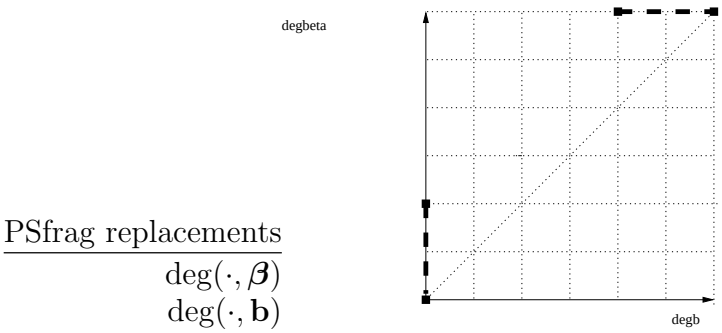


FIGURE 3. Second case when a decomposition does not exist.

Let $\mathbf{p}$ be a monomial such that $\deg(\mathbf{p}, \mathbf{b}) = \deg(\mathbf{p}, \boldsymbol{\beta}) = k+1$. If the graph of $\mathbf{p}$ crosses the diagonal (in dotted line in the figures) at other points than the end points $(0,0)$ and $(k+1, k+1)$, then there exists a decomposition $\mathbf{p} = \mathbf{p_1}\mathbf{p_2}$ where $\mathbf{p_1} \in \boldsymbol{\mathcal{Q}}_i$ with $i \leq k$ and $\mathbf{p_2} \in \boldsymbol{\mathcal{Q}}_j$ with $j \leq k$. It is the case for example of the expression (3.38) with

$$\mathbf{p_1} = \mathbf{m_0}\mathbf{b}\mathbf{m_1}\mathbf{b}\boldsymbol{\beta}(\boldsymbol{t})^2 \text{ and } \mathbf{p_2} = \mathbf{m_2}\boldsymbol{\beta}(\boldsymbol{t})\mathbf{b}^2\mathbf{m_3}\boldsymbol{\beta}(\boldsymbol{t})^2.$$

Therefore, it is possible to estimate $\mathbf{p}$ using twice the induction assumption, once for $\mathbf{p_1}$, once for $\mathbf{p_2}$.

If the graph does not cross the diagonal, the expression begins and ends as in figures 2 or 3. But these two cases are symmetric; it suffices to study the first one. We see that $\mathbf{p}$ can be written as

$$\mathbf{p} = \mathbf{q_0}\boldsymbol{\beta_1}\mathbf{q_1}\boldsymbol{\beta_2}\cdots\mathbf{q_k}\boldsymbol{\beta_{k+1}}\mathbf{q_{k+1}}$$

with $\deg(\mathbf{q_0}, \mathbf{b}) \geq 1$ and $\deg(\mathbf{q_{k+1}}, \mathbf{b}) = 0$ and we estimate it as in the first proof for the case $j_1 = 0$.

Finally, let $\mathbf{p}$ be a monomial such that $\deg(\mathbf{p}, \mathbf{b}) < \deg(\mathbf{p}, \boldsymbol{\beta}) = k+1$; to estimate it, it is sufficient to commute enough $\mathbf{b}$ terms as was done in the first proof in order to write $\mathbf{p} = \mathbf{p_1}\mathbf{p_2} + \mathbf{r}$ where $\mathbf{p_1}$ is such that $\deg(\mathbf{p_1}, \mathbf{b}) = \deg(\mathbf{p_1}, \boldsymbol{\beta})$, $\mathbf{p_2}$ contains only $\boldsymbol{\beta}$ terms and words of $\boldsymbol{\mathcal{M}}^*$, and $\mathbf{r}$ is a term containing all the commutators; thus, we estimate $\Phi(\mathbf{p_1})$ according to the above proof, $\Phi(\mathbf{p_2})$ is bounded and we estimate

$\Phi(\mathbf{r})$ thanks to the induction assumption. This concludes the second proof. $\square$

3.4.2. *Estimates on $R(t)$ and $\widetilde{R}(t)$.* In order to simplify the proofs of the estimates needed later, we introduce an equivalence. For that purpose, we introduce the degree of an operator p as

$$\deg(p, b) = \min\{\deg(\mathbf{p}, \mathbf{b}) : p = \Phi(\mathbf{p})\};$$

we can define by an analogous way the degree with respect to β.

Let us now define the equivalence : two expressions p_1 and p_2 of $\mathcal{A}$ are said to be equivalent if

$$\deg(p_1 - p_2, b) < \deg(p_1, b) = \deg(p_2, b)$$

and

$$\deg(p_1, \beta) = \deg(p_2, \beta).$$

We will write then $p_1 \approx p_2$. This definition will enable us to simplify the calculations, since two equivalent expressions satisfy the same estimate; for example, if p_1 and p_2 are of degree greater than 1 in b and if they differ by a commutator $[b, m]$, they are equivalent.

In particular, we need the following equivalents:

$$a = bm_1 + m_2 \approx bm_1, \tag{3.39}$$

and also

$$\begin{aligned} A = a^2 &= bm_1bm_1 + m_2bm_1 + bm_1m_2 + m_2^2 \\ &= b^2m_1^2 + bm_1'm_1 + m_2bm_1 + bm_1m_2 + m_2^2 \approx b^2m_1^2 \end{aligned} \tag{3.40}$$

and in the same fashion,

$$a^3 \approx b^3m_1^3.$$

Next lemma gives a list of equivalences of commutators needed later in this section

LEMMA 3.8. *There exist m_i, $3 \leq i \leq 6$ such that*

$$[a, B] \approx b^2m_3, \tag{3.41}$$

$$[A, B] \approx b^3m_4, \tag{3.42}$$

$$\big[[a, B], B\big] \approx b^3m_5, \tag{3.43}$$

and

$$\big[a, [a, B]\big] \approx b^2m_6. \tag{3.44}$$

PROOF. This is a straightforward calculation which is left to the reader, provided that assumption (3.31) is used to obtain (3.44). $\square$

Now let us prove some estimates on $R(t)$ and $\widetilde{R}(t)$ that we need to prove the stability of the unsymmetrized form P_k.

The first straightforward result, using equivalents and the sandwich lemma, is the following lemma:

LEMMA 3.9. *The expressions* $R(t) = t\beta(t)^2 A$ *and* $\widetilde{R}(t) = ta\beta(t)^2 a$ *are bounded for* $t \in [0,1]$.

We will need the following identity for any operator Q,

$$\left[Q, \beta(t)^2\right] = t\tau\beta(t)^2[B,Q]\beta(t)^2. \tag{3.45}$$

Now let us prove the following estimate needed later:

LEMMA 3.10. *For all* t *in* $[0,1]$, *for all* $n \in \mathbb{N}$,

$$\left\| R(t)^n - \widetilde{R}(t)^n \right\| \leq C_n \sqrt{t}. \tag{3.46}$$

PROOF. We prove this by induction: for $n = 1$, we calculate $R(t) - \widetilde{R}(t)$, introducing commutators:

$$\begin{aligned} R(t) - \widetilde{R}(t) &= t\left(\beta(t)^2 a^2 - a\beta(t)^2 a\right) \\ &= t\left[\beta(t)^2, a\right] a. \end{aligned}$$

We now use equation (3.45) and equivalent (3.41) to obtain

$$\begin{aligned} R(t) - \widetilde{R}(t) &= t^2\tau\beta(t)^2[a,B]\beta(t)^2 a \\ &\approx t^2\tau\beta(t)^2 b^2 m_3 \beta(t)^2 b m_1. \end{aligned}$$

Let $\mathbf{p} = \boldsymbol{\beta(t)}^2\mathbf{b}^2\mathbf{m_3}\boldsymbol{\beta(t)}^2\mathbf{bm_1} \in \boldsymbol{\mathcal{Q}}_{4,3}$, so we can apply the sandwich lemma, which gives the following estimate:

$$\left\| R(t) - \widetilde{R}(t) \right\| \leq Ct^2 t^{-3/2} = C\sqrt{t}.$$

Assuming that $\left\| R(t)^{n-1} - \widetilde{R}(t)^{n-1} \right\| \leq C_{n-1}\sqrt{t}$, we show the same result for n, writing:

$$R(t)^n - \widetilde{R}(t)^n = R(t)\left(R(t)^{n-1} - \widetilde{R}(t)^{n-1}\right) + \left(R(t) - \widetilde{R}(t)\right)\widetilde{R}(t)^{n-1}$$

and we use the result for $n = 1$ and $n-1$ and the fact that $R(t)$ and $\widetilde{R}(t)$ are bounded to conclude. □

Let us prove an estimate on a commutator built on $\widetilde{R}$ and R:

LEMMA 3.11. *Let* $k \in \mathbb{N}$, $t \in [0,1]$ *and* $s \in [0, n_k t]$, *then*

$$\left\| \left[\widetilde{R}(s), R(t) - \widetilde{R}(t)\right] \right\| \leq Ct. \tag{3.47}$$

PROOF. We calculate this term by developing the commutator which gives

$$\begin{aligned} &\left[\widetilde{R}(s), R(t) - \widetilde{R}(t)\right] = st^2\tau\left[a\beta(s)^2 a, \beta(t)^2[a,B]\beta(t)^2 a\right] \\ &= \Bigg(\underbrace{st^2\tau\left[a, \beta(t)^2[a,B]\beta(t)^2 a\right]\beta(s)^2 a}_{C_1} + \underbrace{st^2\tau a\left[\beta(s)^2, \beta(t)^2[a,B]\beta(t)^2 a\right] a}_{C_2} \\ &\quad + \underbrace{st^2\tau a\beta(s)^2\left[a, \beta(t)^2[a,B]\beta(t)^2 a\right]}_{C_3}\Bigg). \end{aligned}$$

We observe that the two terms C_1 and C_3 contain the same terms in different order. Thus it is sufficient to study C_1 and C_2. Let us develop them:

$$C_1 = \underbrace{st^2\tau[a,\beta(t)^2][a,B]\beta(t)^2a\beta(s)^2a}_{C_{1,1}} + \underbrace{st^2\tau\beta(t)^2\big[a,[a,B]\big]\beta(t)^2a\beta(s)^2a}_{C_{1,2}}$$
$$+ \underbrace{st^2\tau\beta(t)^2[a,B][a,\beta(t)^2]a\beta(s)^2a}_{C_{1,3}}$$

and

$$C_2 = \underbrace{st^2\tau a\beta(t)^2\big[\beta(s)^2,[a,B]\big]\beta(t)^2A}_{C_{2,1}} + \underbrace{st^2\tau a\beta(t)^2[a,B]\beta(t)^2[\beta(s)^2,a]a}_{C_{2,2}}.$$

We also notice here that $C_{1,1}$, $C_{1,3}$ and $C_{2,2}$ contain the same terms in a different order. It remains to estimate $C_{1,1}$, $C_{1,2}$ and $C_{2,1}$.

We use equation (3.45) and equivalent (3.39) to estimate $C_{1,1}$:

$$C_{1,1} \approx st^3\tau^2\beta(t)^2[B,a]\beta(t)^2[a,B]\beta(t)^2bm_1\beta(s)^2bm_1$$

which, with the help of (3.41), is equivalent to

$$\approx st^3\tau^2D_{1,1},$$

where

$$D_{1,1} = \beta(t)^2b^2m_3\beta(t)^2b^2m_3\beta(t)^2bm_1\beta(s)^2bm_1;$$

we find that

$$\mathbf{D_{1,1} = \boldsymbol{\beta(t)}^2b^2m_3\boldsymbol{\beta(t)}^2b^2m_3\boldsymbol{\beta(t)}^2bm_1\boldsymbol{\beta(s)}^2bm_1 \in \mathcal{Q}_{8,6}},$$

and the sandwich lemma gives:

$$\|D_{1,1}\| \le Ct^{-3} \text{ and } \|C_{1,1}\| \le Ct^4t^{-3} = Ct.$$

Now, let us calculate $C_{1,2}$; using equation (3.45), equivalents (3.39) and (3.44), we infer that

$$\begin{aligned} C_{1,2} &\approx st^2\tau\beta(t)^2\big[a,[a,B]\big]\beta(t)^2bm_1\beta(s)^2bm_1 \\ &\approx st^2\tau D_{1,2}, \end{aligned}$$

with

$$D_{1,2} = \beta(t)^2b^2m_6\beta(t)^2bm_1\beta(s)^2bm_1.$$

Since

$$\mathbf{D_{1,2} = \boldsymbol{\beta(t)}^2b^2m_6\boldsymbol{\beta(t)}^2bm_1\boldsymbol{\beta(s)}^2bm_1 \in \mathcal{Q}_{6,4}},$$

we can apply the sandwich lemma, which gives the estimate:

$$\|C_{1,2}\| \le Ct^3t^{-2} = Ct.$$

In virtue of (3.45) and with the help of (3.43), we find

$$\begin{aligned} C_{2,1} &= s^2t^2\tau^2a\beta(t)^2\beta(s)^2\big[[a,B],B\big]\beta(s)^2\beta(t)^2A \\ &\approx s^2t^2\tau^2D_{2,1}, \end{aligned}$$

with

$$D_{2,1} = bm_1\beta(t)^2\beta(s)^2b^3m_5\beta(s)^2\beta(t)^2b^3m_1^3;$$

since

$$\mathbf{D_{2,1}} = \mathbf{bm_1}\boldsymbol{\beta(t)}^2\boldsymbol{\beta(s)}^2\mathbf{b}^3\mathbf{m_5}\boldsymbol{\beta(s)}^2\boldsymbol{\beta(t)}^2\mathbf{b}^3\mathbf{m_1}^3 \in \boldsymbol{\mathcal{Q}}_{8,6},$$

the sandwich lemma gives:

$$\|C_{2,1}\| \leq Ct^4t^{-3} = Ct.$$

Thus, the proof of the lemma is complete. □

LEMMA 3.12. *Let $k \in \mathbb{N}$, $t \in [0,1]$ and $s \in [0, n_k t]$, then for all $n \in \mathbb{N}$,*

$$\left\|[R(t), \widetilde{R}(s)^n]\right\| = O(\sqrt{t}). \tag{3.48}$$

PROOF. We develop this commutator, writing it as a sum of terms, each of which containing powers of $\widetilde{R}(s)$, which are bounded, and a term $[R(t), \widetilde{R}(s)]$, which we must estimate. An intermediate step consists in proving

$$\left\|[R(t), \widetilde{R}(s)]\right\| = O(\sqrt{t}). \tag{3.49}$$

Developing the commutator, we find

$$\begin{aligned}[R(t), \widetilde{R}(s)] &= ts[\beta(t)^2A, a\beta(s)^2a] \\ &= ts\big(\beta(t)^2a[A, \beta(s)^2]a + [\beta(t)^2, a]\beta(s)^2a^3 \\ &\quad + a\beta(s)^2[\beta(t)^2, a]A\big).\end{aligned}$$

We transform this relation using equation (3.45):

$$\begin{aligned}[R(t), \widetilde{R}(s)] = ts\big(&s\tau\beta(t)^2a\beta(s)^2[B, A]\beta(s)^2a \\ &+ t\tau\beta(t)^2[a, B]\beta(t)^2\beta(s)^2a^3 \\ &+ t\tau a\beta(s)^2\beta(t)^2[a, B]\beta(t)^2A\big);\end{aligned}$$

thus, using equations (3.41) and (3.42), we find

$$\begin{aligned}[R(t), \widetilde{R}(s)] \approx ts\big(&s\tau\beta(t)^2bm_1\beta(s)^2b^3m_4\beta(s)^2bm_1 \\ &+ t\tau\beta(t)^2Bm_3\beta(t)^2\beta(s)^2b^3m_1^3 \\ &+ t\tau bm_1\beta(s)^2\beta(t)^2Bm_3\beta(t)^2Bm_1^2\big).\end{aligned}$$

This proves that there exists $\mathbf{p} \in \boldsymbol{\mathcal{Q}}_6$ satisfying $\deg(\mathbf{p}, \mathbf{b}) = 5$ such that

$$\left\|[R(t), \widetilde{R}(s)]\right\| \leq Ct^3 \left\|\Phi(\mathbf{p})\right\|.$$

Therefore the sandwich lemma gives (3.49). □

Let us now prove the last estimate involving R and $\widetilde{R}$:

LEMMA 3.13. *Let $k \in \mathbb{N}$, $t \in [0,1]$ and $s \in [0, n_k t]$, then for all $n \in \mathbb{N}$ and for all $m \in \mathbb{N}$,*

$$\left\| \left[\widetilde{R}(s)^n, \widetilde{R}(t)^m \right] \right\| = O(\sqrt{t}).$$

PROOF. We develop as for (3.48), we write the commutator as a sum of terms involving powers of $\widetilde{R}(s)$ and of $\widetilde{R}(t)$, which are bounded, and the term $[\widetilde{R}(s), \widetilde{R}(t)]$, which is bounded by $O(\sqrt{t})$ thanks to (3.49) and (3.46). □

3.4.3. *Stability of the unsymmetrized form P_k.* Define T_k by

$$T_k = P_k - \widetilde{P}_k. \tag{3.50}$$

We first need an estimate on the norm of T_k:

LEMMA 3.14. *There exists $C > 0$ such that for all $t \in [0,1]$*

$$\|T_k(t)\| \leq C\sqrt{t}. \tag{3.51}$$

PROOF. Using expression (3.15) of $\widetilde{P}_k$ and expression (3.12) of P_k, we can see that T_k given by (3.50) is a finite sum of terms like $R(t/n_j)^i - \widetilde{R}(t/n_j)^i$ with $i \geq 1$.

We therefore use estimate (3.46) of Lemma 3.10 to conclude the proof. □

Estimate (3.51) is of course insufficient to infer the stability; we need an estimate on $\widetilde{P}_k^* T_k + T_k^* \widetilde{P}_k$:

LEMMA 3.15. *There exists $C > 0$ such that for all $t \in [0,1]$*

$$\left\| \widetilde{P}_k^* T_k + T_k^* \widetilde{P}_k \right\| \leq Ct.$$

PROOF. Using expression (3.15) of $\widetilde{P}_k$ and the remark in the proof of Lemma 3.14 about the expression of T_k, we can see that the expression $\widetilde{P}_k^* T_k + T_k^* \widetilde{P}_k$ is a finite sum of terms like

$$E_{m,n} = (\widetilde{R}^*(s))^n \left(R(t)^m - \widetilde{R}(t)^m \right) + \left(R^*(t)^m - \widetilde{R}^*(t)^m \right) (\widetilde{R}(s))^n \tag{3.52}$$

with $m \geq 1$ and $n \in \mathbb{N}$.

As in lemma 3.14, it is sufficient to show that for all $t \in [0,1]$, for all $s \in [0, n_k t]$ and for all $n, m \in \mathbb{N}$,

$$\|E_{m,n}\| \leq Ct. \tag{3.53}$$

We will show this by induction on m, with an embedded induction on n.

For $m = 1$ and $n = 1$, using expressions (3.5) and (3.13), we can write

$$\begin{aligned} E_{1,1} &= tsa\beta(s)a(\beta(t)A - a\beta(t)a) + ts(A\beta(t) - a\beta(t)a)a\beta(s)a \\ &= ts\beta(s)a[\beta(t), a]a + tsa[a, \beta(t)]a\beta(s)a. \end{aligned}$$

Using (3.45), we can rewrite the expression $E_{1,1}$ as follows:
(3.54)

$$E_{1,1} = t^2 s\tau a\beta(s)^2 a\beta(t)^2[a,B]\beta(t)^2 \overset{\downarrow}{\times} a - t^2 s\tau a\beta(t)^2[a,B]\beta(t)^2 a\beta(s)^2 a.$$

We now write this difference as a sum of terms, each of which contains a commutator. For this purpose we commute as many terms as necessary in the first term of expression (3.54) : we move the $\beta(s)^2$ term between the last $\beta(t)^2$ term and the last a term, i.e. at the place enhanced by the cross with an arrow above; we also notice that $\beta(t)$ and $\beta(s)$ commute.

In the first term of the right hand side of (3.54), we commute the $\beta(s)^2$ term and the second a term. Define

$$C_1 = t^2 s\tau a[\beta(s)^2, a]\beta(t)^2[a,B]\beta(t)^2 a;$$

then (3.54) is equal to
(3.55)

$$C_1 + t^2 s\tau a^2\beta(t)^2\beta(s)^2[a,B]\beta(t)^2 a - t^2 s\tau a\beta(t)^2[a,B]\beta(t)^2 a\beta(s)^2 a.$$

Then, in the second term of (3.55), we commute the $\beta(s)^2$ term with the $[a,B]$ term. Let

$$C_2 = t^2 s\tau a^2\beta(t)^2\big[\beta(s)^2, [a,B]\big]\beta(t)^2 a;$$

therefore (3.55) is equal to

$$\begin{aligned} C_1 + C_2 + t^2 s\tau a^2\beta(t)^2[a,B]\beta(t)^2 \overset{\downarrow}{\times} \beta(s)^2 a \\ - t^2 s\tau a\beta(t)^2[a,B]\beta(t)^2 a\beta(s)^2 a. \end{aligned} \tag{3.56}$$

In the third term of (3.56), we commute one of the first a terms moving it to the left of $\beta(s)^2$ to the place with the cross and the arrow; we first commute this a term with the first $\beta(t)^2$ term. Let

$$C_3 = t^2 s\tau a[a, \beta(t)^2][a,B]\beta(t)^2\beta(s)^2 a;$$

therefore (3.56) is equal to

$$\begin{aligned} C_1 + C_2 + C_3 + t^2 s\tau a\beta(t)^2 a[a,B]\beta(t)^2\beta(s)^2 a \\ - t^2 s\tau a\beta(t)^2[a,B]\beta(t)^2 a\beta(s)^2 a. \end{aligned} \tag{3.57}$$

Next, we commute the same a term as above with the $[a,B]$ term. Let

$$C_4 = t^2 s\tau a\beta(t)^2\big[a, [a,B]\big]\beta(t)^2\beta(s)^2 a;$$

therefore (3.57) is equal to

$$\begin{aligned} C_1 + C_2 + C_3 + C_4 + t^2 s\tau a\beta(t)^2[a,B]a\beta(t)^2\beta(s)^2 a \\ - t^2 s\tau a\beta(t)^2[a,B]\beta(t)^2 a\beta(s)^2 a. \end{aligned} \tag{3.58}$$

Finally, recognizing a commutator, we let

$$C_5 = t^2 s\tau a\beta(t)^2[a,B][a, \beta(t)^2]\beta(s)^2 a,$$

and (3.58) is equal to

$$E_{1,1} = C_1 + C_2 + C_3 + C_4 + C_5.$$

We now estimate these five terms one after another. We use (3.45) to calculate C_1:

$$\begin{aligned} C_1 &= t^2 s\tau a[\beta(s)^2, a]\beta(t)^2[a, B]\beta(t)^2 a \\ &= t^2 s\tau a\beta(s)^2[a, B]\beta(s)^2\beta(t)^2[a, B]\beta(t)^2 a. \end{aligned}$$

Therefore, using equivalents (3.39) and (3.41) and equation (3.45), we find

$$\begin{aligned} C_1 &\approx t^2 s^2\tau^2 bm_1\beta(s)^2 b^2 m_3\beta(s)^2\beta(t)^2 b^2 m_3\beta(t)^2 bm_1 \\ &= t^2 s^2\tau^2 D_1, \end{aligned}$$

with

$$D_1 = bm_1\beta(s)^2[a, B]\beta(s)^2\beta(t)^2[a, B]\beta(t)^2 bm_1$$

and

$$\mathbf{D_1 = bm_1\boldsymbol{\beta(s)}^2 b^2 m_3\boldsymbol{\beta(s)}^2\boldsymbol{\beta(t)}^2 b^2 m_3\boldsymbol{\beta(t)}^2 bm_1} \in \boldsymbol{\mathcal{Q}}_{8,6};$$

we may apply the sandwich lemma, which gives the estimate:

$$\|D_1\| \leq Ct^{-3} \text{ and } \|C_1\| \leq Ct^4 t^{-3} = Ct.$$

We use equation (3.45) and equivalents (3.39), (3.40) and (3.43) to estimate C_2:

$$\begin{aligned} C_2 &= t^2 s\tau a^2\beta(t)^2\big[\beta(s)^2, [a, B]\big]\beta(t)^2 a \\ &\approx t^2 s^2\tau^2 b^2 m_1^2\beta(t)^2\beta(s)^2\big[[a, B], B\big]\beta(s)^2\beta(t)^2 bm_1; \\ &= t^2 s^2\tau^2 D_2, \end{aligned}$$

where

$$D_2 = b^2 m_1^2\beta(t)^2\beta(s)^2 b^3 m_5\beta(s)^2\beta(t)^2 bm_1.$$

Therefore,

$$\mathbf{D_2 = b^2 m_1{}^2\boldsymbol{\beta(t)}^2\boldsymbol{\beta(s)}^2 b^3 m_5\boldsymbol{\beta(s)}^2\boldsymbol{\beta(t)}^2 bm_1} \in \boldsymbol{\mathcal{Q}}_{8,6},$$

and the sandwich lemma gives:

$$\|C_2\| \leq Ct^4 t^{-3} = Ct.$$

The expression

$$C_3 = t^2 s\tau a[a, \beta(t)^2][a, B]\beta(t)^2\beta(s)^2 a$$

contains exactly the same terms as C_1 in a different order, so that $\|C_3\| = O(t)$.

In virtue of equivalents (3.39) and (3.44),

$$\begin{aligned} C_4 &= t^2 s\tau a\beta(t)^2\big[a, [a, B]\big]\beta(t)^2\beta(s)^2 a \\ &\approx t^2 s\tau bm_1\beta(t)^2 b^2 m_6\beta(t)^2\beta(s)^2 bm_1 \\ &= t^2 s\tau D_4 \end{aligned}$$

with

$$\mathbf{D_4} = \mathbf{bm_1}\boldsymbol{\beta(t)}^2\mathbf{b}^2\mathbf{m_6}\boldsymbol{\beta(t)}^2\boldsymbol{\beta(s)}^2\mathbf{bm_1} \in \boldsymbol{\mathcal{Q}}_{6,4};$$

the sandwich lemma gives:

$$\|C_4\| \leq Ct^3t^{-2} = Ct.$$

Finally,

$$C_5 = t^2 s\tau a\beta(t)^2[a,B][a,\beta(t)^2]\beta(s)^2 a$$

contains exactly the same terms as C_1, and thus $\|C_5\| = O(t)$. This concludes the case $m = 1$ and $n = 1$.

Assume the result holds for $m = 1$ and $n-1$; then we express $E_{1,n}$ in terms of $E_{1,n-1}$ in order to apply the induction assumption:

(3.59)

$$\begin{aligned} E_{1,n} &= (\widetilde{R}^*(s))^{n-1}\Big(\widetilde{R}^*(s)\big(R(t)-\widetilde{R}(t)\big) - \big(R(t)-\widetilde{R}(t)\big)\widetilde{R}(s)\Big) \\ &+ \Big((\widetilde{R}^*(s))^{n-1}\big(R(t)-\widetilde{R}(t)\big) + \big(R^*(t)-\widetilde{R}^*(t)\big)(\widetilde{R}(s))^{n-1}\Big)\widetilde{R}(s); \end{aligned}$$

we want to prove that (3.59) is an $O(t)$.

The second of the two terms in the right hand side of (3.59) is equal to $E_{1,n-1}\widetilde{R}(s)$ and thus is an $O(t)$ thanks to the induction assumption and the fact that $\widetilde{R}(s)$ is bounded.

Let us study the first term

$$(\widetilde{R}^*(s))^{n-1}\Big(\widetilde{R}^*(s)\big(R(t)-\widetilde{R}(t)\big) - \big(R(t)-\widetilde{R}(t)\big)\widetilde{R}(s)\Big). \tag{3.60}$$

The term $(\widetilde{R}^*(s))^{n-1}$ is bounded; as $\widetilde{R}(s)$ is self-adjoint, we can write

$$\widetilde{R}^*(s)\big(R(t)-\widetilde{R}(t)\big) - \big(R(t)-\widetilde{R}(t)\big)\widetilde{R}(s) = \big[\widetilde{R}(s), R(t)-\widetilde{R}(t)\big].$$

Using estimate (3.47) proved in Lemma 3.11, we see that (3.60) is an $O(t)$ and this concludes the proof of the case $m = 1$.

Assume that (3.53) holds for all $n \in \mathbb{N}$ and some index $m-1$, and let us show it for the index m. Thus, we study

(3.61)

$$E_{m,n} = (\widetilde{R}^*(s))^n\left(R(t)^m - \widetilde{R}(t)^m\right) + \left(R^*(t)^m - \widetilde{R}^*(t)^m\right)(\widetilde{R}(s))^n.$$

Using the fact that

$$\begin{aligned} R^*(t)^m - \widetilde{R}^*(t)^m &= \left(R^*(t)^{m-1} - \widetilde{R}^*(t)^{m-1}\right)R^*(t) \\ &+ \widetilde{R}^*(t)^{m-1}\big(R^*(t)-\widetilde{R}^*(t)\big), \end{aligned}$$

and that the analogous equality holds replacing R^* with R and $\widetilde{R}^*$ with $\widetilde{R}$, we rewrite (3.61) as

(3.62)
$$(\widetilde{R}^*(s))^n\Big(\big(R(t)^{m-1}-\widetilde{R}(t)^{m-1}\big)R(t)+\widetilde{R}(t)^{m-1}\big(R(t)-\widetilde{R}(t)\big)\Big)$$
$$+\Big(\big(R^*(t)^{m-1}-\widetilde{R}^*(t)^{m-1}\big)R^*(t)+\widetilde{R}^*(t)^{m-1}\big(R^*(t)-\widetilde{R}^*(t)\big)\Big)(\widetilde{R}(s))^n.$$

Define

$$\begin{aligned}C_1 =&(\widetilde{R}^*(s))^n\big(R(t)^{m-1}-\widetilde{R}(t)^{m-1}\big)R(t)\\&+\big(R^*(t)^{m-1}-\widetilde{R}^*(t)^{m-1}\big)R^*(t)(\widetilde{R}(s))^n\end{aligned}$$

and

$$\begin{aligned}C_2 =&\ (\widetilde{R}^*(s))^n\widetilde{R}(t)^{m-1}\big(R(t)-\widetilde{R}(t)\big)\\&+\widetilde{R}^*(t)^{m-1}\big(R^*(t)-\widetilde{R}^*(t)\big)(\widetilde{R}(s))^n.\end{aligned}$$

Then, $E_{m,n}=C_1+C_2$. We estimate separately C_1 and C_2; in C_1, we display $E_{m-1,n}$:

(3.63)
$$\begin{aligned}C_1 =&\Big(\widetilde{R}^*(s)^n\big(R(t)^{m-1}-\widetilde{R}(t)^{m-1}\big)\\&+\big(R^*(t)^{m-1}-\widetilde{R}^*(t)^{m-1}\big)\widetilde{R}(s)^n\Big)R(t)\\&+\big(R^*(t)^{m-1}-\widetilde{R}^*(t)^{m-1}\big)\big(R^*(t)\widetilde{R}(s)^n-\widetilde{R}(s)^nR(t)\big)\\=&\ E_{m-1,n}R(t)+\big(R^*(t)^{m-1}-\widetilde{R}^*(t)^{m-1}\big)\big(R^*(t)\widetilde{R}(s)^n-\widetilde{R}(s)^nR(t)\big).\end{aligned}$$

Using the induction assumption and the fact that $R(t)$ is bounded, we see that the first term of the right hand side of (3.63) is $O(t)$.

The second term of (3.63)

$$R^*(t)^{m-1}-\widetilde{R}^*(t)^{m-1}$$

is $O(\sqrt{t})$ as proved in Lemma 3.10 and it suffices now to prove that

$$R^*(t)\widetilde{R}(s)^n-\widetilde{R}(s)^nR(t)=(R^*(t)-R(t))\widetilde{R}(s)^n+[R(t),\widetilde{R}(s)^n] \tag{3.64}$$

is $O(\sqrt{t})$. This is clear for the first term of the right hand side of (3.64) using (3.46); we use estimate (3.48) of Lemma 3.12 for the second term of the right hand side of (3.64). Thus, we have proved the expected estimate for (3.64) and thus for C_1.

We estimate finally C_2, where we make $E_{1,n}$ visible in the right hand side as follows:

$$\begin{aligned}
C_2 &= \widetilde{R}^*(s)^n \widetilde{R}(t)^{m-1}\big(R(t)-\widetilde{R}(t)\big) + \widetilde{R}^*(t)^{m-1}\big(R^*(t)-\widetilde{R}^*(t)\big)\widetilde{R}(s)^n \\
&= \Big(\big(\widetilde{R}^*(s)^n \widetilde{R}(t)^{m-1} - \widetilde{R}^*(t)^{m-1}\widetilde{R}^*(s)^n\big)\big(R(t)-\widetilde{R}(t)\big)\Big) \\
&+ \widetilde{R}^*(t)^{m-1}\Big(\big(R^*(t)-\widetilde{R}^*(t)\big)\widetilde{R}(s)^n + \widetilde{R}^*(s)^n\big(R(t)-\widetilde{R}(t)\big)\Big) \\
&= \Big(\big(\widetilde{R}^*(s)^n \widetilde{R}(t)^{m-1} - \widetilde{R}^*(t)^{m-1}\widetilde{R}^*(s)^n\big)\big(R(t)-\widetilde{R}(t)\big)\Big) \\
&\quad + \widetilde{R}^*(t)^{m-1} E_{1,n}(t).
\end{aligned}$$

The second term of the last expression of C_2 is an $O(t)$, thanks to the case $m=1$. Let us study the first term; as $\widetilde{R}(t)$ is self-adjoint, the first term is equal to

$$\big[\widetilde{R}(s)^n, \widetilde{R}(t)^{m-1}\big]\big(R(t)-\widetilde{R}(t)\big);$$

thanks to equation (3.46) we have

$$\Big\|\big(R(t)-\widetilde{R}(t)\big)\Big\| = O(\sqrt{t})$$

and thanks to Lemma 3.13, we have

$$\Big\|\big[\widetilde{R}(s)^n, \widetilde{R}(t)^{m-1}\big]\Big\| = O(\sqrt{t}).$$

This concludes the proof. □

Now with the help of Theorem 3.5 and of Lemmas 3.14 and 3.15, we can complete the proof of our main result, i.e. the proof of the stability of the unsymmetrized form of RSS:

THEOREM 3.16. *For all $k \in \mathbb{N}$, for any choice of integers $1 \le n_1 < n_2 < \cdots < n_k$, there exists $\tau_0 > 0$ such that for all $\tau \ge \tau_0$, P_k is stable, and more precisely, there exists $C > 0$ such that for all $t \in [0,1]$ the following inequality holds in the sense of quadratic forms:*

$$-(1+Ct)\mathbf{1} \prec P_k(t) \prec (1+Ct)\mathbf{1}. \tag{3.65}$$

PROOF. From definition (3.50) of T_k, we have $P_k = \widetilde{P}_k + T_k$. Therefore, we can write

$$\begin{aligned}
\|P_k\|^2 = P_k^* P_k &= \Big(\widetilde{P}_k + T_k\Big)^* \Big(\widetilde{P}_k + T_k\Big) \\
&= \|\widetilde{P}_k\|^2 + \Big(\widetilde{P}_k^* T_k + T_k^* \widetilde{P}_k\Big) + \|T_k\|^2.
\end{aligned}$$

Theorem 3.5 enables us to estimate $\|\widetilde{P}_k\|^2$ by 1; thanks to Lemma 3.15, we estimate the term $\Big(\widetilde{P}_k^* T_k + T_k^* \widetilde{P}_k\Big)$ by Ct and thanks to Lemma 3.14, we estimate $\|T_k\|^2$ by Ct. This completes the proof of the theorem. □

4. Stabilité, convergence et ordre du schéma par régularisation du résidu et de ses extrapolations en norme d'énergie

Stability, convergence and order of the extrapolations of the Residual Smoothing Scheme in energy norm

Magali Ribot and Michelle Schatzman

Abstract : *We show that the RSS scheme which amounts to including a preconditioning of the spatial discretization in the time discretization is stable, convergent and of order one in energy norm. We also prove that its k-th Richardson extrapolation is stable and of order k.*

4.1. Introduction. The time integration of parabolic systems of equations is dominated by the dilemma between explicit methods, subject to the Courant-Friedrichs-Lewy (CFL) condition and implicit methods which require the use of efficient solvers, and make use of preconditioners.

Preconditioners are often left for the computer science and implementation side of scientific computation; for elliptic problems, preconditioners have been actively studied, the aim being to obtain a better convergence rate for iterative methods. In the case of time dependent problems, most of the literature considers that after applying the method of lines (i.e. semidiscretization in time), the preconditioning for time dependent problems is reduced to preconditioning for space dependent problems with a matrix $I + \Delta t A$ instead of A, like in Bornemann [**17, 18**], Mulholland and Sloan [**85**] or Brown and Woodward [**21**]. Then, usually there is no estimate of the error committed when time integration is performed with the help of a preconditioner.

In this article, we consider instead that preconditioning is an essential step from the analytical and numerical points of view, and we give a convergence and error analysis for a class of time integration schemes. More precisely, let A be a self-adjoint operator in a Hilbert space; we assume that A is bounded from below and we consider the problem

$$\begin{cases} \dfrac{du}{dt} + Au = f(t), \\ \qquad u(0) = u_0. \end{cases} \tag{4.1}$$

Without loss of generality, we may assume that for all u in the domain $D(A)$ of A, we have

$$(Au, u) \geq |u|^2 \tag{4.2}$$

Indeed, if A is bounded from below, there exists C in $\mathbb{R}$ such that

$$(Ax, x) \geq C|x|^2,$$

we set $v = ue^{-\lambda t}$ in (3.1) and we obtain

$$\begin{cases} \dfrac{dv}{dt} + (A+\lambda)v = 0, \\ \qquad\qquad v(0) = u_0, \end{cases}$$

that is to say a system analogous to (3.1) with $\widetilde{A} = A + \lambda$ instead of A. We choose λ such that $C + \lambda \geq 1$ and therefore inequality (3.2) holds for $\widetilde{A}$.

Denote by V the closure of $D(A)$ for the prehilbertian norm (Au, u). Assume that B is a self – adjoint unbounded operator which has the same domain as A and which satisfies

$$c^{-1}(Bu, u) \leq (Au, u) \leq c(Bu, u) \tag{4.3}$$

for some strictly positive constant c.

The residual smoothing scheme has been considered in Averbuch et al. [**4**] as an alternative to the backward Euler scheme; it is given by

$$\frac{U_{n+1} - U_n}{\Delta t} + \tau B(U_{n+1} - U_n) + AU_n = F_n, \tag{4.4}$$

where τ is a parameter which can be chosen to enforce stability.

In [**4**], the authors show that the scheme (4.4) is unconditionally stable for τ large enough if (4.3) holds.

We define $P(t)$ as

$$P(t) = \mathbf{1} - t\left(\mathbf{1} + t\tau B\right)^{-1} A. \tag{4.5}$$

Thanks to the definition (4.4) of the residual smoothing scheme, if F vanishes, we have

$$U^n = \left(P(\Delta t)\right)^n u_0.$$

Given any choice of integers $1 \leq n_1 < n_2 < \cdots < n_k$, the Richardson extrapolation of $P(t)$ is

$$P_k(t) = \sum_{j=1}^{k} \ell_j^k(0) P(t/n_j)^{n_j}$$

where ℓ_j^k are the elements of the Lagrange interpolation basis with knots $1/n_j$.

In this article, A and B are general operators, but we apply them in [**97**] to a spectral method preconditioned by a finite element method. In [**97**], we prove that the matrices A and B are equivalent in that particular case and we calculate the consistency error. The preconditioning of spectral methods by finite differences or finite elements method has been widely studied in the literature, theoretically by Orszag [**87**] or Haldenwang et al. [**63**] and numerically, for example, by Canuto and Quarteroni [**24**] or Deville and Mund [**53, 54**]. However, these articles only deal with elliptic problems.

In [**100**], we have already proved the unconditional stability of the Residual Smoothing Scheme (4.4) and its extrapolations provided that τ is large enough for the usual norm. In this article, we consider the same results in energy norm. The energy norm is defined by $|x|_A = (x^*Ax)^{1/2}$; this norm is convenient since to study the energy norm of $P(t)$, we have to study the operator $A^{-1/2}P^*APA^{-1/2}$ which is self-adjoint unlike P.

We therefore define an order relation between two operators to make precise the notion of two equivalent operators as in equation (4.3). We then prove that if A is dominated by B and for τ large enough the Residual Smoothing Scheme is stable, that is to say that the energy norm of P is bounded by 1. We then extend this result to the Richardson's extrapolations of P. If furthermore A is equivalent to B, we can prove the convergence of RSS and of its Richardson's extrapolations for τ large enough. To show this result, we use the theory of approximation of continuous semi-groups by discrete semi-groups described in Kato [**72**]. Finally, we show that if $A^{-k}B^k$ and $B^{-k}A^k$ are bounded, the scheme defined with the help of P_k is of order k; we use for that purpose the article of Dia and Schatzman [**55**] dealing with the algebraic point of view on extrapolation.

Let us explain the organization of the article: in section 4.2, we define a relation between operators and we study its properties; in section 4.3, we define some regularity spaces related to operators A and B and we give conditions for their equivalence. After these preliminary results, in section 4.4, we prove the stability of RSS in energy norm and we extend the proof of the stability to the extrapolations of RSS in section 4.5. Then, in section 4.6, we study the conditional stability; in section 4.7, we prove the convergence of RSS and its extrapolations and eventually in section 4.8, the orders of these schemes.

4.2. An order on self-adjoint operators. Let us first define in a very precise way the equivalence of two operators and study the properties of the order relation. This relation has already been introduced and studied in [**100**] but we recall here the definition and the properties for the reader's convenience.

In this article, we denote by **1** the identity operator in any vector space. We recall that every self-adjoint operator T in a Hilbert space possesses a spectral decomposition

$$T = \int_{\mathbb{R}} \lambda dP(\lambda),$$

where $dP(\lambda)$ is the spectral measure associated to T. We will say that a self-adjoint operator T is positive semi-definite if for all $x \in D(T)$, $x^*Tx \geq 0$. If T is positive semi-definite, the square root of T is

defined by

$$\sqrt{T} = \int_{\mathbb{R}} \sqrt{\lambda} dP(\lambda).$$

We define as follows a partial order relation between self-adjoint and bounded from below operators in a Hilbert space H:

(4.6)

$$T_1 \prec T_2 \Longrightarrow D(T_2) \subset D(T_1) \text{ and } \forall x \in D(T_2), \quad x^* T_1 x \leq x^* T_2 x.$$

We will need also a less precise order relation, when $T_2 \geq 0$:

$$T_1 \precsim T_2 \Longrightarrow \exists r \in (0, \infty) : \quad T_1 \prec r T_2. \tag{4.7}$$

If T_2 is positive and no assumption on the sign of the self-adjoint operator T_1 is made, definitions (4.6) and (4.7) still make sense; moreover, we may define for $T_1 \precsim T_2$:

$$[T_1 : T_2] = \sup_{T_2 x \neq 0} \frac{x^* T_1 x}{x^* T_2 x}. \tag{4.8}$$

With this definition, we always have for $T_2 \geq 0$ and $T_1 \precsim T_2$

$$T_1 \prec [T_1 : T_2] T_2.$$

We define the relations $\succ$ and $\succsim$ to be the opposite relations to $\prec$ and $\precsim$; if T_1 and T_2 are positive self-adjoint operators in H, the relation $T_1 \sim T_2$ means that $T_1 \precsim T_2 \precsim T_1$.

We may relate these equivalence relations to algebraic operations; in particular, if S is a self-adjoint operator which is bounded from below, it is plain that

$$T_1 \prec T_2 \Longrightarrow T_1 + S \prec T_2 + S.$$

If S is any bounded operator from a Hilbert space H_1 to H, and if the domain of $S^* T_j S$ for $j = 1, 2$ is defined as $S^{-1} D(T_j)$, we have also:

$$T_1 \prec T_2 \Longrightarrow S^* T_1 S \prec S^* T_2 S. \tag{4.9}$$

The proof is performed through the change of variable $x = Sy$.

Another important fact is the following:

Lemma 4.1. *If T_1 and T_2 are positive self-adjoint and injective, then*

$$T_1 \prec T_2 \Longrightarrow T_2^{-1} \prec T_1^{-1}. \tag{4.10}$$

Proof. This can be deduced from the proof of Theorem VI.2.21 in Kato's book [**72**]. □

Observe that if $T_1 \prec T_2$ then for any powers $\alpha \in]0, 1[$, $T_1^\alpha \prec T_2^\alpha$. Indeed a formula of Balakrishnan in [**5**] which is given in Yosida's book [**125**] gives the representation of T^α:

$$x \in D(T) \Rightarrow T^\alpha x = \frac{\sin(\alpha\pi)}{\pi} \int_0^\infty \lambda^{\alpha-1} (\lambda \mathbf{1} + T)^{-1} T x d\lambda.$$

The relation

$$(\lambda\mathbf{1}+T)^{-1}T=\mathbf{1}-\lambda\,(\lambda\mathbf{1}+T)^{-1}$$

is classical; we infer from Lemma 4.1 that

$$(\lambda\mathbf{1}+T_1)^{-1}T_1\prec(\lambda\mathbf{1}+T_2)^{-1}T_2;$$

therefore, it is plain that $T_1^\alpha\prec T_2^\alpha$.

However $T_1\prec T_2$ does not imply $T_1^n\prec T_2^n$ for all n in $\mathbb{N}$; a counter-example is, for instance

$$T_1=\begin{pmatrix}2\varepsilon & 0\\ 0 & 2/\varepsilon\end{pmatrix},\quad T_2=\begin{pmatrix}\varepsilon & 1/2\\ 1/2 & 1/\varepsilon\end{pmatrix}.$$

The reader will check that for all positive ε, $T_1\succ T_2$, while for all small enough ε, it is not true that $T_1^2\succ T_2^2$. However, if the self-adjoint, positive operators T_1 and T_2 commute, and in particular if one of them is scalar, the conclusion is true and this can be checked simply with the help of the spectral theorem.

4.3. Some preliminary results. Let us introduce the domains of the powers of A and B; we define, as in [**96**], $\mathcal{H}_A^n$ for $n\in\mathbb{Z}$ as

$$\mathcal{H}_A^0=H$$

and

$$\mathcal{H}_A^n=A^{-n/2}H=D(A^{n/2}),\ \text{for } n\in\mathbb{N}^*$$

and

$$\mathcal{H}_A^n=\big(\mathcal{H}_A^{-n}\big)^*,\ \text{for } n\in\mathbb{Z}\setminus\mathbb{N},$$

where the star denote the dual space.

The norm over $\mathcal{H}_A^n$ is defined by

$$|u|_{\mathcal{H}_A^n}=|A^{n/2}u|_0.$$

The same definitions hold replacing A by B. We will write for simplicity

$$\mathcal{H}^0=\mathcal{H}_A^0=\mathcal{H}_B^0.$$

We have the following inclusions

$$\cdots\mathcal{H}_A^{-n}\supset\cdots\mathcal{H}_A^{-2}\supset\mathcal{H}_A^{-1}\supset\mathcal{H}_A^0=H\supset\mathcal{H}_A^1\supset\mathcal{H}_A^2\cdots\supset\mathcal{H}_A^n\cdots$$

and

$$\cdots\mathcal{H}_B^{-n}\supset\cdots\mathcal{H}_B^{-2}\supset\mathcal{H}_B^{-1}\supset\mathcal{H}_B^0=H\supset\mathcal{H}_B^1\supset\mathcal{H}_B^2\cdots\supset\mathcal{H}_B^n\cdots$$

and for $s>t$, $\mathcal{H}_A^s$ (resp. $\mathcal{H}_B^s$) is dense in $\mathcal{H}_A^t$ (resp. $\mathcal{H}_B^t$).

Indeed, if n is even, $n=2p$, for the density of $\mathcal{H}_A^n$ in $\mathcal{H}^0$, it suffices to consider

$$x(t)=\frac{p!}{t^p}\int_0^p\int_0^{t_1}\cdots\int_0^{t_{p-1}}e^{-t_pA}x\,dt_p\cdots dt_1$$

which belongs to $D(A^p)$ and converges to $x(0) = x$ as t tends to zero. Moreover, if n is odd, $\mathcal{H}_A^n$ contains $\mathcal{H}_A^{n+1}$ which is dense in $\mathcal{H}^0$ from the previous case.

Moreover, $A \sim B$ implies the equality $\mathcal{H}_A^1 = \mathcal{H}_B^1$ and the equivalence of the norms $|A^{1/2}x|_0$ and $|B^{1/2}x|_0$ over $\mathcal{H}_A^1$. Let us give now some conditions for the isomorphism between $\mathcal{H}_A^n$ and $\mathcal{H}_B^n$.

LEMMA 4.2. *For $n \in \mathbb{N}$, the following propositions are equivalent:*

(1) *$B^{-n/2}A^{n/2}$ and $A^{-n/2}B^{n/2}$ are bounded in $\mathcal{L}(\mathcal{H}^0)$,*
(2) *$\mathcal{H}_A^n$ and $\mathcal{H}_B^n$ are isomorphic and their norms are equivalent,*
(3) *$\mathcal{H}_A^{-n}$ and $\mathcal{H}_B^{-n}$ are isomorphic and their norms are equivalent.*

PROOF. Observe that *a priori* $A^{-n/2}B^{n/2}$ is defined on $\mathcal{H}_B^n$; hypothesis (1) states that this operator is bounded for the operator norm subordinate to the norm of $\mathcal{H}^0$; consequently, it admits a unique extension to all $\mathcal{H}^0$.

It is immediate by duality that (2) is equivalent to (3). Let us prove that hypothesis (1) implies (3).

Assume $y \in \mathcal{H}^0$ and $x = A^{-n/2}y \in \mathcal{H}_A^n \subset \mathcal{H}^0$. Since

$$\left\|B^{-n/2}A^{n/2}\right\|_{\mathcal{L}(\mathcal{H}^0)} \leq C_n,$$

we have

$$|B^{-n/2}y|_0 = |B^{-n/2}A^{n/2}x|_0 \leq C_n|x|_0 = C_n|A^{-n/2}y|_0,$$

that is to say

$$|y|_{\mathcal{H}_B^{-n}} \leq C_n\,|y|_{\mathcal{H}_A^{-n}}\,;$$

and thus by density of $\mathcal{H}^0$ in $\mathcal{H}_A^{-n}$

$$\mathcal{H}_A^{-n} \subset \mathcal{H}_B^{-n}.$$

The opposite inclusion is obtained by exchanging A and B.

To complete the proof, let us prove that hypothesis (3) implies (1): $A^{n/2}$ is an isomorphism from $\mathcal{H}^0$ to $\mathcal{H}_A^{-n}$ and $B^{n/2}$ is an isomorphism from $\mathcal{H}^0$ to $\mathcal{H}_B^{-n}$, thus using hypothesis (3), $B^{-n/2}A^{n/2}$ is an isomorphism from $\mathcal{H}^0$ to $\mathcal{H}^0$ and consequently the same holds true for $A^{-n/2}B^{n/2}$. □

LEMMA 4.3. *If for an integer n , $\mathcal{H}_A^n$ is isomorphic to $\mathcal{H}_B^n$, then for all $s \in \{-|n|, \cdots, |n|\}$, $\mathcal{H}_A^s$ is isomorphic to $\mathcal{H}_B^s$.*

PROOF. This is a result of interpolation; see, for example Definition 2.1 and Remark 2.3 of [**79**]. □

Define

$$\beta(t) = (\mathbf{1} + t\tau B)^{-1} \tag{4.11}$$

such that $P(t) = \mathbf{1} - t\beta(t)A$. We remark that here $\beta(t)$ is the square of the element $\beta(t)$ previously defined in [**100**].

Let us prove that it converges strongly to $\mathbf{1}$ as t tends to zero.

LEMMA 4.4. *If $B^{-n/2}A^{n/2}$ and $A^{-n/2}B^{n/2}$ are bounded in $\mathcal{L}(\mathcal{H}^0)$, then for all $s \in \{-n, \cdots, 0\}$,*

$$\beta(t) \to \mathbf{1} \text{ strongly in } \mathcal{H}_A^s \text{ as } t \to 0^+.$$

PROOF. Since $\beta(t) \in \mathcal{L}\left(\mathcal{H}_B^s, \mathcal{H}_B^{s+2}\right)$ and thanks to Lemma 4.2, we see that $\beta(t) \in \mathcal{L}(\mathcal{H}_A^s, \mathcal{H}_A^s)$.

We know already that $\beta(t)$ converges strongly in $\mathcal{H}^0$ to $\mathbf{1}$, i.e. for all $v \in \mathcal{H}^0$, $\beta(t)v \to v$ in $\mathcal{H}^0$ and that $\mathcal{H}^0$ is dense in $\mathcal{H}_A^s$; to conclude by density it suffices to prove that $\|\beta(t)\|_{\mathcal{L}(\mathcal{H}_A^s)}$ is bounded. This is clearly true, since in virtue of Lemma 4.3,

$$\|\beta(t)\|_{\mathcal{L}(\mathcal{H}_A^s)} \leq C_n \|\beta(t)\|_{\mathcal{L}(\mathcal{H}_B^s)};$$

the operator norm of $\beta(t)$ in $\mathcal{L}(\mathcal{H}_B^s)$ is equal to its operator norm in $\mathcal{L}(\mathcal{H}^0)$, giving therefore the conclusion. □

4.4. Stability of RSS. Let us prove the stability of the Residual Smoothing Scheme defined in equation (4.4).

We will systematically write $a = A^{1/2}$ and $b = B^{1/2}$.

Define the energy norm by

$$|x|_A = (x^* A x)^{1/2}.$$

This norm is the above defined norm over $\mathcal{H}_A^1 = D(a)$. The corresponding operator norm is

$$\|L\|_A^2 = \sup_{x \in \mathcal{H}_A^1} \frac{x^* L^* A L x}{x^* A x}.$$

It is clear that the energy operator norm of L is bounded iff the ordinary operator norm $\|a^{-1}L^*ALa^{-1}\|$ is bounded, where the double norm $\| \quad \|$ denotes from now on the operator norm $\mathcal{L}(H)$.

We remark that unlike the operator P, the operator $a^{-1}P^*APa^{-1}$ is self-adjoint, which simplifies a lot the proof.

We have the following stability result on (4.4):

THEOREM 4.5. *Let A and B be positive definite self-adjoint operators in H satisfying $A \precsim B$ and let $P(t)$ be defined by* (4.5)*. Then, for τ larger than $[A : B]/2$, the energy norm of $P(t)$ is at most equal to* 1.

PROOF. The energy operator norm of $P(t)$ is

$$\|P(t)\|_A = \|a^{-1}P(t)^*AP(t)a^{-1}\|^{1/2},$$

and a straightforward computation gives

$$a^{-1}P(t)^*AP(t)a^{-1} = \mathbf{1} - 2ta\beta a + t^2 (a\beta a)^2. \tag{4.12}$$

It is convenient to let

$$Q(t) = a^{-1}P(t)^*AP(t)a^{-1}. \tag{4.13}$$

It is clear that $Q(t)$ is semi definite positive. We see that $Q(t) \prec \mathbf{1}$ iff

$$2ta\beta a \succ t^2 (a\beta a)^2$$

and this will be true if

$$2\mathbf{1} \succ ta\beta a. \tag{4.14}$$

Let us check that for all $t > 0$ the following inequality holds:

$$ta\beta a \prec \frac{[A:B]}{\tau}\mathbf{1}. \tag{4.15}$$

We have indeed the inequalities

$$tA \prec [A:B]tB \prec (\mathbf{1} + t\tau B)\frac{[A:B]}{\tau};$$

if we apply (4.10), we find that

$$\frac{\tau}{[A:B]}\beta \prec t^{-1}A^{-1}. \tag{4.16}$$

We multiply (4.16) on the left and on the right by a and we find immediately that (4.15) holds. Therefore, if

$$\tau \geq [A:B]/2,$$

the inequality $Q(t) \prec \mathbf{1}$ will be satisfied, proving thus the stability of $P(t)$ in energy norm. □

4.5. The general proof of stability in energy norm of the extrapolation of RSS. Let us now extend the result of the previous section to the Richardon's extrapolations of RSS; we need before to show some algebraic lemmas. The lemmas of subsection 4.5.1 have already been proved in [**100**], but we recall them and their proof.

4.5.1. *A preliminary inequality.* Given k distinct strictly positive integers $1 \leq n_1 < n_2 < \cdots < n_j < \cdots < n_k$, we define the coefficients of Richardson's extrapolation as follows: let ℓ_j^k be the Lagrange basis relative to the nodes $1/n_j$, $1 \leq j \leq k$:

$$\ell_j^k(t) = \prod_{\{i:i\neq j\}} \frac{t - 1/n_i}{(1/n_j) - (1/n_i)}. \tag{4.17}$$

Some well-known choices for these nodes are

- the harmonic sequence $n_j = j$,
- the Romberg sequence $n_j = 2^j$,
- the Bulirsch sequence

$$1, 2, 3, 4, 6, 8, 12, 16, \cdots, 2^j, \frac{3}{2}2^j, 2^{j+1}, \frac{3}{2}2^{j+1}, \cdots.$$

By interpolation of $1, t, \cdots, t^{k-1}$, we have the equalities:

$$\sum_{j=1}^{k} \ell_j^k(t) = 1, \tag{4.18}$$

$$\forall p = 1, \cdots, k-1, \quad \sum_{j=1}^{k} \ell_j^k(t) \frac{1}{n_j^p} = t^p,$$

from which we infer that at $t = 0$

$$\forall p = 0, \cdots, k-1, \quad \sum_{j=1}^{k} \frac{\ell_j^k(0)}{n_j^p} = \delta_{0p}. \tag{4.19}$$

The following function

$$\phi_k(s,t) = \sum_{j=1}^{k} \frac{\ell_j^k(t)}{1 + s/n_j}$$

will play an essential role in our analysis; $\phi_k(s, \cdot)$ interpolates the function $f : t \mapsto 1/(1+st)$ at the points $1/n_j$, $1 \le j \le k$.

Lemma 4.6. *The function $s \mapsto \phi_k(s, 0)$ is strictly positive over $\mathbb{R}^+$.*

Proof. For any function g, denote by $g[x_1, \cdots, x_n]$ the divided difference of the function g at the knots $x_1, \cdots, x_n$.

We use Newton's form of interpolation:

$$\begin{aligned}\phi_k(s,t) = {} & f(1/n_1) + f[1/n_1, 1/n_2](t - 1/n_1) + \cdots \\ & + f[1/n_1, 1/n_2, \cdots, 1/n_k](t - 1/n_1)(t - 1/n_2) \cdots (t - 1/n_{k-1}).\end{aligned}$$

In particular,

$$\begin{aligned}\phi_k(s,0) = {} & f(1/n_1) - \frac{f[1/n_1, 1/n_2]}{n_1} + \frac{f[1/n_1, 1/n_2, 1/n_3]}{n_1 n_2} + \cdots \\ & + \frac{(-1)^{k-1} f[1/n_1, 1/n_2, \cdots, 1/n_k]}{n_1 n_2 \cdots n_{k-1}}.\end{aligned} \tag{4.20}$$

If S^j is the simplex

$$S^j = \{x \in (\mathbb{R}^+)^j : x_1 + \cdots + x_j \le 1\},$$

the divided differences are given by the integral representation

$$f[a_1, \cdots, a_{j+1}] = \int_{S^j} f^{(j)}(a_1 + t_1(a_2 - a_1) + \cdots + t_j(a_{j+1} - a_j)). \tag{4.21}$$

But in our case,

$$f^{(j)}(t) = \frac{j!(-s)^j}{(1+st)^{j+1}}. \tag{4.22}$$

The term $(-1)^{j-1} f[1/n_1, \cdots, 1/n_j]$ involves $f^{(j-1)}$ in the integral representation; it is therefore positive, and the lemma is proved. □

We need another algebraic fact:

LEMMA 4.7. *For all $k = 1, 2, \cdots$ the following identity holds:*

$$\sum_{j=1}^{k} n_j \ell_j^k(0) = \sum_{j=1}^{k} n_j. \tag{4.23}$$

PROOF. Write

$$T_k = \sum_{j=1}^{k} n_j \ell_j^k(0).$$

We infer from formula (4.17) that $\ell_j^k(0)$ is given as

$$\ell_j^k(0) = (-n_j)^{k-1} \prod_{\{i:1\le i\le k, i\neq j\}} \frac{1}{n_i - n_j}. \tag{4.24}$$

Therefore, we have the relation

$$T_k - T_{k-1} = n_k \ell_k^k(0) + \sum_{j=1}^{k-1} A_j$$

with

$$A_j = (-n_j)^{k-1}\left(n_j \prod_{\{i:1\le i\le k, i\neq j\}} \frac{1}{n_i - n_j} + \prod_{\{i:1\le i\le k-1, i\neq j\}} \frac{1}{n_i - n_j}\right).$$

We remark that the factor of $(-n_j)^{k-1}$ in A_j contains $k-2$ common factors; therefore, it is equal to

$$\left(\frac{n_j}{n_k - n_j} + 1\right) \prod_{\{i:1\le i\le k-1, i\neq j\}} \frac{1}{n_i - n_j},$$

and therefore

$$A_j = (-n_j)^{k-1} n_k \prod_{\{i:1\le i\le k, i\neq j\}} \frac{1}{n_i - n_j}$$

and with the help of (4.24),

$$A_j = n_k \ell_j^k(0);$$

therefore,

$$T_k - T_{k-1} = n_k \sum_{j=1}^{k} \ell_j^k(0).$$

which concludes the proof, thanks to (4.19). □

Lemma 4.6 says that the function $\psi_k(s) = \phi_k(s, 0)$ is non negative on $\mathbb{R}^+$; lemma 4.7 enables us to find an equivalent of ψ_k at infinity:

$$\psi_k(s) \sim \sum_{j=1}^{k} \frac{n_j \ell_j^k(0)}{s} = \frac{\sum_{j=1}^{k} n_j}{s},$$

and therefore, the following corollary holds:

COROLLARY 4.8. *There exist $c_k > 0$ and $C_k > 0$ such that for all $s > 0$:*

$$\frac{c_k}{1+s/n_k} \leq \sum_{j=1}^{k} \frac{\ell_j^k(0)}{1+s/n_j} \leq \frac{C_k}{1+s/n_k}. \tag{4.25}$$

4.5.2. *Proof of the stability of the extrapolation.* We introduce the notation

$$\beta_j(t) = \beta(t/n_j).$$

The purpose of this section is to show that the extrapolation of RSS is unconditionally stable in energy norm for large enough values of τ.

THEOREM 4.9. *Let A and B be positive definite self-adjoint operators in H satisfying $A \precsim B$. For all $k \in \mathbb{N}$ and for any choice of integers $1 \leq n_1 < n_2 < \cdots < n_k$, there exists $\tau_0 > 0$ such that for all $\tau \geq \tau_0$ the following estimate holds:*

$$\forall t > 0, \quad \|P_k(t)\|_A \leq 1.$$

PROOF. We expand $P(t/n_j)^{n_j}$ according to the binomial formula. Therefore, if $\binom{n}{p}$ is set equal to zero for $p < 0$ or $p > n$, we have

$$P_k(t) = \sum_{j=1}^{k} \ell_j^k(0) \sum_{i=0}^{n_k} (-1)^i \binom{n_j}{i} \left(\frac{t}{n_j}\beta_j A\right)^i.$$

If we define

$$p_0(t) = \mathbf{1}, \tag{4.26a}$$

$$p_1(t) = \sum_{j=1}^{k} \ell_j^k(0)\beta_j A, \tag{4.26b}$$

and for all $i = 2, \cdots, n_k$

$$p_i(t) = \sum_{\{j:1\leq j\leq k, n_j \geq i\}} \binom{n_j}{i} \ell_j^k(0) \frac{(\beta_j A)^i}{n_j^i}, \tag{4.26c}$$

the expression of P_k can be rewritten

$$P_k(t) = \sum_{i=0}^{n_k} (-t)^i p_i(t). \tag{4.27}$$

The energy norm of the operator P_k is equal to

$$\left\| a^{-1} \sum_{i=0}^{n_k} (-t)^i p_i(t)^* A \sum_{l=0}^{n_k} (-t)^l p_l(t) a^{-1} \right\|;$$

the operator inside the norm symbol can be rewritten as

$$\mathbf{1} - 2\sum_{j=1}^{k} t\ell_j^k(0) a\beta_j a + \sum_{i+l\geq 2} (-t)^{i+l} a^{-1} p_i(t)^* A p_l(t) a^{-1}.$$

As in the proof of Theorem 4.5, this operator is semi-definite positive. Therefore, the stability in energy norm will hold if

$$0 \prec 2\sum_{j=1}^{k} t\ell_j^k(0)a\beta_j a - \sum_{i+l\geq 2}(-t)^{i+l}a^{-1}p_i(t)^*Ap_l(t)a^{-1}.$$

We can deduce from equation (4.25) that

$$c_k a\beta_k a \prec \sum_{j=1}^{k}\ell_j^k(0)a\beta_j a \prec C_k a\beta_k a.$$

Therefore, it suffices to find values of τ for which

$$2tc_k a\beta_k a \succ \sum_{i+l\geq 2}(-t)^{i+l}a^{-1}p_i(t)^*Ap_l(t)a^{-1}. \tag{4.28}$$

Condition (4.28) holds if

$$2c_k\mathbf{1} \succ -\sum_{i+l\geq 2}(-t)^{i+l-1}\beta_k^{-1/2}A^{-1}p_i(t)^*Ap_l(t)A^{-1}\beta_k^{-1/2}.$$

Therefore, we have to estimate the operator norm of L_{iljr} given by

$$L_{iljr} = C_{iljr}M_{iljr}$$

with

$$C_{iljr} = \frac{\ell_j^k(0)\ell_r^k(0)}{n_j^i n_r^l}\binom{n_j}{i}\binom{n_r}{l},$$
$$M_{iljr} = (-t)^{i+l-1}\beta_k^{-1/2}A^{-1}(A\beta_j)^iA(\beta_r A)^lA^{-1}\beta_k^{-1/2}.$$

The terms M_{iljr} can be rewritten for $\min(i,l)\geq 1$

$$M_{iljr} = (-t)^{i+l-1}\beta_k^{-1/2}\beta_j^{1/2}\left(\beta_j^{1/2}A\beta_j^{1/2}\right)^{i-1}\beta_j^{1/2}A\beta_r^{1/2}\left(\beta_r^{1/2}A\beta_r^{1/2}\right)^{l-1}$$
$$\times\beta_r^{1/2}\beta_k^{-1/2}.$$

We infer from the obvious inequality

$$t\tau A \leq [A:B]\,(\mathbf{1}+t\tau B)$$

that

$$t\tau\beta^{1/2}A\beta^{1/2} \leq [A:B]\mathbf{1}.$$

Therefore, we have the estimate

$$\beta_j^{1/2}A\beta_j^{1/2} \prec \frac{[A:B]n_j}{\tau t}\mathbf{1}; \tag{4.29}$$

on the other hand, by the spectral theorem,

$$\|\beta_k^{-1/2}\beta_j^{1/2}\| \leq 1. \tag{4.30}$$

We deduce from the inequality

$$\|\beta_j^{1/2}A\beta_r^{1/2}\| \leq \|\beta_j^{1/2}A\beta_j^{1/2}\|^{1/2}\|\beta_r^{1/2}A\beta_r^{1/2}\|^{1/2}$$

the following estimate

$$\|\beta_j^{1/2} A \beta_r^{1/2}\| \leq \frac{[A:B]\sqrt{n_j n_r}}{\tau t}. \tag{4.31}$$

We put together the estimates (4.29), (4.30) and (4.31) and we find that

$$\|M_{iljr}\| \leq \left(\frac{[A:B]}{\tau}\right)^{i+l-1} n_j^{i-1/2} n_r^{l-1/2}$$

and therefore that

$$\|L_{iljr}\| \leq \left(\frac{[A:B]}{\tau}\right)^{i+l-1} n_j^{i-1/2} n_r^{l-1/2} |C_{iljr}|.$$

If i vanishes, the expression for M_{iljr} is even simpler:

$$M_{0l0r} = (-t)^{l-1} \beta_k^{-1/2} (\beta_r A)^l A^{-1} \beta_k^{-1/2},$$

and the norm of L_{0l0r} can be estimated by

$$\|L_{0l0r}\| \leq \left(\frac{[A:B]}{\tau}\right)^{l-1} n_r^{l-1} |C_{0l0r}|.$$

Let us write

$$\nu_{il} = \sum_{\substack{\{j:n_j \geq i\} \\ \{r:n_r \geq l\}}} n_j^{i-1/2} n_r^{l-1/2} |C_{iljr}| \text{ and } \nu_{0l} = \nu_{l0} = \sum_{\{r:n_r \geq l\}} n_r^{l-1} |C_{0l0r}|.$$

There is a finite number of terms to estimate, and therefore, a sufficient condition for (4.28) to hold is

$$2c_k \geq \sum_{\substack{0 \leq i \leq k \\ 0 \leq l \leq k \\ i+l \geq 2}} \nu_{il} \left(\frac{[A:B]}{\tau}\right)^{i+l-1}.$$

It is clear that for large enough values of τ, (4.28) is satisfied, which completes the proof of the theorem. □

4.6. Conditional stability. If the operator A is bounded and in particular in the finite dimension case, we may be interested by conditional stability results. We start with an improvement of Theorem 4.5.

LEMMA 4.10. *Let A and B be self-adjoint positive definite operators such that $A \precsim B$ and A is bounded. Then for all $\tau > 0$, $\|P(t)\|_A \leq 1$ if*

$$t\|A\| \left(1 - \frac{2\tau}{[A:B]}\right) \leq 2. \tag{4.32}$$

PROOF. It suffices to find a condition under which

$$ta\beta a \prec 2\mathbf{1};$$

using 4.1 and (4.9), it is realized provided that

$$tA \prec 2\left(\mathbf{1} + t\tau B\right). \tag{4.33}$$

Observe that

$$A \prec [A : B]B,$$

and therefore (4.33) will hold provided that

$$tA \prec 2\left(\mathbf{1} + t\tau[A : B]^{-1}A\right)$$

which is true if

$$t\left(1 - 2\tau[A : B]^{-1}\right)A \prec 2\mathbf{1}$$

and the conclusion is clear. □

Relation (4.32) is typical of a conditional stability condition, since it could have been obtained for the fully explicit scheme, corresponding to $\tau = 0$,

$$u_{n+1} = u_n - tAu_n,$$

where the stability condition reads

$$\|\mathbf{1} - tA\| \leq 1;$$

this inequality is satisfied under condition (4.32).

Let us prove now the conditional stability for the Richardson's extrapolation of RSS.

Lemma 4.11. *Under the assumption of Lemma 4.10, for all sequence of distinct positive integers $n_1 < n_2 < \cdots < n_k$, if A and B are positive semi-definite operators, A is bounded and $A \precsim B$, there exists $\varepsilon_k > 0$ such that*

$$t\|A\|\left(1 - \frac{\tau\varepsilon_k}{n_k[A : B]}\right) \leq \varepsilon_k$$

implies $\|P_k(t)\|_A \leq 1$.

Proof. As in the proof of Theorem 4.9, and with the same notations, it suffices to prove

$$\sum_{\substack{0\leq i\leq k\\ 0\leq l\leq k\\ i+l\geq 2}} \sum_{\substack{\{j:n_j\geq i\}\\ \{r:n_r\geq l\}}} |C_{iljr}|\|M_{iljr}\| \leq 2c_k.$$

We observe that

$$\left(1 + \frac{t\tau\|A\|}{[A : B]}\right)A \prec \|A\|(\mathbf{1} + t\tau B)$$

and therefore

$$A \prec \frac{\|A\|}{1 + t\tau\|A\|/[A : B]}(\mathbf{1} + t\tau B)$$

which implies immediately

$$\beta^{1/2}A\beta^{1/2} \prec \frac{\|A\|}{1 + t\tau\|A\|/[A : B]}\mathbf{1}.$$

Therefore, we have now the estimates

$$\|M_{iljr}\| \leq \left(\frac{t\|A\|}{1+t\tau\|A\|/n_j[A:B]}\right)^{i-1/2}\left(\frac{t\|A\|}{1+t\tau\|A\|/n_r[A:B]}\right)^{l-1/2}$$

and

$$\|M_{0l0r}\| \leq \left(\frac{t\|A\|}{1+t\tau\|A\|/n_r[A:B]}\right)^{l-1}.$$

We write

$$\nu_{i,l} = \sum_{\substack{\{j:n_j\geq i\}\\ \{r:n_r\geq l\}}} |C_{iljr}| \text{ and } \nu_{l,0} = \nu_{0,l} = \sum_{\{r:n_r\geq l\}} |C_{0l0r}|.$$

Therefore it suffices to have the estimate

$$\sum_{i+l\geq 2} \nu_{il}\left(\frac{t\|A\|}{1+t\tau\|A\|/n_k[A:B]}\right)^{i+l-1} \leq 2c_k.$$

The polynomial

$$\sum_{i+l\geq 2} \nu_{il}x^{i+l-1}$$

vanishes at 0; if we denote by ε_k the smallest positive real for which it takes the value $2c_k$, we see that $\|P_k(t)\|_A$ is at most equal to 1 provided that

$$t\|A\| \leq \varepsilon_k\left(1+\frac{t\tau\|A\|}{n_k[A:B]}\right).$$

□

4.7. Convergence of RSS in energy norm and general proof of the convergence of the extrapolation of RSS. We first prove the convergence of the residual smoothing scheme:

THEOREM 4.12. *Assume $A \sim B$. Then for $\tau \geq [A:B]/2$, $P(t_n)^n$ converges strongly to e^{-tA} in energy norm, as n tends to $+\infty$, t_n tends to 0 and nt_n tends to t.*

PROOF. We use Theorem IX.3.6 of Kato [**72**] which describes the theory of approximation of continuous semi-groups by discrete semi-groups.

Define indeed

$$A_n = \frac{1}{t_n}\left(\mathbf{1} - P(t_n)\right).$$

Theorem 4.5 implies that

$$U_n^k = \left(\mathbf{1} - t_nA_n\right)^k = P(t_n)^k$$

is norm bounded by 1 for all integers k and n. Therefore, it suffices to find a complex number ζ such that

$$(A_n+\zeta)^{-1} \xrightarrow{s} (A+\zeta)^{-1} \text{ in } \mathcal{H}_A^1$$

to conclude the proof of the theorem.

We choose $\zeta = 1$, and we prove first that for $f \in \mathcal{H}_A^1 = \mathcal{H}_B^1$, it is possible to find a solution $u(t)$ of

$$\left(\mathbf{1} + \frac{\mathbf{1} - P(t)}{t}\right) u(t) = f.$$

By definition,

$$\mathbf{1} + \frac{\mathbf{1} - P(t)}{t} = \mathbf{1} + \beta(t)A,$$

and therefore it suffices to find a solution of

$$u(t) + (t\tau B + A)\, u(t) = f + t\tau B f. \tag{4.34}$$

Thanks to our assumptions on f, A and B, $f + t\tau B f$ belongs to $\mathcal{H}_A^{-1} = \mathcal{H}_B^{-1}$, and Lax-Milgram's Lemma gives a unique solution of (4.34); moreover, this solution belongs to $\mathcal{H}_A^1$.

We may rewrite (4.34) as

$$(\mathbf{1} + t\tau B + A)\,(u(t) - f) = -Af. \tag{4.35}$$

If we multiply scalarly (4.35) by $u(t) - f$, we obtain the estimate

$$|u(t) - f|^2 + |u(t) - f|_A^2 \leq |f|_A\, |u(t) - f|_A$$

which implies immediately

$$|u(t)|_A \leq 2\, |f|_A\,.$$

For any sequence t_n decreasing toward 0, we select a subsequence, still denoted by t_n, such that

$$u(t_n) \rightharpoonup u_0 \text{ weakly in } \mathcal{H}_A^1.$$

Clearly, $t_n \tau B u(t_n)$ tends to zero weakly in $\mathcal{H}_A^{-1}$ and $t_n \tau B f$ tends to zero strongly in $\mathcal{H}_A^{-1}$. Therefore, in the limit, we must have

$$u_0 + Au_0 = f \tag{4.36}$$

since $Au(t_n)$ converges to Au_0 weakly in $\mathcal{H}_A^{-1}$.

Lax-Milgram's Lemma shows that there exists a unique u_0 satisfying (4.36). In order to show the strong convergence of $u(t_n)$ to u_0 in $\mathcal{H}_A^1$, we multiply (4.34) by $u(t_n)^*$, getting thus the identity

$$|u(t_n)|^2 + t_n \tau u(t_n)^* B u(t_n) + u(t_n)^* A u(t_n) = u(t_n)^* f + t_n \tau u(t_n)^* B f. \tag{4.37}$$

On the one hand, we infer from (4.37)

$$\begin{aligned}\overline{\lim_{n\to+\infty}} |u(t_n)|^2 + u(t_n)^* A u(t_n) &\leq \overline{\lim_{n\to+\infty}} u(t_n)^* f + t_n \tau u(t_n)^* B f \\ &\leq u_0^* f = |u_0|^2 + u_0^* A u_0.\end{aligned}$$

On the other hand, general theorems imply

$$\underline{\lim_{n\to+\infty}} |u(t_n)|^2 + u(t_n)^* A u(t_n) \geq |u_0|^2 + u_0^* A u_0.$$

Therefore,

$$\lim_{n\to+\infty} |u(t_n)|^2 + u(t_n)^* A u(t_n) = |u_0|^2 + u_0^* A u_0,$$

proving the desired strong convergence.

Moreover, since the sequence t_n was arbitrary and u_0 is unique we have the stronger result

$$\lim_{t\to 0} |u(t) - u_0|_A = 0.$$

□

We will prove now the convergence of the extrapolation P_k of RSS.

THEOREM 4.13. *If $A \sim B$, for all $k \in \mathbb{N}$, for any choice of integers $1 \le n_1 < n_2 < \cdots < n_k$, there exists $\tau_0 > 0$ such that for all $\tau \ge \tau_0$, $P_k(t_n)^n$ converges strongly to e^{-tA} in energy norm, as n tends to $+\infty$, t_n tends to 0 and nt_n tends to t.*

PROOF. As in Theorem 4.12, we will use Theorem IX.3.6 of [**72**]. Theorem 4.9 yields that $\{\|P_k(t_n)\|_A\}_n$ is bounded uniformly by 1 and we have to prove that for all f in $\mathcal{H}_A^1$,

$$\left(\mathbf{1} + \frac{\mathbf{1} - P_k(t)}{t}\right)^{-1} f \xrightarrow[t\to 0]{s} (\mathbf{1} + A)^{-1} f.$$

We show first that for $f \in \mathcal{H}_A^1$, we can find a solution $u(t)$ of

$$\left(\mathbf{1} + \frac{\mathbf{1} - P_k(t)}{t}\right) u(t) = f. \tag{4.38}$$

Using equation (4.27), we find that

$$\mathbf{1} + \frac{\mathbf{1} - P_k(t)}{t} = \mathbf{1} + \sum_{i=1}^{n_k} (-t)^{i-1} p_i(t)$$

and therefore after applying A to equation (4.38), we can rewrite this equation as

$$Au(t) + Ap_1(t)u(t) + \sum_{i=2}^{n_k} (-t)^{i-1} Ap_i(t)u(t) = Af. \tag{4.39}$$

In order to apply Lax-Milgram's Lemma, let us show first that there exists $\tau_0 > 0$ such that for all $\tau \ge \tau_0$, for all $t > 0$,

$$Ap_1(t) + \sum_{i=2}^{n_k} (-t)^{i-1} Ap_i(t) \succ 0, \tag{4.40}$$

that is to say, using equations (4.26) and multiplying equation (4.40) on the left and on the right by A^{-1}, that, for τ large enough,

$$\sum_{j=1}^{k} \ell_j^k(0)\beta_j \succ -\sum_{i=2}^{n_k} (-t)^{i-1} \sum_{\{j:n_j\ge i\}} \binom{n_j}{i} \frac{\ell_j^k(0)}{n_j^i} \beta_j^{1/2} \left(\beta_j^{1/2} A \beta_j^{1/2}\right)^{i-1} \beta_j^{1/2}.$$

Since, from equation (4.25),

$$\sum_{j=1}^{k} \ell_j^k(0)\beta_j \succ c_k\beta_k,$$

it suffices to show that for τ large enough,

(4.41)

$$c_k\mathbf{1} \succ -\sum_{i=2}^{n_k}(-t)^{i-1}\sum_{\{j:n_j\geq i\}}\binom{n_j}{i}\frac{\ell_j^k(0)}{n_j^i}\beta_k^{-1/2}\beta_j^{1/2}\left(\beta_j^{1/2}A\beta_j^{1/2}\right)^{i-1}\beta_j^{1/2}\beta_k^{-1/2}.$$

Let us write

$$\mu_i = \sum_{\{j:n_j\geq i\}}\binom{n_j}{i}\frac{|\ell_j^k(0)|}{n_j}$$

and

$$Q(x) = \sum_{i=2}^{n_k}\mu_i x^{i-1}.$$

Using estimates (4.29) and (4.30), equation (4.41) and therefore equation (4.40) are satisfied if

$$c_k \geq \sum_{i=2}^{n_k}\mu_i\left(\frac{[A:B]}{\tau}\right)^{i-1} = Q\left(\frac{[A:B]}{\tau}\right),$$

which is true if $\tau \geq [A:B]/\varepsilon_k$, where ε_k is the positive real such that $Q(\varepsilon_k) = c_k$.

Therefore we have proved equation (4.40) and Lax-Milgram's Lemma yields the existence and the uniqueness of a solution $u(t)$ of (4.39), which belongs to $\mathcal{H}_A^1$.

We multiply now equation (4.39) by $u(t)$ and we obtain

(4.42)

$$u(t)^*Au(t) + u(t)^*Ap_1(t)u(t) + \sum_{i=2}^{n_k}(-t)^{i-1}u(t)^*Ap_i(t)u(t) = u(t)^*Af.$$

Using equation (4.40), we find that

$$u(t)^*Au(t) \leq u(t)^*Af$$

and therefore that

$$|u(t)|_A \leq |f|_A. \tag{4.43}$$

Thus, for any subsequence t_n, decreasing toward 0, we extract a subsequence, still denoted by t_n, such that

$$u(t_n) \rightharpoonup u_0 \text{ weakly in } \mathcal{H}_A^1. \tag{4.44}$$

We pass to the limit in equation

$$u(t_n) + p_1(t_n)u(t_n) + \sum_{i=2}^{n_k}(-t_n)^{i-1}p_i(t_n)u(t_n) = f. \tag{4.45}$$

Since

$$p_1(t_n)u(t_n) = \sum_{j=1}^{k} \ell_j^k(0)\beta_j Au(t_n),$$

and $Au(t_n) \rightharpoonup Au_0$ weakly in $\mathcal{H}_A^{-1}$, we deduce from Lemma 4.4 that

$$p_1(t_n)u(t_n) \rightharpoonup \sum_{j=1}^{k} \ell_j^k(0)Au_0 = Au_0 \text{ weakly in } \mathcal{H}_A^{-1}. \tag{4.46}$$

Moreover, since for all j, $1 \leq j \leq k$,

$$\|\beta_j\|_{\mathcal{L}(\mathcal{H}_B^{-1},\mathcal{H}_B^1)} \leq 1, \tag{4.47}$$

there exists $C > 0$ such that

$$|\beta_j Au(t_n)|_A \leq C\,|u(t_n)|_A \tag{4.48}$$

and therefore the term $p_i(t_n)u(t_n)$ is bounded by $C\,|f|_A$ in $\mathcal{H}_A^1$, where C is a positive constant. Thus, the following limit holds true:

$$(-t_n)^{i-1}p_i(t_n)u(t_n) \to 0 \text{ strongly in } \mathcal{H}_A^1. \tag{4.49}$$

Thus, from limits (4.44), (4.46) and (4.49), equation (4.45) yields

$$u_0 + Au_0 = f.$$

To conclude, as in the proof of Theorem 4.12, that $u(t)$ converges strongly to u_0 in $\mathcal{H}_A^1$, we have to prove that

$$|u(t_n)|^2 + u(t_n)^*Au(t_n) \to |u_0|^2 + u_0^*Au_0. \tag{4.50}$$

We deduce from equation (4.45) multiplied on the left by $u(t_n)^*$ that

$$\begin{aligned} |u(t_n)|^2 + u(t_n)^*Au(t_n) &= u(t_n)^*Au(t_n) - u(t_n)^*p_1(t_n)u(t_n) \\ &\quad - \sum_{i=2}^{n_k}(-t_n)^{i-1}u(t_n)^*p_i(t_n)u(t_n) + u(t_n)^*f. \end{aligned} \tag{4.51}$$

The last term on the right hand side of (4.51) converges:

$$u(t_n)^*f \to u_0^*f = |u_0|^2 + u_0^*Au_0. \tag{4.52}$$

Now let us prove that the sum of the other terms of the right hand side of (4.51) converges to zero. For that purpose, we first prove that $u(t_n)^*p_i(t_n)u(t_n)$ is bounded. We remark that, thanks to hypothesis (4.2),

$$|u(t_n)^*p_i(t_n)u(t_n)| \leq |u(t_n)|\,|p_i(t_n)u(t_n)| \leq |u(t_n)|_A\,|p_i(t_n)u(t_n)|_A\,;$$

the $\mathcal{H}_A^1$ norm of $p_i(t_n)u(t_n)$ is bounded by $C\,|f|_A$ from equation (4.48), as explained above, and $|u(t_n)|_A$ is bounded by $|f|_A$ thanks to equation (4.43). Therefore,

$$(-t_n)^{i-1}u(t_n)^*p_i(t_n)u(t_n) \to 0. \tag{4.53}$$

Now let us compute $A - p_1(t)$ in order to factorize it by t. Hence, a simple computation leads to

$$\begin{aligned}
&\mathbf{1} - \sum_{j=1}^{k} \ell_j^k(0)\beta_j \\
&\quad = \prod_{i=1}^{k} \beta_i \left(\prod_{l=1}^{k} \left(\mathbf{1} + \frac{t\tau}{n_l} B \right) - \sum_{j=1}^{k} \ell_j^k(0) \prod_{\{l:1\le l\le k, l\neq j\}} \left(\mathbf{1} + \frac{t\tau}{n_l} B \right) \right) \\
&\quad = \prod_{i=1}^{k} \beta_i \left(\mathbf{1} - \sum_{j=1}^{k} \ell_j^k(0)\mathbf{1} + tS(B) \right)
\end{aligned}$$

where S is a polynomial of degree k, with coefficients depending continuously on t. Therefore, using equation (4.18), we find that

$$A - p_1(t) = t \prod_{i=1}^{k} \beta_i S(B) A.$$

Thus, we can estimate for l, $1 \le l \le k$ the term $u(t_n)^* \prod_i \beta_i B^l A u(t_n)$ as follows:

$$\left| u(t_n)^* \prod_{i=1}^{k} \beta_i B^l A u(t_n) \right| \le \left| B^l \prod_{i=1}^{k} \beta_i u(t_n) \right|_A |u(t_n)|_A \,;$$

using the fact that $|\beta_j B u(t_n)|_A$ is bounded by $C\,|u(t_n)|_A$ by equation (4.47) and using equation (4.43), we conclude that

$$u(t_n)^* \prod_{i=1}^{k} \beta_i S(B) A u(t_n)$$

is bounded and therefore that the term

$$u(t_n)^* A u(t_n) - u(t_n)^* p_1(t_n) u(t_n) = t_n u(t_n)^* \prod_{i=1}^{k} \beta_i S(B) A u(t_n)$$

converges to 0. Using the two other limits (4.52) and (4.53), we can conclude that (4.50) holds true and that $u(t)$ converges strongly to u_0 in $\mathcal{H}_A^1$; the proof of Theorem 4.13 is complete.

□

4.8. Order of the residual smoothing scheme and of its extrapolations. In the following, we will denote by $C(|u|_{\mathcal{H}_A^n})$ a constant depending only on $|u|_{\mathcal{H}_A^n}$ and by $C(|u|_{\mathcal{H}_A^n}, \tau)$ a constant depending on $|u|_{\mathcal{H}_A^n}$ and τ.

Let us prove that the residual smoothing scheme is of order one in time:

THEOREM 4.14. *Suppose that $\mathcal{H}_B^2$ and $\mathcal{H}_A^2$ are isomorphic. There exists $t_0 > 0$ such that for all $u \in \mathcal{H}_A^5$, for all τ larger than $[A : B]/2$ and for all $T > 0$, there exists $C(|u|_{\mathcal{H}_A^5}, T, \tau)$ such that for all $t \in (0, t_0]$ and for all n such that $nt \leq T$,*

$$\left|P(t)^n u - e^{-ntA} u\right|_A \leq C(|u|_{\mathcal{H}_A^5}, T, \tau) t.$$

PROOF. Let $u \in \mathcal{H}_A^5$,

$$\left|e^{-tA} u - u + tAu\right|_A \leq Ct^2 \left|A^2 u\right|_A = Ct^2 \left|A^{5/2} u\right|. \tag{4.54}$$

We also have the following equality:

$$\beta(t) = \mathbf{1} - t\tau\beta(t)B;$$

thus $P(t)u$ can be expressed as follows:

$$P(t)u = u - t\beta(t)Au = u - tAu + t^2\tau\beta(t)BAu. \tag{4.55}$$

Equations (4.54) and (4.55) lead to the following estimate:

$$\left|P(t)u - e^{-tA} u\right|_A \leq Ct^2 \left(\left|A^2 u\right|_A + \tau \left|BAu\right|_A\right) \tag{4.56}$$

and as $\mathcal{H}_B^2$ and $\mathcal{H}_A^2$ are isomorphic,

$$\leq Ct^2\tau \left|A^2 u\right|_A.$$

Let us now consider n iteration steps. Using the triangle inequality, we obtain

$$\left|P(t)^n u - e^{-ntA} u\right|_A \leq \sum_{j=0}^{n-1} \left|P(t)^{n-j-1} \left(P(t) - e^{-tA}\right) e^{-jtA} u\right|_A$$

using Theorem 4.5 and estimate (4.56),

$$\begin{aligned} &\leq \sum_{j=0}^{n-1} C\tau t^2 \left|A^2 e^{-jtA} u\right|_A \\ &\leq C\tau n t^2 |u|_{\mathcal{H}_A^5} \\ &\leq CTt\tau |u|_{\mathcal{H}_A^5} \end{aligned}$$

and the proof is complete. □

Now let us prove that the extrapolation P_k of P is of order k in time.

THEOREM 4.15. *Suppose that $\mathcal{H}_B^{2k+2}$ and $\mathcal{H}_A^{2k+2}$ are isomorphic. There exist $p \in \mathbb{N}$, $\tau_0 > 0$ and $t_0 > 0$ such that for all $u \in \mathcal{H}_A^p$, for all $\tau \geq \tau_0$ and for all $T > 0$, there exists $C(|u|_{\mathcal{H}_A^p}, T, \tau)$ such that for all $t \in (0, t_0]$ and for all n such that $nt \leq T$,*

$$\left|P_k(t)^n u - e^{-ntA} u\right|_A \leq C(|u|_{\mathcal{H}_A^p}, T, \tau) t^k.$$

PROOF. We will use the Theorem 3.1 of [**55**]. A is an operator with bounded inverse and $Y = \cap_{k\in\mathbb{Z}} D(A^k)$ is dense in $\mathcal{H}^0$. As in [**55**], we will denote by $\mathcal{Z}_k$ the set of operators L such that $Y \subset D(L)$, $L : Y \to Y$ and

$$\text{for all } m \in \mathbb{Z},\ |L|_{m,k} = \sup_{u\in Y\setminus\{0\}} \frac{\left|A^{m-k} L u\right|_0}{|A^m u|_0} < \infty.$$

$\mathcal{Z} = \cup_{k\in\mathbb{Z}} \mathcal{Z}_k$ is a subalgebra of the algebra of linear operators from Y to itself. We can remark that for all $k \in \mathbb{N}$, $A^k \in \mathcal{Z}_k \subset \mathcal{Z}$.

A is also the generator of a strongly continuous semigroup $\exp(-tA)$ which satisfies the following estimate:

$$\text{for all } m \in \mathbb{Z}, \left|\exp(-tA) - \sum_{j=0}^{k} \frac{(-tA)^j}{j!}\right|_{m,k+1} = O(t^{k+1}). \tag{4.57}$$

The formula

$$\beta(t) = \sum_{l=0}^{k} (-t\tau)^l B^l + (-t\tau)^{k+1} \beta(t) B^{k+1}$$

enables us to develop $P(t)u$ as follows:

$$P(t)u = u + \sum_{l=1}^{k+1} (-t)^l \tau^{l-1} B^{l-1} A u + (-t)^{k+2} \tau^{k+1} \beta(t) B^{k+1} A u. \tag{4.58}$$

Let us define

$$f_l = (-1)^l \tau^{l-1} B^{l-1} A \text{ and } \varepsilon_{k+1}(t) = (-1)^{k+2} \tau^{k+1} t \beta(t) B^{k+1} A.$$

Then,

$$P(t) = \mathbf{1} + \sum_{l=1}^{k+1} t^l f_l + t^{k+1} \varepsilon_{k+1}(t).$$

We can remark that $f_1 = -A$ as required.

As $\mathcal{H}_B^{2k+2}$ and $\mathcal{H}_A^{2k+2}$ are isomorphic, $f_l \in \mathcal{Z}_l$ and $\varepsilon_{k+1}(t) \in \mathcal{Z}_{k+2}$. We also have

$$\text{for all } m \in \mathbb{Z},\ \lim_{t\to 0} |\varepsilon_{k+1}(t)|_{m,k+2} = 0.$$

Finally, thanks to estimate (4.57) and equation (4.58), we obtain

$$\text{for all } m \in \mathbb{Z}, t^{-2} \left|P(t) - \exp(-tA)\right|_{m,2} = C(|u|_{\mathcal{H}_A^4}).$$

Finally, we can adapt the proof of Theorem 3.1 of [**55**] and we find that there exists i large enough such that, for all $m \in \mathbb{Z}$,

$$t^{-(k+1)} \left| \sum_{j=1}^{k} \ell_j^k(0) P(t/n_j)^{n_j} - \exp(-tA) \right|_{m,i} = O(1).$$

And in particular, for $m = i+1$, we obtain, using equation (4.2), that

$$\left| \left(\sum_{j=1}^{k} \ell_j^k(0) P(t/n_j)^{n_j} - \exp(-tA) \right) u \right|_A \leq C t^{k+1} \left| u \right|_{\mathcal{H}_A^{2i+2}}.$$

For n time steps, the end of the proof is similar to Theorem 4.14. $\square$

5. Application du schéma par régularisation du résidu au préconditionnement d'une méthode spectrale par une méthode d'éléments finis

Application of the Residual Smoothing Scheme to the preconditioning of spectral methods by finite elements methods

Magali Ribot [1]

Abstract : *Preconditioning of spectral methods by finite elements operators has been used since 1984* [**24**], [**54**]*; in this article, we prove the spectral equivalence of the corresponding stiffness matrices for a Gauss-Lobatto-Legendre method in one dimension, and the operator* $-d^2/dx^2$*. We also obtain results relative to combinations of the mass and stiffness matrices of both methods; these results are used for proving order and consistency for time integration by the residual smoothing scheme defined by Averbuch et al.* [**4**] *and studied theoretically in* [**99**].

5.1. Introduction. Spectral methods produce full matrices; therefore, their numerical efficiency depends on the introduction of appropriate preconditioners. In the case of a Laplace — or more generally an elliptic — operator, methods have been proposed for preconditioning: in 1980, Orszag [**87**] suggested preconditioning by finite differences; he gave an argument for the spectral equivalence between the spectral and finite differences stiffness matrices in the case of periodic boundary conditions and a Fourier basis, and stated that this equivalence still holds in many other cases. Haldenwang et al. [**63**] give an argument for spectral equivalence between the stiffness matrices for finite differences and Chebyshev spectral approximation. In [**24**], Canuto and Quarteroni tested a large number of preconditioners for Chebyshev spectral calculations, including preconditioning by finite elements, and gave numerical estimates of the spectral radii of the different numerical methods; in [**53**] Deville and Mund test a variety of finite elements methods for the same problem and in [**54**], they extend their ideas to more general classes of orthogonal polynomials.

Preconditioning by finite elements produces self-adjoint operators, without any extra algorithmic cost relatively to finite differences; moreover more theoretical results are known for self-adjoint preconditioning than for non-self adjoint preconditioning.

Denote by K_S the stiffness matrix associated to a Legendre–Gauss–Lobatto method for $-d^2/dx^2$ with Dirichlet boundary conditions, and

[1] I would like to thank very warmly Michelle Schatzman for pointing me out this subject and for many helpful discussions. Many thanks are due to Seymour Parter and David Gottlieb for their numerous advice and encouragements.

by K_F the stiffness matrix associated to the P_1 finite elements method on the nodes of this spectral method.

Let M_S be the mass matrix of the spectral method, and let M_F be the mass-lumped mass matrix of the P_1 finite elements method constructed on the nodes of the spectral method. We want to solve the one-dimensional heat equation with Dirichlet boundary conditions. Using a spectral method in space and an Euler backward scheme in time, the discretization can be written as

$$M_S \frac{u^{n+1} - u^n}{\Delta t} + K_S u^n = 0,$$

or as

$$\frac{u^{n+1} - u^n}{\Delta t} + M_S^{-1} K_S u^n = 0. \tag{5.1}$$

M_S is a diagonal positive definite matrix, so it is possible to define and to compute easily M_S^{-1}, $M_S^{1/2}$ and $M_S^{-1/2}$. Let $v^n = M_S^{1/2} u^n$, equation (5.1) becomes

$$\frac{v^{n+1} - v^n}{\Delta t} + M_S^{-1/2} K_S M_S^{-1/2} v^n = 0. \tag{5.2}$$

We introduce a preconditioner with the help of the finite element method and the new scheme can be written as follows, where τ is a parameter to be chosen:

$$\frac{v^{n+1} - v^n}{\Delta t} + \tau M_F^{-1/2} K_F M_F^{-1/2} (v^{n+1} - v^n) + M_S^{-1/2} K_S M_S^{-1/2} v^n = 0. \tag{5.3}$$

We define operators A and B defined as follows:

$$A = M_S^{-1/2} K_S M_S^{-1/2} \tag{5.4}$$

and

$$B = M_F^{-1/2} K_F M_F^{-1/2} \tag{5.5}$$

and we recognize the Residual Smoothing Scheme introduced in [**4**] and studied in [**100, 99**] for general operators A and B.

In [**100**], we prove that the Residual Smoothing Scheme and its Richardson's extrapolations are stable for the usual norm for general operators A and B under the assumption that A and B are equivalent. In [**99**], we prove the stability, the convergence and we compute the order of RSS and its extrapolations for the energy norm still under the unique assumption that A and B are equivalent operators and for τ large enough.

Because of the self-adjointness of M_S, M_F, K_S and K_F, A and B are both self adjoint; moreover, they are positive since K_F and K_S are both positive. To satisfy all the hypothesis of [**99**], it suffices to prove now that A and B are equivalent in the sense of quadratic forms.

Recent results of Parter [**90**] give the following bounds:

$$\frac{1}{C} \leq \frac{\mathrm{Re}\,(K_F M_S M_F^{-1} U, U)}{(K_S U, U)} \leq \frac{|(K_F M_S M_F^{-1} U, U)|}{(K_S U, U)} \leq C.$$

Here $(\,,\,)$ denotes the hermitian scalar product. These results are based on [**89**], which itself builds on Gatteschi's results from [**61**]. When M_F is not mass-lumped, Parter proves an analogous result in [**91**]. These results of Parter are quite analogous to ours, but are not exactly the same.

The main result (Proposition 5.5) of the present article is the spectral equivalence between

$$M_S^{1/2} M_F^{-1/2} K_F M_F^{-1/2} M_S^{1/2}$$

and K_S uniformly with respect to the number N of discretization nodes. Since M_F and M_S are both diagonal and therefore commute, this implies that A, defined at equation (5.4), and B, defined at equation (5.5), are equivalent.

As a consequence of a theorem of Parter and Rothman [**92**], which proves that the stiffness matrices K_F and K_S are equivalent, it suffices to prove the spectral equivalence between K_F and

$$M_S^{1/2} M_F^{-1/2} K_F M_F^{-1/2} M_S^{1/2}.$$

It turns out that when we started working on this question, we were not aware of Parter's results, and we did not consult the recent literature on orthogonal polynomials; in [**98**], instead of using a Sturm method or a descent method, as is done by most authors in this field, we took the classical integral representation formula for ultra-spherical polynomials (4.10.3) from Szegő [**118**], and we applied to this formula a stationary phase strategy, in a region where the classical expansions cannot be applied; this method gives an expansion at all orders, with estimates for the error bound. Let us point out that this is not a classical stationary phase method, since the exponential term is a non linear function of the large parameter and of the integration variable.

Though our result on preconditioning can be obtained with Parter's method, we feel that our treatment of the asymptotics is novel.

Let us describe briefly how we use formulas of Szegő [**118**] and of [**98**] to prove our main result (Proposition 5.5). Denote by ξ_k, $k \in \{1, \cdots, N-1\}$ the nodes of the spectral method; they are the zeroes of the derivative of the Legendre polynomial L_N. The diagonal elements of $M_F^{-1} M_S$ are called σ_k and they can be given in terms of $L_N(\xi_k)^2$ and of the ξ_j's. We prove that the elements σ_k are bounded independently of k and N.

Define the discrete H^1 norm by

$$\|U\|_{\mathbb{H}^1_N} = \left(\sum_{k=0}^{N-1} \frac{|U_{k+1} - U_k|^2}{\xi_{k+1} - \xi_k} \right)^{1/2};$$

our result is equivalent to the existence of a constant $C > 0$ such that

$$\|U\|_{\mathbb{H}^1_N}/C \leq \|M_F^{-1/2} M_S^{1/2} U\|_{\mathbb{H}^1_N} \leq C\|U\|_{\mathbb{H}^1_N}.$$

We define

$$\mu_k = \frac{2 - |\xi_k| - |\xi_{k+1}|}{\xi_{k+1} - \xi_k} \left| \frac{1}{\sigma_{k+1}} - \frac{1}{\sigma_k} \right|^2$$

and as explained in section 5.4.1, we are reduced to estimate

$$\Sigma_N = \sum_{k=0}^{N-1} \mu_k. \tag{5.6}$$

Proposition 5.5 states that this quantity is bounded independently of N.

Denote by $\lfloor r \rfloor$ the largest integer at most equal to the real r. By symmetry, it suffices to deal with $k \leq \lfloor (N-1)/2 \rfloor = N'$. We partition the interval $\{0, \cdots, N'\}$ into three subintervals: $\{0, \cdots, K\}$, $\{K+1, \cdots, \lfloor \Lambda N \rfloor\}$ and $\{\lfloor \Lambda N \rfloor + 1, \cdots, N'\}$ where K is bounded and will be chosen later, and Λ belongs to the open interval $(0, 1/2)$; to estimate μ_k, we need some precise asymptotics of the nodes ξ_k and of the polynomial L_N and we obtain these expansions by different ways on each subinterval.

First, in the region $1 \leq k \leq K$, which is the easiest to handle, it suffices to find the limit of μ_k for N tending to infinity. We use asymptotics for the Legendre polynomials and their derivatives available in Szegő [**118**].

Then, in the region $\lfloor \Lambda N \rfloor \leq k \leq N'$, we use estimates of [**98**]; the asymptotics for the zeroes ξ_k of L'_N are computed from limited expansion of L_N and its derivatives, given by Szegő's book [**118**] thanks to a quantitative implicit function theorem. Let us recall the asymptotics of the polynomials $L_N^{(k)}$ of Szegő [**118**] and the expansions of the zeroes ξ_k computed in [**98**].

Denote by $P_N^{(\lambda)}$ the ultra-spherical polynomial of degree N, i.e. the orthogonal polynomial of degree N related to the weight $(1-x^2)^{\lambda - 1/2}$. The Legendre polynomial L_N is equal to $P_N^{(1/2)}$ and its derivatives are therefore given by formula (4.7.14) of Szegő [**118**], that is to say

$$\frac{d}{dx} P_N^{(\lambda)}(x) = 2\lambda P_{N-1}^{(\lambda+1)}(x). \tag{5.7}$$

We set

$$\omega_{N,\lambda} = \binom{N + \lambda - 1}{N} = \frac{\Gamma(N+\lambda)}{\Gamma(N+1)\Gamma(\lambda)}; \tag{5.8}$$

formula (8.21.14) of Szegő [**118**] is the following:

$$
\begin{aligned}
P_N^{(\lambda)}(\cos\theta) = \frac{2\omega_{N,\lambda}}{(2\sin\theta)^\lambda}\sum_{\nu=0}^{p-1}\omega_{\nu,\lambda}\frac{(1-\lambda)\cdots(\nu-\lambda)}{(N+\lambda-1)\cdots(N+\lambda-\nu)} \\
\times\frac{\cos\big((N-\nu+\lambda)\theta-(\nu+\lambda)\pi/2\big)}{(2\sin\theta)^\nu}+O(N^{\lambda-p-1});
\end{aligned}
\tag{5.9}
$$

this formula is uniform with respect to θ in $[\Lambda/2,\pi/2]$ and to N.

In [**98**], we find, using an implicit function theorem, the following lemma:

Lemma 5.1. *Define*

$$\theta_{0,k}=\frac{\pi/4+k\pi}{N+3/2}.$$

Then for all $\Lambda\in(0,1/2)$, there exist C, C' such that for all $N\geq 2$ and for all integer k in $\{\lfloor\Lambda N\rfloor,\cdots,\lceil(1-\Lambda)N\rceil\}$, there exists a unique zero θ_k of $P_N^{(3/2)}(\cos\theta)$ in a ball of radius C'/N^2 about $\theta_{0,k}$; moreover the following estimate holds

$$\left|\theta_k-\theta_{0,k}+\frac{3}{8N^2\tan\theta_{0,k}}-\frac{9}{8N^3\tan\theta_{0,k}}\right|\leq CN^{-4}.$$

We can infer from this lemma an expansion in terms of k and N of the zero η_k of $\theta\mapsto L_N'(\cos\theta)$ which lies in a neighborhood of size $O(N^{-2})$ about

$$\eta_{0,k}=\pi-\frac{\pi/4+k\pi}{N+1/2}=\frac{(N-k)\pi+\pi/4}{N+1/2}. \tag{5.10}$$

At last, in the region $\{K,\cdots,\lfloor\Lambda N\rfloor\}$, we use asymptotics of $L_N^{(k)}$ and ξ_k both computed in [**98**]. The general form of the asymptotics of $P_N^{(\lambda)}$ are given in Theorem 3.15 of [**98**] and we find some explicit values for $\lambda=1/2,3/2,5/2$ and $7/2$ in Corollary 3.16. Then, we compute the expansion of the discretization nodes ξ_k, still using an implicit functions theorem and we obtain the following lemma:

Lemma 5.2. *Define*

$$\theta_{0,k}=\frac{\pi(N-k+1/4)}{N+1/2}.$$

Then for all $K>0$ and for all $\Lambda\in(0,1/2)$, there exist C, C' such that for all $N\geq 2$ and for all integer k in $\{K,\cdots,\lfloor\Lambda N\rfloor\}$, there exists a unique zero θ_k of $L_N'(\cos\theta)$ in a ball of radius C'/N^2 about $\theta_{0,k}$; moreover the following estimate holds

$$\left|\theta_k-\theta_{0,k}-\frac{13}{8N^2\tan\theta_{0,k}}+\frac{49}{12N^3\tan\theta_{0,k}}\right|\leq C\left((N^{-1}+K^{-1})^4\right). \tag{5.11}$$

All these results enable us to prove that the quantity (5.6) is bounded independently of N and therefore that $M_F^{-1/2}M_S^{1/2}K_FM_S^{1/2}M_F^{-1/2}$ and K_F are equivalent.

Once this is done, we generalize all these results to the two–dimensional case.

The article is organized as follows: in section 5.2, we give all our notations, definitions ; in section 5.3, we show the spectral equivalence of the mass matrices in L^∞ norm. Section 5.4 contains the proof of the main estimate (5.6) which leads to the spectral equivalence of the mass matrices in discrete H^1 norm. Then, results are generalized to the two–dimensional case in section 5.5. Eventually, we give numerical results in section 5.6

5.2. Notations, definitions. Let us denote by $\mathbb{P}_N$ the space of polynomial functions of degree N defined over $[-1,1]$ and by $\mathbb{P}_N^0$ the subspace of functions p belonging to $\mathbb{P}_N$ and verifying $p(-1)=p(1)=0$. We also denote by $C^0([-1,1])$ the space of continuous functions over $[-1,1]$ and $C_0^0([-1,1])$ the subspace of the functions f of $C^0([-1,1])$ vanishing at -1 and 1.

Let us denote by L_N the Legendre polynomial of degree N and let $-1=\xi_0<\xi_1<\cdots<\xi_{N-1}<\xi_N=1$ be the roots of $(1-x^2)L_N'$; let ρ_k, $0\le k\le N$ be the weights of the quadrature formula associated to the nodes ξ_k; since this is a Gauss-Lobatto formula, we shall have

$$\forall\Phi\in\mathbb{P}_{2N-1},\ \int_{-1}^{1}\Phi(x)dx=\sum_{k=0}^{N}\Phi(\xi_k)\rho_k; \tag{5.12}$$

the weights ρ_k are strictly positive.

Bernardi and Maday [**9**] give explicit expressions of the ρ_k's:

$$\begin{aligned}\rho_0=\rho_N&=\frac{2}{N(N+1)},\\ \rho_k&=\frac{2}{N(N+1)L_N^2(\xi_k)},\ 1\le k\le N-1.\end{aligned} \tag{5.13}$$

For k in $\{0,\cdots,N\}$, we denote by l_k the elements of the Lagrange basis built on the nodes ξ_k. They form a basis of $\mathbb{P}_N^0$.

Let u and v belong to $\mathbb{P}_N^0$; their scalar product associated to the collocation method on the nodes ξ_k is

$$(u,v)_N=\sum_{k=0}^{N}u(\xi_k)v(\xi_k)\rho_k.$$

Then the mass matrix M_S is given by

$$(M_S)_{i,j}=\delta_{i,j}\rho_j,\ 1\le i,j\le N-1$$

and the stiffness matrix K_S is defined by

$$(K_S)_{i,j} = \sum_{k=0}^{N} \rho_k l_i'(\xi_k) l_j'(\xi_k). \tag{5.14}$$

Let $\mathbf{1}_{[a,b]}$ be the characteristic function of $[a,b]$. The space V_N of P_1 finite elements is an (N-1)-dimensional subspace of $C_0^0([-1,1])$ spanned by the hat functions ϕ_k, for k in $\{1,\cdots,N-1\}$,

$$\phi_k(x) = \frac{x-\xi_{k-1}}{\xi_k-\xi_{k-1}}\mathbf{1}_{[\xi_{k-1},\xi_k]}(x) + \frac{\xi_{k+1}-x}{\xi_{k+1}-\xi_k}\mathbf{1}_{[\xi_k,\xi_{k+1}]}(x). \tag{5.15}$$

For algorithmic reasons, we choose to lump the mass in the finite elements method; then the mass matrix M_F is the $(N-1)\times(N-1)$ diagonal matrix with diagonal coefficients

$$(\xi_2-\xi_0)/2,\cdots,(\xi_{i+1}-\xi_{i-1})/2,\cdots,(\xi_N-\xi_{N-2})/2.$$

The stiffness matrix is a $(N-1)\times(N-1)$ tridiagonal matrix K_F given by

$$(K_F)_{i,j} = \int_{-1}^{1} \phi_i'(x)\phi_j'(x)dx. \tag{5.16}$$

The non vanishing coefficients are explicitly given by

$$(K_F)_{i,j} = \begin{cases} \dfrac{1}{\xi_i-\xi_{i-1}} + \dfrac{1}{\xi_{i+1}-\xi_i}, & \text{if } i=j, \\ \dfrac{1}{\xi_{i-1}-\xi_i}, & \text{if } i=j+1, \\ \dfrac{1}{\xi_i-\xi_{i+1}}, & \text{if } i=j-1. \end{cases} \tag{5.17}$$

The operator α_N^F maps a vector r of $\mathbb{R}^{N-1}$ to the piecewise affine function which interpolates the values r_j at the nodes ξ_j, i.e.:

$$\alpha_N^F(r) = \sum_{k=1}^{N-1} r_k\phi_k. \tag{5.18}$$

Let $U=(U_1,\cdots,U_{N-1})$ be an element of $\mathbb{R}^{N-1}$; whenever needed we set $U_0=U_N=0$. Let us denote by $|.|_1$ the H_0^1 semi-norm on $[-11]$. The two following identities are obvious:

$$\left|\alpha_N^F U\right|_1^2 = U^* K_F U \text{ and} \tag{5.19}$$

$$U^* K_F U = \sum_{k=0}^{N-1} \frac{(U_{k+1}-U_k)^2}{\xi_{k+1}-\xi_k}. \tag{5.20}$$

We define η_k by

$$\eta_k = \mathrm{Arccos}(\xi_k). \tag{5.21}$$

Since we have

$$-1 = \xi_0 < \xi_1 < \cdots < \xi_{N-1} < \xi_N = 1,$$

we infer that

$$0 = \eta_N < \eta_{N-1} < \cdots < \eta_1 < \eta_0 = \pi.$$

The matrices M_S and M_F are diagonal; we define the diagonal elements of $M_F^{-1} M_S$ as:

$$\sigma_k = \frac{2\rho_k}{\xi_{k+1} - \xi_{k-1}}, \quad \text{for } 1 \le k \le N-1. \tag{5.22}$$

From the beginning of the article up to Lemma 5.6, we set $\sigma_0 = \sigma_N = 0$; after Lemma 5.7, we set $1/\sigma_0 = 1/\sigma_N = 0$.

Remark 5.3. *Since L_N is even (resp. odd) when N is even (resp. odd), we see that*

$$\xi_{N-k} = -\xi_k, \quad \textit{for } 1 \le k \le N-1 \tag{5.23}$$

and, using equation (5.21),

$$\eta_{N-k} = \pi - \eta_k, \quad \textit{for } 1 \le k \le N-1. \tag{5.24}$$

Thus, using formulas (5.13) *and* (5.22),

$$\rho_{N-k} = \rho_k, \quad \textit{for } 1 \le k \le N-1 \tag{5.25}$$

and

$$\sigma_{N-k} = \sigma_k, \quad \textit{for } 1 \le k \le N-1. \tag{5.26}$$

We will need estimates on ξ_k.

Szegő's book [**118**] contains a large number of estimates and asymptotics of the zeroes of L_N; in particular, we deduce from his theorems 6.21.2 and 6.21.3 the following estimates on η_k for $k \in \{1, \cdots, N-1\}$,

$$\pi\left(1 - \frac{2k+2}{2N+1}\right) \le \eta_k \le \pi\left(1 - \frac{2k-1}{2N+1}\right) \tag{5.27}$$

and

$$\pi\left(1 - \frac{k+1/2}{N}\right) \le \eta_k \le \pi\left(1 - \frac{k-1}{N}\right). \tag{5.28}$$

The use of either (5.27) or (5.28) is a question of algebraic convenience.

For the reader's convenience, it is advisable to consult the fourth edition of Szegő's book [**118**], which is the most complete.

5.3. Equivalence between M_F and M_S in L^∞ operator norm. We want to show that the diagonal matrices $M_F^{-1}M_S$ are bounded from above and from below independently of N for the L^∞ operator norm, and therefore it is sufficient to prove that the coefficients σ_k of the diagonal of $M_F^{-1}M_S$ are bounded.

THEOREM 5.4. *There exists $\tau > 0$ such that for all $N \geq 2$, for all $k \in \{1 \cdots N-1\}$,*

$$\tau^{-1} \leq \sigma_k \leq \tau. \tag{5.29}$$

PROOF. The symmetry relation (5.23) shows that it suffices to prove the assertion of the lemma for $1 \leq k \leq N/2$. Lemma 1.14 of Chapter III of [**9**] gives the following estimate of ρ_k as a function of ξ_k:

$$cN^{-1}(1-\xi_k^2)^{1/2} \leq \rho_k \leq c'N^{-1}(1-\xi_k^2)^{1/2}, \text{ for } k \text{ in } \{1, \cdots, N-1\}. \tag{5.30}$$

Therefore, it is equivalent to estimate σ_k or $\sqrt{1-\xi_k^2}/\big(N(\xi_{k+1}-\xi_{k-1})\big)$. We use the numbers η_k defined in (5.21), and we introduce the numbers $\eta_k^\pm$:

$$\eta_k^+ = \frac{\eta_{k-1}+\eta_{k+1}}{2} \text{ and } \eta_k^- = \frac{\eta_{k-1}-\eta_{k+1}}{2};$$

it is plain that $\sqrt{1-\xi_k^2} = \sin\eta_k$; elementary trigonometry gives $\xi_{k+1} - \xi_{k-1} = 2\sin\eta_k^+ \sin\eta_k^-$. Therefore, it suffices to show that the expression

$$\sin\eta_k / N(\sin\eta_k^+ \sin\eta_k^-)$$

is bounded from above and from below.

Estimate (5.28) implies

$$\pi\frac{k-1}{N} \leq \pi - \eta_k \leq \pi\frac{k+1/2}{N}, \tag{5.31a}$$

$$\pi\frac{k-1}{N} \leq \pi - \eta_k^+ \leq \pi\frac{k+1/2}{N}, \tag{5.31b}$$

$$\frac{\pi}{4N} \leq \eta_k^- \leq \frac{7\pi}{4N}. \tag{5.31c}$$

In order to apply the inequality $2x/\pi \leq \sin x \leq x$ to (5.31), we have to check that the upper bounds belong to $[0, \pi/2]$, the lower bounds being obviously non negative; if $N \geq 4$, the upper bound in (5.31c) is less than or equal to $\pi/2$; we infer then the inequalities

$$\frac{1}{2} \leq N\sin\eta_k^- \leq \frac{7\pi}{4}; \tag{5.32}$$

if $k \leq \lfloor (N-1)/2 \rfloor$, we have $k+1/2 \leq N/2$, and therefore, the upper bound in (5.31a) and (5.31b) is at most equal to $\pi/2$, whence

$$\frac{2(k-1)}{N} \leq \min(\sin\eta_k, \sin\eta_k^+) \leq \max(\sin\eta_k, \sin\eta_k^+) \leq \frac{\pi(k+1/2)}{N}. \tag{5.33}$$

Then, a straightforward computation gives for $N \geq 4$ and $2 \leq k \leq \lfloor (N-1)/2 \rfloor$:

$$\frac{8}{7\pi^2}\frac{k-1}{k+1/2} \leq \frac{\sin \eta_k}{N \sin \eta_k^+ \sin \eta_k^-} \leq \frac{\pi(k+1/2)}{k-1}. \tag{5.34}$$

If $k = 1$, the lower bound in (5.33) vanishes; therefore, instead of (5.28), we must use (5.27), whence

$$\frac{\pi}{2N+1} \leq \pi - \eta_1 \leq \frac{4\pi}{2N+1}; \tag{5.35}$$

we also use (5.27) and $\eta_0 = \pi$ to establish

$$\frac{3\pi}{2(2N+1)} \leq \pi - \eta_1^+ \leq \frac{3\pi}{2N+1}. \tag{5.36}$$

We infer from (5.35), (5.36) and (5.32) the inequality

$$\frac{8}{21\pi^2} \leq \frac{1}{N}\frac{\sin \eta_1}{\sin \eta_1^- \sin \eta_1^+} \leq \frac{8\pi}{3}. \tag{5.37}$$

If N is even and $k = N/2$, we replace the upper bound in (5.33) by 1, and we obtain immediately the inequalities

$$\frac{4}{7\pi}\left(1 - \frac{2}{N}\right) \leq \frac{\sin \eta_{N/2}}{N \sin \eta_{N/2}^- \sin \eta_{N/2}^+} \leq 2\left(1 - \frac{2}{N}\right)^{-1};$$

finally the cases $N = 2, 3, 4$ contribute a finite number of strictly positive answers, and the lemma is proved. □

5.4. Equivalence between $M_F^{1/2}$ and $M_S^{1/2}$ in discrete H^1 norm. In this section, we prove the equivalence between K_F and

$$M_F^{-1/2} M_S^{1/2} K_F M_S^{1/2} M_F^{-1/2},$$

which is equivalent to the following result:

Proposition 5.5. *The matrix $M_F^{-1/2} M_S^{1/2}$ and its inverse are bounded uniformly in the discrete H^1 norm, i.e. there exists a constant C such that for all $N \geq 2$, and for all vector $U \in \mathbb{R}^{N-1}$ the following inequalities hold:*

$$C^{-1} \leq \frac{\|M_F^{-1/2} M_S^{1/2} U\|_{\mathbb{H}_N^1}}{\|U\|_{\mathbb{H}_N^1}} \leq C.$$

5.4.1. *Reduction to a simpler problem.* Let us show that proving Proposition 5.5 is equivalent to proving that (5.6) is bounded.

We introduce now the discrete H^1 norm: the space $\mathbb{R}^{N-1}$ is equipped with the norm

$$\|U\|_{\mathbb{H}_N^1} = \left|\alpha_N^F U\right|_1$$

and using equations (5.19) and (5.20), we have the following equality

$$\|U\|_{\mathbb{H}_N^1} = \left(\sum_{k=0}^{N-1} \frac{|U_{k+1} - U_k|^2}{\xi_{k+1} - \xi_k}\right)^{1/2}, \tag{5.38}$$

where U_0 and U_N are set equal to zero throughout this section. We will denote by $\mathbb{H}_N^1$ the space $\mathbb{R}^{N-1}$ equipped with norm (5.38).

To obtain the consistency error [**99**]– [**100**] of the residual smoothing scheme, we need to prove that the mass matrices $M_F^{1/2}$ and $M_S^{1/2}$ are equivalent as operators from $\mathbb{H}_N^1$ to $\mathbb{H}_N^1$. Indeed, the remainder of this section is devoted to proving that as an operator from $\mathbb{H}_N^1$ to itself, $M_F^{-1/2} M_S^{1/2}$ and its inverse are bounded independently of N.

The square of the norm of $M_F^{-1/2} M_S^{1/2}$ is given by the expression

$$\max\left\{\sum_{k=0}^{N-1} \frac{\left|\sqrt{\sigma_{k+1}} U_{k+1} - \sqrt{\sigma_k} U_k\right|^2}{\xi_{k+1} - \xi_k} : \|U\|_{\mathbb{H}_N^1} = 1\right\}. \tag{5.39}$$

We will transform this unpalatable expression into something more tractable. The first step comes from the identity

$$\begin{aligned}\sqrt{\sigma_{k+1}} U_{k+1} - \sqrt{\sigma_k} U_k = &\left(\frac{\sqrt{\sigma_{k+1}} + \sqrt{\sigma_k}}{2}\right)(U_{k+1} - U_k) \\ &+ \left(\frac{\sqrt{\sigma_{k+1}} - \sqrt{\sigma_k}}{2}\right)(U_{k+1} + U_k)\,;\end{aligned}$$

we infer from the inequality $(a+b)^2 \le 2a^2 + 2b^2$ that

$$\begin{aligned}\sum_{k=0}^{N-1} \frac{\left|\sqrt{\sigma_{k+1}} U_{k+1} - \sqrt{\sigma_k} U_k\right|^2}{\xi_{k+1} - \xi_k} \le 2 &\sum_{k=0}^{N-1} \left(\frac{\sqrt{\sigma_{k+1}} + \sqrt{\sigma_k}}{2}\right)^2 \frac{|U_{k+1} - U_k|^2}{\xi_{k+1} - \xi_k} \\ &+ \sum_{k=0}^{N-1} \frac{\left(\sqrt{\sigma_{k+1}} - \sqrt{\sigma_k}\right)^2}{2} \frac{|U_{k+1} + U_k|^2}{\xi_{k+1} - \xi_k}.\end{aligned} \tag{5.40}$$

The first sum on the right hand side of (5.40) can be estimated by

$$2 \max_{0 \le k \le N} \sigma_k \|U\|_{\mathbb{H}_N^1}^2,$$

and since, according to Theorem 5.4, σ_k is bounded independently of N and k, there remains to estimate the second sum in the right hand side of (5.40).

We will produce now an algebraically simpler upper bound of this sum. The first step consists in observing that the discrete Hölder constant of $U \in \mathbb{H}_N^1$ can be estimated by $\|U\|_{\mathbb{H}_N^1}$; more precisely, we prove:

LEMMA 5.6. *Let U belong to $\mathbb{H}_N^1$. For all i and j in $\{0, \cdots, N\}$, the following estimate holds*

$$|U_j - U_i| \leq \|U\|_{\mathbb{H}_N^1} |\xi_j - \xi_i|^{1/2}. \tag{5.41}$$

PROOF. Let us assume that $j > i$. We use the triangle inequality to write

$$|U_j - U_i| \leq \sum_{k=i}^{j-1} |U_{k+1} - U_k| \tag{5.42}$$

We decompose $|U_{k+1} - U_k|$ as $\dfrac{|U_{k+1} - U_k|}{\sqrt{\xi_{k+1} - \xi_k}} \sqrt{\xi_{k+1} - \xi_k}$ and then use Cauchy - Schwarz inequality; equation (5.42) becomes:

$$|U_j - U_i| \leq \left(\sum_{k=i}^{j-1} \frac{|U_{k+1} - U_k|^2}{\xi_{k+1} - \xi_k} \right)^{1/2} \left(\sum_{k=i}^{j-1} (\xi_{k+1} - \xi_k) \right)^{1/2}$$

and using equation (5.38)

$$|U_j - U_i| \leq \|U\|_{\mathbb{H}_N^1} |\xi_j - \xi_i|^{1/2},$$

which concludes the proof. □

If $1 \leq k \leq N/2$, we use (5.41) with $j = k$ and $i = 0$; we obtain, since $U_0 = 0$ and $\xi_0 = -1$ and since ξ_k is non positive,

$$|U_k| \leq \|U\|_{\mathbb{H}_N^1} |\xi_k + 1|^{1/2} = \|U\|_{\mathbb{H}_N^1} (1 - |\xi_k|)^{1/2}. \tag{5.43}$$

If $N/2 \leq k \leq N$, we use (5.41) with $j = k$ and $i = N$ and we obtain, by the same way, equation (5.43).

Therefore, we infer from equation (5.43) the estimate:

$$|U_k + U_{k+1}|^2 \leq 2 \left(2 - |\xi_k| - |\xi_{k+1}|\right) \|U\|_{\mathbb{H}_N^1}^2 .$$

Next lemma shows that $\left|\sqrt{\sigma_{k+1}} - \sqrt{\sigma_k}\right|$, $\left|1/\sqrt{\sigma_{k+1}} - 1/\sqrt{\sigma_k}\right|$ and $|1/\sigma_{k+1} - 1/\sigma_k|$ are equivalent quantities.

LEMMA 5.7. *Let τ be as in Theorem 5.4. We have the following inequalities for $N \geq 2$ and for k in $\{1, \cdots, N-2\}$:*

(5.44)

$$\tau^{-2} |\sqrt{\sigma_k} - \sqrt{\sigma_{k+1}}|^2 \leq \left| \frac{1}{\sqrt{\sigma_{k+1}}} - \frac{1}{\sqrt{\sigma_k}} \right|^2 \leq \tau^2 |\sqrt{\sigma_k} - \sqrt{\sigma_{k+1}}|^2 ,$$

$$4\tau^{-5} |\sqrt{\sigma_k} - \sqrt{\sigma_{k+1}}|^2 \leq \left| \frac{1}{\sigma_{k+1}} - \frac{1}{\sigma_k} \right|^2 \leq 4\tau^5 |\sqrt{\sigma_k} - \sqrt{\sigma_{k+1}}|^2 \tag{5.45}$$

and

$$\tau^{-2}\sigma_1 \le \frac{1}{\sigma_1} \le \tau^2\sigma_1, \quad \tau^{-2}\sigma_{N-1} \le \frac{1}{\sigma_{N-1}} \le \tau^2\sigma_{N-1},$$
$$4\tau^{-5}\sigma_1 \le \frac{1}{\sigma_1^2} \le 4\tau^5\sigma_1, \quad 4\tau^{-5}\sigma_{N-1} \le \frac{1}{\sigma_{N-1}^2} \le 4\tau^5\sigma_{N-1}.$$

Proof. We observe that

$$\left|\frac{1}{\sqrt{\sigma_{k+1}}} - \frac{1}{\sqrt{\sigma_k}}\right|^2 = \frac{\left|\sqrt{\sigma_k} - \sqrt{\sigma_{k+1}}\right|^2}{\sigma_k\sigma_{k+1}},$$

and, using Theorem 5.4, inequalities (5.44) hold.

We also have

$$\left|\frac{1}{\sigma_{k+1}} - \frac{1}{\sigma_k}\right|^2 = |\sqrt{\sigma_k} - \sqrt{\sigma_{k+1}}|^2 \left|\frac{\sqrt{\sigma_k} + \sqrt{\sigma_{k+1}}}{\sigma_k\sigma_{k+1}}\right|^2$$

which implies (5.45) with the help of Theorem 5.4. □

Therefore, in order to estimate $M_F^{-1/2}M_S^{1/2}$ in $\mathbb{H}_N^1$ operator norm, it suffices to estimate

$$\Sigma_N = \sum_{k=0}^{N-1} \frac{(2 - |\xi_{k+1}| - |\xi_k|)}{\xi_{k+1} - \xi_k} \left|\frac{1}{\sigma_{k+1}} - \frac{1}{\sigma_k}\right|^2. \tag{5.46}$$

We make henceforth the convention that $1/\sigma_0 = 0$ and $1/\sigma_N = 0$, and we will need for future reference

$$\mu_k = \frac{(2 - |\xi_{k+1}| - |\xi_k|)}{\xi_{k+1} - \xi_k} \left|\frac{1}{\sigma_{k+1}} - \frac{1}{\sigma_k}\right|^2. \tag{5.47}$$

Observe that of $M_F^{1/2}M_S^{-1/2}$ can be estimated along the same route: instead of (5.39), we consider

$$\max\left\{\sum_{k=0}^{N-1} \frac{\left|U_{k+1}/\sqrt{\sigma_{k+1}} - U_k/\sqrt{\sigma_k}\right|^2}{\xi_{k+1} - \xi_k} : \|U\|_{\mathbb{H}_N^1} = 1\right\}.$$

We argue as for (5.40), and as σ_k is bounded away from 0 uniformly in k and N, it suffices to estimate

$$\sum_{k=0}^{N-1} \left(\frac{1}{\sqrt{\sigma_{k+1}}} - \frac{1}{\sqrt{\sigma_k}}\right)^2 \frac{|U_{k+1} + U_k|^2}{\xi_{k+1} - \xi_k}.$$

In virtue of Lemma 5.7, $M_F^{1/2}M_S^{-1/2}$ will be bounded in $\mathbb{H}_N^1$ operator norm if (5.46) is bounded independently of N.

The estimate of Σ_N will depend on very precise asymptotics of ξ_k and L_N given in [**98**].

We deduce from the symmetries (5.23) and (5.26):

$$\mu_{N-k} = \mu_{k-1}, \quad 1 \le k \le N.$$

Define $N' = \left\lfloor \frac{N-1}{2} \right\rfloor$; it suffices to estimate

$$\Sigma'_N = \sum_{k=0}^{N'} \mu_k$$

since $\Sigma_N \leq 2\Sigma'_N$.

We partition the interval $\{0, \cdots, N'\}$ into three subintervals: $\{0, \cdots, K\}$, $\{K+1, \cdots, \lfloor \Lambda N \rfloor\}$ and $\{\lfloor \Lambda N \rfloor + 1, \cdots, N'\}$ where K is bounded and will be chosen later, and Λ belongs to the open interval $(0, 1/2)$. We will study separately these three regions, one region by subsection.

5.4.2. *The region* $0 \leq k \leq K$. First, we prove that the sum of the μ_k's is bounded when k is in a bounded region around 0. For that purpose, we use some asymptotics of Szegő's book [**118**].

THEOREM 5.8. *For all $K > 0$, there exists $C > 0$ such that for all $N \geq 2$,*

$$\sum_{0 \leq k \leq K} \mu_k \leq C, \tag{5.48}$$

where μ_k is defined in equation (5.47).

PROOF. Since K is finite, it suffices to find a bound of each μ_k, $0 \leq k \leq K$ as $N \to +\infty$.

First, when $k = 0$, μ_0 has a simpler expression since $\xi_1 - \xi_0$ is equal to $1 - |\xi_1|$ which is itself equal to $2 - |\xi_1| - |\xi_0|$. Therefore, $\mu_0 = 1/\sigma_1^2$, and we know from equation (5.29) that μ_0 is bounded.

Let z_k be the k-th positive zero of the Bessel function J_1; Theorem 8.1.2 of Szegő [**118**] says that, for $k \geq 1$,

$$\pi - \eta_k = \frac{z_k}{N} + o(1/N), \text{ as } N \text{ tends to } \infty. \tag{5.49}$$

Therefore, $\xi_k = -1 + 2\sin^2(z_k/2N) + o(1/N)$, so that

$$2 - |\xi_k| - |\xi_{k+1}| = 2\big(\sin^2(z_k/2N) + \sin^2(z_{k+1}/2N) + o(1/N)\big),$$

and in a similar fashion

$$\xi_{k+1} - \xi_k = 2\big(\sin^2(z_{k+1}/2N) - \sin^2(z_k/2N) + o(1/N)\big);$$

therefore, we find the limit of the ratio of these two quantities as N tends to infinity:

$$\lim_{N\to\infty} \frac{2 - |\xi_k| - |\xi_{k+1}|}{\xi_{k+1} - \xi_k} = \frac{z_{k+1}^2 + z_k^2}{z_{k+1}^2 - z_k^2}.$$

According to (5.13) and (5.22), the limit of $1/\sigma_k$ as N tends to infinity is given by

$$\lim_{N\to\infty} \frac{N(N+1)(\xi_{k+1} - \xi_k)}{4} L_N^2(\xi_k) = \frac{z_{k+1}^2 - z_{k-1}^2}{8} \lim_{N\to\infty} L_N^2(\xi_k).$$

At this point, we use Theorem 8.1.1 of [**118**], and we conclude that

$$\lim_{N\to+\infty}\frac{1}{\sigma_k}=\frac{(z_{k+1}^2-z_{k-1}^2)}{8}J_0^2(z_k).$$

It is then clear that μ_k has a limit given by

$$\lim_{N\to+\infty}\mu_k=\frac{z_{k+1}^2+z_k^2}{64(z_{k+1}^2-z_k^2)}\left|(z_{k+1}^2-z_{k-1}^2)J_0^2(z_k)-(z_{k+2}^2-z_k^2)J_0^2(z_{k+1})\right|^2$$

and thus μ_k, $1\le k\le K$, is bounded independently of N, which proves the Theorem. □

5.4.3. *The region* $\lfloor\Lambda N\rfloor\le k\le N'$. Now, we estimate the sum of the μ_k's when k is close to its upper bound N', i.e. in the region $\lfloor\Lambda N\rfloor\le k\le N'$. We use some asymptotics proved in [**98**] using an implicit functions theorem and expansions with uniform remainders of Szegő [**118**].

Lemma 5.9. *For all* $\Lambda\in(0,1/2)$, *there exists* $C>0$ *such that for all* $N\ge 2$,

$$\sum_{\lfloor\Lambda N\rfloor\le k\le N'}\mu_k\le CN^{-4}. \tag{5.50}$$

Proof. Since, from formula (4.7.14) of Szegő [**118**], L_N' is equal to $P_{N-1}^{(3/2)}$, we pass from the root θ_k of $P_N^{(3/2)}$, expanded in Theorem 2.1 of [**98**], to the root η_k of $P_{N-1}^{(3/2)}$ by decreasing N by 1. According to (5.10), we also have to replace k by $N-k$; therefore

$$\eta_{0,k}=\frac{(N-k+1/4)\pi}{N+1/2}$$

and we have the asymptotic

$$\eta_k=\eta_{0,k}-\frac{3}{8N^2\tan\eta_{0,k}}\left(1-\frac{1}{N}\right)+O(N^{-4}), \tag{5.51}$$

the error term being uniform in N and in $k\in\{\lfloor\Lambda N\rfloor,\cdots,N'\}$. The asymptotic expansion of $L_N=P_N^{(1/2)}$ given by formula (8.21.14) of Szegő [**118**] is

$$\begin{aligned}
P_N^{(1/2)}(\cos\theta)&=\frac{2\omega_{N,1/2}}{\sqrt{2\sin\theta}}(T_1+T_2+T_3)+O(N^{-7/2}),\\
T_1&=\cos\big((N+1/2)\theta-\pi/4\big)\\
T_2&=\frac{1}{2(2N-1)}\frac{\cos\big((N-1/2)\theta-3\pi/4\big)}{2\sin\theta},\\
T_3&=\frac{9}{8(2N-1)(2N-3)}\frac{\cos\big((N-3/2)\theta-5\pi/4\big)}{(2\sin\theta)^2}.
\end{aligned}$$

In this subsection, we write for simplicity

$$t=\tan(\eta_{0,k}).$$

We infer from the asymptotic (5.51) the following asymptotics for each of the terms T_1, T_2 and T_3 when $\theta = \eta_k$:

$$T_1 = (-1)^{N-k}\left(1 - \frac{9}{128N^2t^2}\right) + O(N^{-3}),$$
$$T_2 = (-1)^{N-k-1}\frac{1}{8N}\left(1 + \frac{1}{2N} + \frac{3}{8Nt^2}\right) + O(N^{-3}),$$
$$T_3 = (-1)^{N-k-1}\frac{9}{128N^2}\left(\frac{1}{t^2} - 1\right) + O(N^{-3}).$$

Therefore, the sum $T_1 + T_2 + T_3$ is

(5.52)
$$T_1 + T_2 + T_3 = (-1)^{N-k}\left(1 - \frac{1}{8N} + \frac{1}{128N^2} - \frac{3}{16N^2t^2}\right) + O(N^{-3}).$$

We need an expansion for $1/\sqrt{\sin\eta_k}$: from the Taylor expansion

$$\sin\eta_k = \sin\eta_{0,k}\left(1 - \frac{3}{8N^2t^2} + O(N^{-3})\right), \tag{5.53}$$

we infer

$$\frac{1}{\sqrt{\sin\eta_k}} = \frac{1}{\sqrt{\sin\eta_{0,k}}}\left(1 + \frac{3}{16N^2t^2} + O(N^{-3})\right). \tag{5.54}$$

We get also an expansion of $\omega_{N,1/2}$ with the help of Stirling's formula:

$$\omega_{N,1/2} = \frac{1}{\sqrt{\pi N}}\left(1 - \frac{1}{8N} + \frac{1}{128N^2} + O(N^{-3})\right). \tag{5.55}$$

We perform the product of (5.52), (5.54) and (5.55), and we find

(5.56)
$$P_N^{(1/2)}(\cos\eta_k) = (-1)^{N-k}\frac{\sqrt{2}}{\sqrt{\pi N\sin\eta_{0,k}}}\left(1 - \frac{1}{4N} + \frac{1}{32N^2} + O(N^{-3})\right).$$

Observe that the error term in (5.56) is uniform in N and in $k \in \{\lfloor \Lambda N\rfloor, \cdots, N'\}$. In order to calculate σ_k, we need an asymptotic of $\xi_{k+1} - \xi_{k-1}$: we write a Taylor expansion of $\xi_{k\pm1} = \cos\eta_{k\pm1}$ at η_k, and we obtain

$$\begin{aligned}\xi_{k+1} - \xi_{k-1} = \\ &\sin\eta_k(\eta_{k-1} - \eta_{k+1}) - \frac{1}{2}\big((\eta_{k+1} - \eta_k)^2 - (\eta_{k-1} - \eta_k)^2\big)\cos\eta_k \\ &+ \frac{1}{6}\big((\eta_{k+1} - \eta_k)^3 - (\eta_{k-1} - \eta_k)^3\big)\sin\eta_k + O(N^{-4}).\end{aligned}$$

Another Taylor expansion gives $\eta_{k\pm1} - \eta_k$:

$$\eta_{k\pm1} - \eta_k = \big(\eta_{0,k\pm1} - \eta_{0,k}\big)\left(1 + \frac{3(1 + 1/t^2)}{8N^2}\right) + O(N^{-4}).$$

Therefore, we obtain with the help of (5.53):

$$\frac{\xi_{k+1}-\xi_{k-1}}{2} = \sin\eta_{0,k}\frac{\pi}{N+1/2}\left(1+\frac{3}{8N^2}-\frac{\pi^2}{6N^2}\right)+O(N^{-4}). \tag{5.57}$$

We put together (5.56) and (5.57) and we obtain the expansion of $1/\sigma_k$, given by

$$\frac{1}{\sigma_k} = 1-\frac{\pi^2}{6N^2}+O(1/N^3). \tag{5.58}$$

We can now estimate the sum

$$\sum_{k=\lfloor\Lambda N\rfloor}^{N'} \frac{(2-|\xi_{k+1}|-|\xi_k|)}{\xi_{k+1}-\xi_k}\left|\frac{1}{\sigma_{k+1}}-\frac{1}{\sigma_k}\right|^2. \tag{5.59}$$

The difference $1/\sigma_{k+1}-1/\sigma_k$ is an $O(N^{-3})$; the term $2-|\xi_k|-|\xi_{k+1}|$ is an $O(1)$ and relation (5.57) shows that $\xi_{k+1}-\xi_k$ is greater than or equal to C/N; since there are $O(N)$ terms, we see that the sum is an $O(N^{-4})$. □

5.4.4. *The region $K \le k \le \lfloor\Lambda N\rfloor$.* Finally, let us prove that the sum of the μ_k's is bounded in the third region thanks to asymptotics of [**98**], which are the most difficult to handle. We proved them using integral representations of Legendre polynomials and a stationary phase method.

Let us show that the sum of the μ_k's for k in $\{K,\cdots,\lfloor\Lambda N\rfloor\}$ is bounded:

LEMMA 5.10. *For $K>0$ large enough and for all $\Lambda\in(0,1/2)$, there exists $C>0$ such that for all $N\ge 2$,*

$$\sum_{K\le k\le\lfloor\Lambda N\rfloor}\mu_k \le C. \tag{5.60}$$

PROOF. We use equation (5.11) of Lemma 5.2.

In this proof, we also write for simplicity $t=\tan\eta_{0,k}$.

Now, we calculate an asymptotic formula for $L_N(\xi_k)=L_N(\cos\eta_k)$; for this purpose, we use the following formula of [**98**]:

$$\begin{aligned}
&P_N^{(1/2)}(\cos(z/N)) = \\
&\quad\sqrt{\frac{2}{\pi}}\frac{1}{\zeta_N^{1/2}}\Bigg[\cos\left(z+\frac{z}{2N}+\frac{3\pi}{4}\right)\left(1-\frac{3}{8N}+\frac{185}{128N^2}\right)\\
&\quad+\frac{1}{\zeta_N}\sin\left(z+\frac{3z}{2N}+\frac{3\pi}{4}\right)\left(\frac{1}{8}-\frac{55}{64N}\right)\\
&\quad-\frac{43}{384}\frac{1}{\zeta_N^2}\cos\left(z+\frac{5z}{2N}+\frac{3\pi}{4}\right)\Bigg]\\
&\quad+O\left((N^{-1}+z^{-1})^3\right),
\end{aligned}$$

where $\zeta_N = N\sin(z/N)$.

We find that

$$L_N(\xi_k) = (-1)^{N+k+1}\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{N\sin\eta_{0,k}}}\left(1-\frac{1}{4N}+\frac{67}{96N^2}-\frac{49}{24N^2t^2}\right) + O\left((N^{-1}+K^{-1})^3\right)$$

and we use equation (5.13) to compute $1/\rho_k$:

$$\frac{1}{\rho_k} = \frac{N}{\pi\sin\eta_{0,k}}\left(1+\frac{1}{2N}+\frac{23}{24N^2}-\frac{49}{12N^2t^2}\right) + O\left((N^{-1}+K^{-1})^2\right).$$

Now, to calculate $1/\sigma_k$, we compute $\xi_{k+1}-\xi_{k-1}$:

$$\xi_{k+1}-\xi_{k-1} = 2\sin\eta_{0,k}\frac{\pi}{N}\left(1-\frac{1}{2N}-\frac{11}{8N^2}-\frac{\pi^2}{6N^2}\right) + O\left((N^{-1}+K^{-1})^4\right)$$

and we obtain

(5.61)

$$\frac{1}{\sigma_k} = \frac{\xi_{k+1}-\xi_{k-1}}{2\rho_k} = 1-\frac{2}{3N^2}-\frac{\pi^2}{6N^2}-\frac{49}{12N^2t^2} + O\left((N^{-1}+K^{-1})^3\right);$$

therefore, there exists $C>0$ such that

$$\left|\frac{1}{\sigma_{k+1}}-\frac{1}{\sigma_k}\right|^2 = \frac{C}{N^6t^2}\left(1+\frac{1}{t^2}\right)^2\left(1+O(N^{-1}+K^{-1})\right).$$

In order to calculate μ_k, we need to compute

$$\frac{(2-|\xi_{k+1}|-|\xi_k|)}{\xi_{k+1}-\xi_k} = \frac{2N}{\pi}\frac{1}{\tan(\eta_{0,k}/2)}\left(1+O(N^{-1}+K^{-1})\right).$$

At last, we compute the value of μ_k, which is

$$\mu_k = \frac{C}{N^5}\frac{\cos^2\eta_{0,k}}{\sin^6\eta_{0,k}}\frac{1}{\tan(\eta_{0,k}/2)}\left(1+O(N^{-1}+K^{-1})\right) \le \frac{C}{k^5}$$

and the sum (5.60) is bounded. □

Remark 5.11. *Observe that* (5.61) *is compatible with* (5.58)*, because the error term in* (5.61) *is large with respect to the error term in* (5.58)*.*

We may now conclude our argument

End of the proof of Proposition 5.5. We complete the proof by choosing Λ in the interval $(0,1/2)$; then the sum of the μ_k's for $\lfloor\Lambda N\rfloor \le k \le N'$ is bounded (Lemma 5.9); for K large enough, the sum of the μ_k's for $K \le k \le \lfloor\Lambda N\rfloor$ is bounded (Lemma 5.10), and finally, the sum of the μ_k's for $k \in \{0,\cdots,K\}$ is bounded (Theorem 5.8). □

5.5. Extension to the two–dimensional case. In this section, we use the previous one-dimensional results to prove some analogous two-dimensional results on the square $[-1,1]\times[-1,1]$.

5.5.1. *Definition of the analogous two-dimensional matrices.* Let us consider the Gauss-Lobatto-Legendre discretization grid, i.e.

$$\Xi_N = \{(\xi_m, \xi_n), 0 \leq m, n \leq N\}.$$

We therefore have a grid made of rectangles and we consider all the meshes obtained by cutting the rectangles into two triangles along the diagonal as is shown in Figure 4. In particular, we consider among all these meshes the two regular meshes of Figure 5.

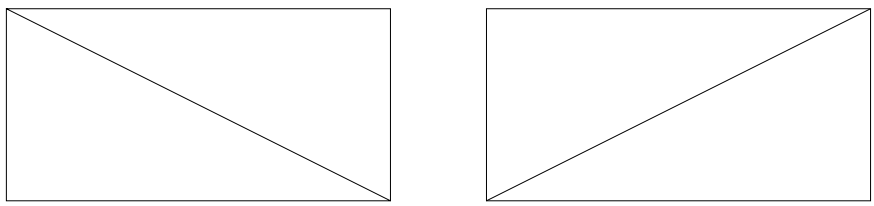


FIGURE 4. Two ways of cutting a rectangle into two triangles

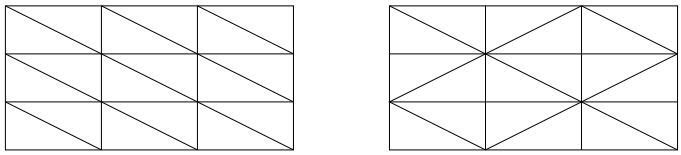


FIGURE 5. The two regular meshes

We notice that the nodes of these meshes are still those of the grid Ξ_N.

The two-dimensional Lagrange basis is the basis of tensorial products $L_{(i,j)} = l_i \otimes l_j$, where the tensorial product $\varphi \otimes \psi$ of two functions φ and ψ defined over $\mathbb{R}$ is the function f defined over $\mathbb{R}^2$ by $f(x, y) = \varphi(x)\psi(y)$. The corresponding scalar product is

$$(u, v)_N = \sum_{0 \leq j,k \leq N} u(\xi_j, \xi_k) v(\xi_j, \xi_k) \rho_j \rho_k.$$

Then the two-dimensional spectral mass matrix, denoted by $M_{2,S}$, is equal to

$$M_{2,S} = M_{1,S} \otimes M_{1,S} \tag{5.62}$$

where $M_{1,S}$ is the one-dimensional spectral mass matrix and the stiffness matrix $K_{2,S}$ is equal to

$$K_{2,S} = M_{1,S} \otimes K_{1,S} + K_{1,S} \otimes M_{1,S}, \tag{5.63}$$

where $K_{1,S}$ is the one-dimensional spectral stiffness matrix.

We denote, for example, by $M_{2,S}(i, j, k, l)$ the coefficient of matrix $M_{2,S}$ corresponding to the scalar product of two basis functions $L_{(i,j)}$ and $L_{(k,l)}$. Thus, $M_{2,S}(i, j, i, j)$ is a diagonal term of the matrix $M_{2,S}$.

We also denote by $M_{1,S}(i,j)$ the (i,j) coefficient of the one-dimensional mass matrix $M_{1,S}$.

Let us compute the finite elements matrices.

In the case of P^1 finite elements, the basis function related to the node (ξ_i, ξ_j) is the function $\phi_{i,j}$ affine on each triangle which is equal to 1 on (ξ_i, ξ_j) and to 0 on the other nodes.

The mass-lumped matrix $M_{2,F}$ is a diagonal matrix, but it is not exactly equal to $M_{1,F} \otimes M_{1,F}$.

Hence, the (i,j,i,j) coefficient of matrix $M_{2,F}$ satisfies the following inequalities:

$$\frac{2}{3}(M_{1,F} \otimes M_{1,F})(i,j,i,j) \leq M_{2,F}(i,j,i,j) \leq \frac{4}{3}(M_{1,F} \otimes M_{1,F})(i,j,i,j).$$

For simplicity reasons, we will therefore take $M_{2,F}$ to be exactly equal to

$$M_{2,F} = M_{1,F} \otimes M_{1,F}. \tag{5.64}$$

Eventually, the stiffness matrix $K_{2,F}$ is independent of the mesh and is equal to

$$K_{2,F} = M_{1,F} \otimes K_{1,F} + K_{1,F} \otimes M_{1,F}, \tag{5.65}$$

where $M_{1,F}$ is still the mass-lumped matrix.

Therefore, we have expressed the two-dimensional mass and stiffness matrices in terms of the one-dimensional matrices; let us now use the theorems of previous sections to show analogous results.

5.5.2. *Equivalence between the stiffness matrices $K_{2,F}$ and $K_{2,S}$.* Parter and Rothman prove in [**92**] the following one-dimensional result:

THEOREM 5.12. *There exists $\kappa > 0$ such that, for all $N \geq 2$ and for all U in $\mathbb{R}^{N-1}$, the following inequality holds:*

$$\kappa^{-1} U^* K_{1,F} U \leq U^* K_{1,S} U \leq \kappa U^* K_{1,F} U.$$

The analogous two-dimensional result is:

THEOREM 5.13. *Let τ defined in Theorem 5.4 and κ in Theorem 5.12. For all $N \geq 2$ and for all U in $\mathbb{R}^{(N-1)\times(N-1)}$, we have the following inequalities:*

$$(\tau\kappa)^{-1} U^* K_{2,F} U \leq U^* K_{2,S} U \leq \tau\kappa\, U^* K_{2,F} U. \tag{5.66}$$

PROOF. Let U belong to $\mathbb{R}^{(N-1)\times(N-1)}$; let us denote

$$U = (U_1 \cdots U_{N-1})$$

where U_i is in $\mathbb{R}^{N-1}$. We denote by $U(i,j)$ the j-th component of the vector U_i and we also define the vector $\widetilde{U}_i$ such that its jth component $\widetilde{U}(i,j)$ is equal to $U(j,i)$.

Thus, $U^* K_{2,S} U$ is equal to the sum of two terms:

$$U^* K_{2,S} U = U^*(M_{1,S} \otimes K_{1,S} + K_{1,S} \otimes M_{1,S}) U.$$

Let us consider the first term on the right-hand side. We will deduce a similar result for the second term.

We use the tensorial product of two matrices, which is defined with our notations by

$$(A \otimes B)(i,j,k,l) = A(i,k)B(j,l),$$

and we obtain

$$\begin{aligned} U^*(M_{1,S} \otimes K_{1,S})U &= \sum_{m,n,p,q} U(m,n)M_{1,S}(m,p)K_{1,S}(n,q)U(p,q) \\ &= \sum_{m,n,q} \rho_m U(m,n)K_{1,S}(n,q)U(m,q) \\ &= \sum_m \rho_m U_m^* K_{1,S} U_m, \end{aligned}$$

where ρ_k is defined at equation (5.13). We then use Theorem 5.4 which proves the equivalence of one-dimensional mass matrices and we obtain:

$$U^*(M_{1,S} \otimes K_{1,S})U \leq \tau \sum_m M_{1,F}(m,m)U_m^* K_{1,S} U_m;$$

we then use Theorem 5.12 on the equivalence of one-dimensional stiffness matrices and we find

$$\begin{aligned} U^*(M_{1,S} \otimes K_{1,S})U &\leq \tau\kappa \sum_m M_{1,F}(m,m)U_m^* K_{1,F} U_m \\ &= \tau\kappa\, U^*(M_{1,F} \otimes K_{1,F})U. \end{aligned}$$

We can prove by the same way, using vectors $\widetilde{U}_i$, that

$$U^*(K_{1,S} \otimes M_{1,S})U \leq \tau\kappa\, U^*(K_{1,F} \otimes M_{1,F})U$$

and therefore that

$$U^* K_{2,S} U \leq \tau\kappa\, U^* K_{2,F} U.$$

We show the first inequality of equation (5.66) in the same fashion and this completes the proof of Theorem 5.13. □

5.5.3. *Equivalence between $M_{2,F}$ and $M_{2,S}$ in L^∞ operator norm.* The matrix $M_{2,F}^{-1}M_{2,S}$ is a diagonal matrix whose coefficients are denoted by $\sigma_2(i,j)$, $1 \leq i,j \leq N-1$. We have the following theorem:

THEOREM 5.14. *Let τ defined at Theorem 5.4. For all $N \geq 2$, for all $i,j \in \{1 \cdots N-1\}$, the following inequalities hold:*

$$\tau^{-2} \leq \sigma_2(i,j) \leq \tau^2. \tag{5.67}$$

PROOF. We use equations (5.62) and (5.64), to obtain

$$\sigma_2(i,j) = \sigma_i \sigma_j.$$

Thanks to Theorem 5.4, we find equation (5.67). □

5.5.4. *Equivalence between $M_{2,F}^{1/2}$ and $M_{2,S}^{1/2}$ in discrete H^1 norm.* At last, let us generalize Proposition 5.5 of section 5.4.

THEOREM 5.15. *Let τ and C be the constants defined at Theorem 5.4 and at Proposition 5.5. For all $N \geq 2$ and for all vector $U \in \mathbb{R}^{(N-1)\times(N-1)}$, the following inequalities hold:*

$$(C\tau)^{-1} \leq \frac{\|M_{2,F}^{-1/2} M_{2,S}^{1/2} U\|_{\mathbb{H}_N^1}}{\|U\|_{\mathbb{H}_N^1}} \leq C\tau. \tag{5.68}$$

PROOF. Let us write $\|M_{2,F}^{-1/2} M_{2,S}^{1/2} U\|_{\mathbb{H}_N^1}$ as

$$U^* M_{2,S}^{1/2} M_{2,F}^{-1/2} K_{2,F} M_{2,F}^{-1/2} M_{2,S}^{1/2} U.$$

We first compute $U^* M_{2,S}^{1/2} M_{2,F}^{-1/2} (M_{1,F} \otimes K_{1,F}) M_{2,F}^{-1/2} M_{2,S}^{1/2} U$, as in the proof of Theorem 5.13 and we obtain

(5.69)

$$\begin{aligned}
&U^* M_{2,S}^{1/2} M_{2,F}^{-1/2} (M_{1,F} \otimes K_{1,F}) M_{2,F}^{-1/2} M_{2,S}^{1/2} U \\
&= \sum_{m,n,p,q} \sqrt{\sigma_2(m,n)\sigma_2(p,q)} U(m,n) M_{1,F}(m,p) K_{1,F}(n,q) U(p,q) \\
&= \sum_{m,n,q} \sigma_m \sqrt{\sigma_n \sigma_q} U(m,n) M_{1,F}(m,m) K_{1,F}(n,q) U(m,q).
\end{aligned}$$

We then use inequality (5.64) to bound the right hand side of equation (5.69) and we obtain

$$\begin{aligned}
&U^* M_{2,S}^{1/2} M_{2,F}^{-1/2} (M_{1,F} \otimes K_{1,F}) M_{2,F}^{-1/2} M_{2,S}^{1/2} U \\
&\leq \frac{3}{2} \sum_{m,n,q} \sigma_m \sqrt{\sigma_n \sigma_q} U(m,n) M_{1,F}(m,m) K_{1,F}(n,q) U(m,q) \\
&\leq \frac{3}{2} \sum_{m} \sigma_m M_{1,F}(m,m) U_m^* M_{1,S}^{1/2} M_{1,F}^{-1/2} K_{1,F} M_{1,F}^{-1/2} M_{1,S}^{1/2} U_m.
\end{aligned}$$

We use Theorem 5.4 and Proposition 5.5 and we find

$$\begin{aligned}
U^* M_{2,S}^{1/2} M_{2,F}^{-1/2} (M_{1,F} \otimes K_{1,F}) M_{2,F}^{-1/2} M_{2,S}^{1/2} U \\
\leq \tau C \sum_m M_{1,F}(m,m) U_m^* K_{1,F} U_m \\
\leq \quad \tau C U^* (M_{1,F} \otimes K_{1,F}) U.
\end{aligned}$$

We show in the same fashion that

$$U^* M_{2,S}^{1/2} M_{2,F}^{-1/2} (K_{1,F} \otimes M_{1,F}) M_{2,F}^{-1/2} M_{2,S}^{1/2} U \leq \tau C U^* (K_{1,F} \otimes M_{1,F}) U$$

and therefore that

$$U^* M_{2,S}^{1/2} M_{2,F}^{-1/2} K_{2,F} M_{2,F}^{-1/2} M_{2,S}^{1/2} U \leq \tau C U^* K_{2,F} U.$$

By the same way, we show the lower bound of (5.68). □

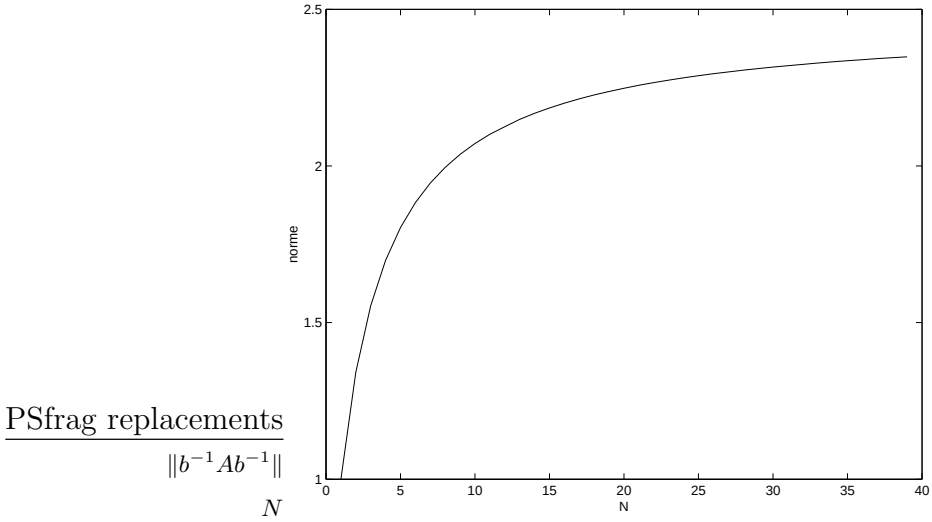


FIGURE 6. The norm $\|b^{-1}Ab^{-1}\|$ as a function of N with A defined by (5.4) and $b = \sqrt{B}$ with B defined by (5.5)

5.6. Numerical simulations. At last, let us present some numerical simulations illustrating the results proved in this article and in [**99**].

First, we check numerically that A and B are spectrally equivalent, as can be seen in Figure 6, which displays the norm $\|b^{-1}Ab^{-1}\|$ in terms of $N \in \{2, \cdots, 40\}$.

A consequence of Lemmas 3.1 and 3.2 of [**99**] is that if $B^{-n/2}A^{n/2}$ is bounded, then for all $s \in \{-n, \cdots, n\}$, $B^{-s/2}A^{s/2}$ is bounded. Figure 7 displays the norm of $B^{-n/2}A^{n/2}$ for $n = j/2$, $j = 1, \cdots, 5$.

These results enable us to prove some general theorems of stability, convergence and order for the Residual Smoothing Scheme and its Richardson's extrapolations.

Now, let us test numerically some results of stability. The strong stability of the second extrapolation P_2 is checked experimentally in Figure 8 for two different values of τ and $N = 20$.For $\tau = 1$, the P_2 scheme is conditionally stable as can be seen for large values of t; it is unconditionally stable for $\tau = 2$.

Then, we can also check the convergence of these schemes. The convergence of RSS and its two first extrapolations P_2 and P_3 is seen experimentally in Figure 9 for $\tau = 2$ and $N = 20$. The order is exactly one for RSS, nearly two for P_2, but we can notice there is a problem with P_3. These extrapolations seem not to be adapted in case of high order.

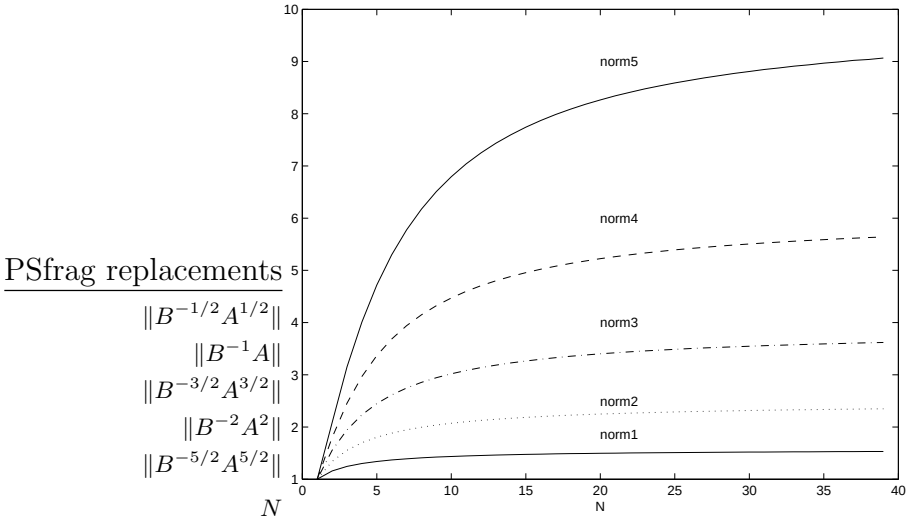


FIGURE 7. The norms of $B^{-n/2}A^{n/2}$ for $n = j/2$, $j = 1, \cdots, 5$

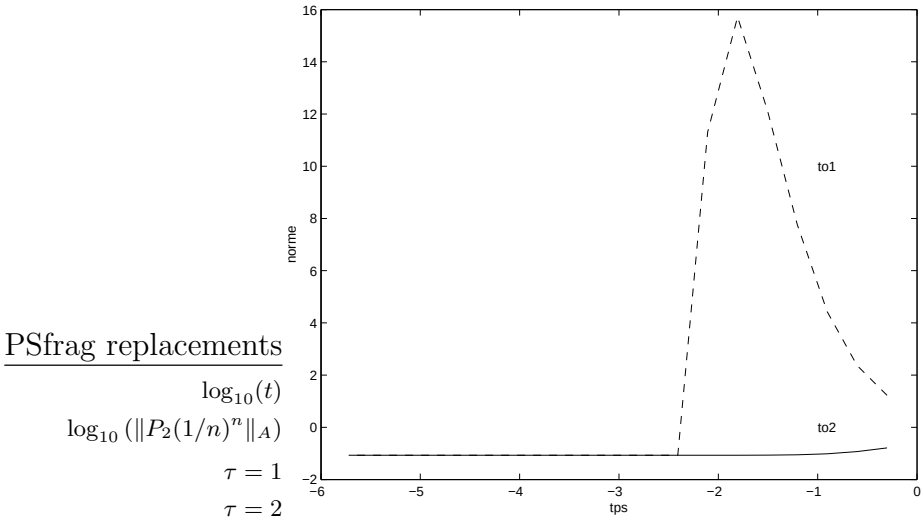


FIGURE 8. Plot of $\|P_2(1/n)^n\|_A$ for $n = 2^j$, $j = 1, \cdots, 19$. The dashed line is the case $\tau = 1$ and the solid line the case $\tau = 2$; for $t \leq 2^{-8}$, the two curves are superposed.

We consider the heat equation

$$\begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = 0,\ x \in]-1, 1[, \\ u(x, 0) = \cos(\pi x/2),\ x \in]-1, 1[. \end{cases} \tag{5.70}$$

with Dirichlet boundary limits.

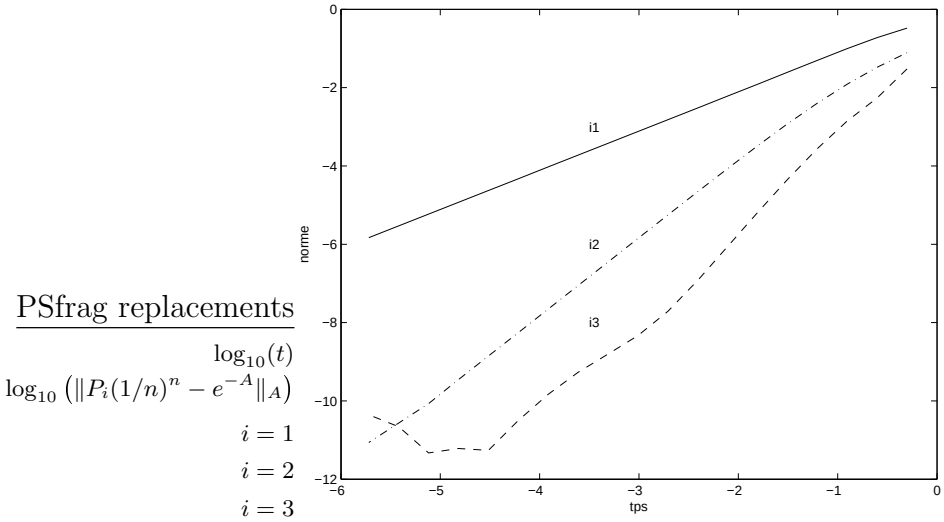


FIGURE 9. Plot of $\|P_i(1/n)^n - e^{-A}\|_A$ for $i = 1, 2, 3$ as a function of $1/n$. The slope of the regression line is about 0.9974 for $i = 1$, about 1.9066 for $i = 2$ and about 2.3329 for $i = 3$, without some of the values for t small.

The exact solution of this equation is

$$u(x,t) = e^{-\pi^2 t/4}\cos(\pi x/2),\ x \in]-1,1[,$$

and we compare it with the solution computed by the residual smoothing scheme. Let us denote the approximation of $u(\xi_j, mt)$ by U_j^m; we consider the following discrete norm of the error:

$$\sqrt{\sum_j |u(\xi_j, mt) - U_j^m|^2 \rho_j}.$$

For $\tau = 2$, the Figure 10 shows the error as a function of t and of $1/N$. We can check here the order one in time of RSS.

Now, we compute the solution of the problem (5.70) with the second extrapolation P_2 of the residual smoothing scheme and with $\tau = 2$. The result can be seen in Figure 11. We can check here the order two of the second Richardson extrapolation.

At last, we still compute the solution of the problem (5.70) with $\tau = 2$ but using the third extrapolation P_3 of the residual smoothing scheme. The error as a function of t and of $1/N$ is plotted in Figure 12. We can check here the order three of the third Richardson extrapolation for N large enough; we notice some problems when the number N of points is too small.

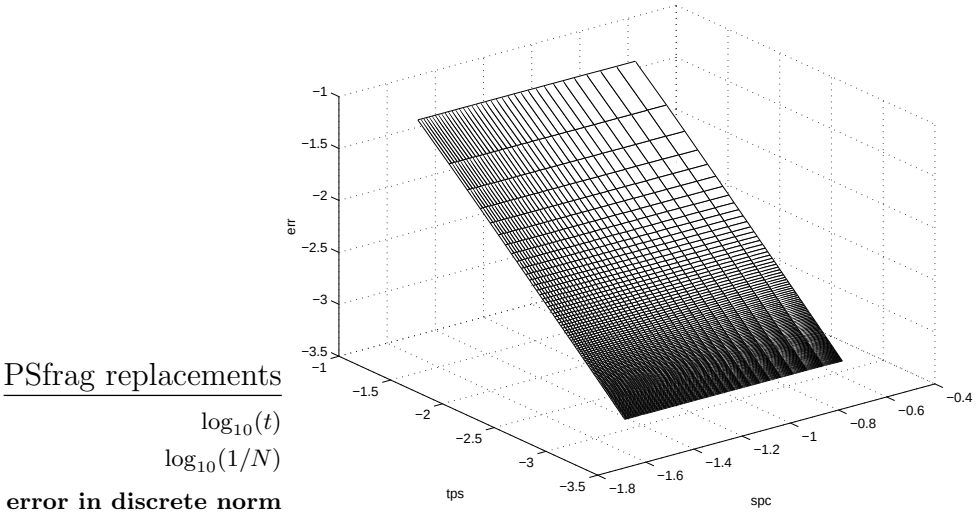


FIGURE 10. Plot of the error as a function of t and $1/N$ in case of problem (5.70); the computation is made with the residual smoothing scheme P. The slope of the logarithm of the error as a function of $\log_{10}(t)$ is nearly 0.9946.

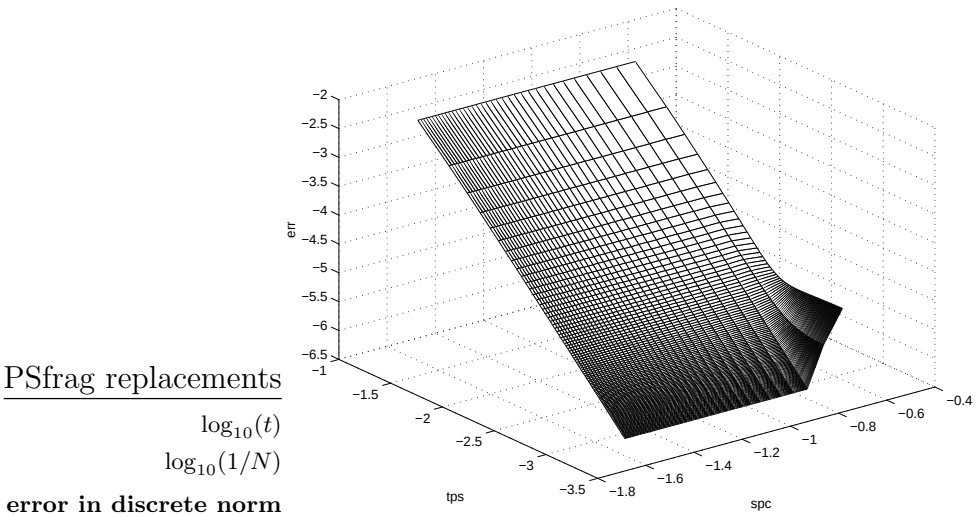


FIGURE 11. Plot of the error as a function of t and $1/N$ in case of problem (5.70); the computation is made with the second extrapolation of the residual smoothing scheme P_2. The slope of the logarithm of the error as a function of $\log_{10}(t)$ is nearly 1.9679.

Now let us check that the space precision obtained by the spectral method has not been deteriorated by the use of the preconditioner.

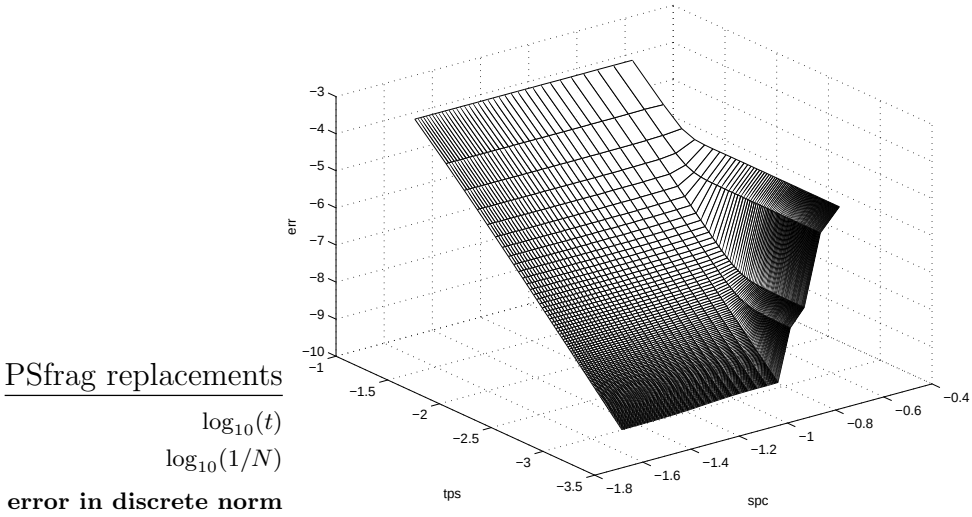


FIGURE 12. Plot of the error as a function of t and $1/N$ in case of problem (5.70); the computation is made with the third extrapolation of the residual smoothing scheme P_3. The slope of the logarithm of the error as a function of $\log_{10}(t)$ is nearly 2.9527.

We consider the equation

$$\begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = 1,\ x \in]-1,1[, \\ u(x,0) = -x^2/2 + 1/2,\ x \in]-1,1[. \end{cases} \tag{5.71}$$

with Dirichlet boundary limits. The exact stationary solution of this equation is equal for all time to the initial data and the difference between the exact solution and the approximation is given in Figure 13. We remark here that for any time step t and any number of discretization points N, the error is negligible.

We consider the same equation but with a non polynomial second member, that is to say

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = (-x^2 + 4x - 1)e^{-x} \tag{5.72}$$

and the exact stationary solution is

$$(x^2 - 1)e^{-x}.$$

The computation is given in Figure 14. We can remark here that unlike the previous computations the error is given mainly by the spatial discretization.

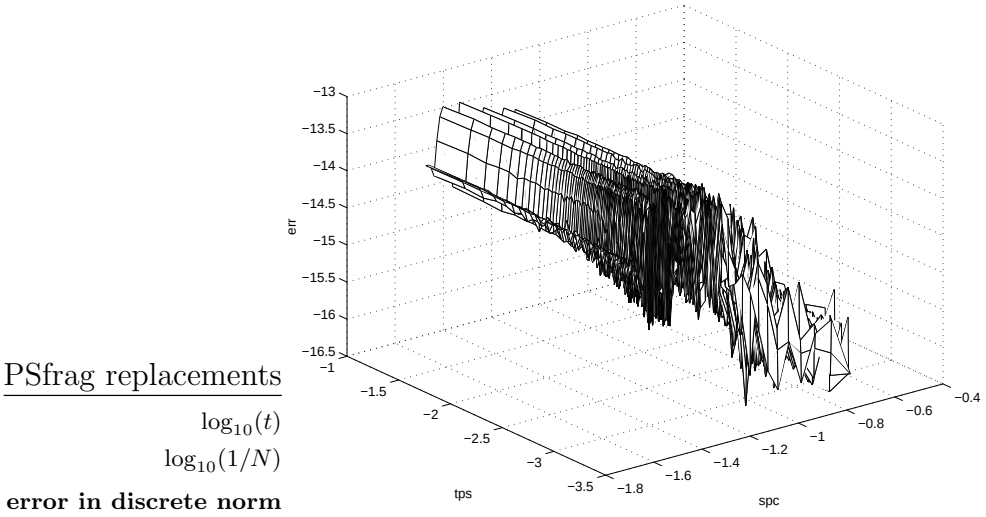


FIGURE 13. Plot of the error as a function of t and $1/N$ in case of problem (5.71); the computation is made with the residual smoothing scheme P. This slightly disgusting graph displays basically the round off error.

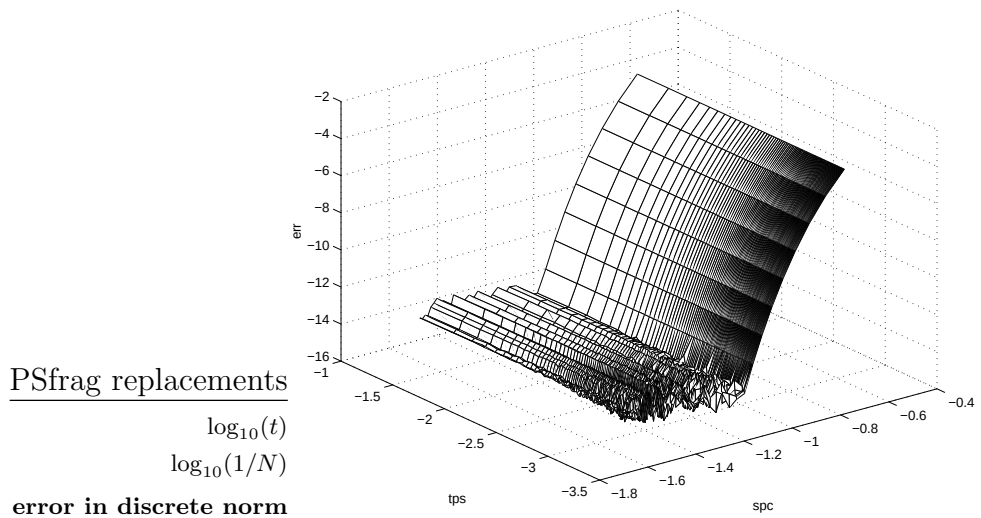


FIGURE 14. Plot of the error as a function of t and $1/N$ in case of problem (5.72).

We consider the same sort of equation with a stationary solution of class C^1 but not C^2, equal to

$$f(x) = \begin{cases} 1-(2x+1)^2, & \text{for } x \in [-1,-1/2] \\ 1-(2x+1)^2/9, & \text{for } x \in [-1/2,1] \end{cases} \tag{5.73}$$

The second member of the equation is thus equal to $-f''$. The result with the RSS scheme is shown in Figure 15 and Figure 16 plots the

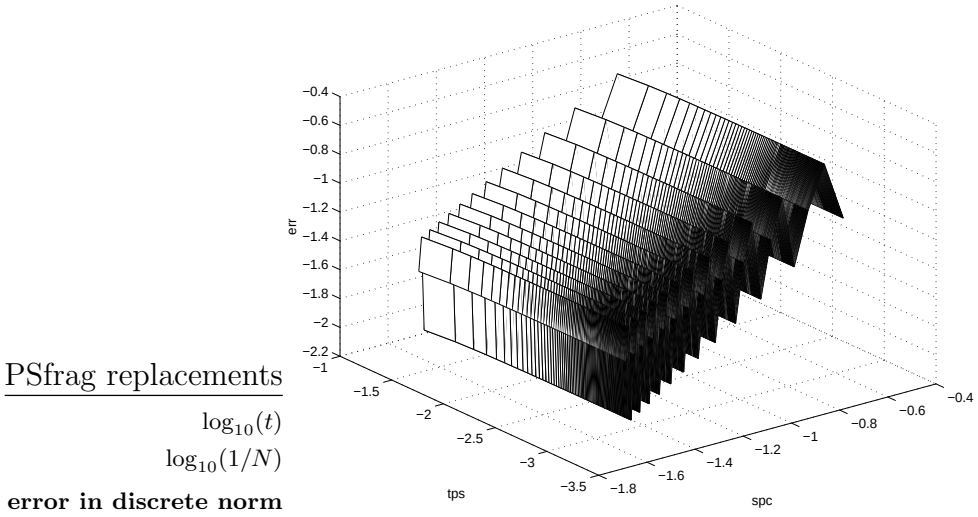


FIGURE 15. Plot of the error as a function of t and $1/N$ in case of the problem with solution (5.73). The space error can be seen in Figure 16 and the time error is too small in comparison with the space error to be seen.

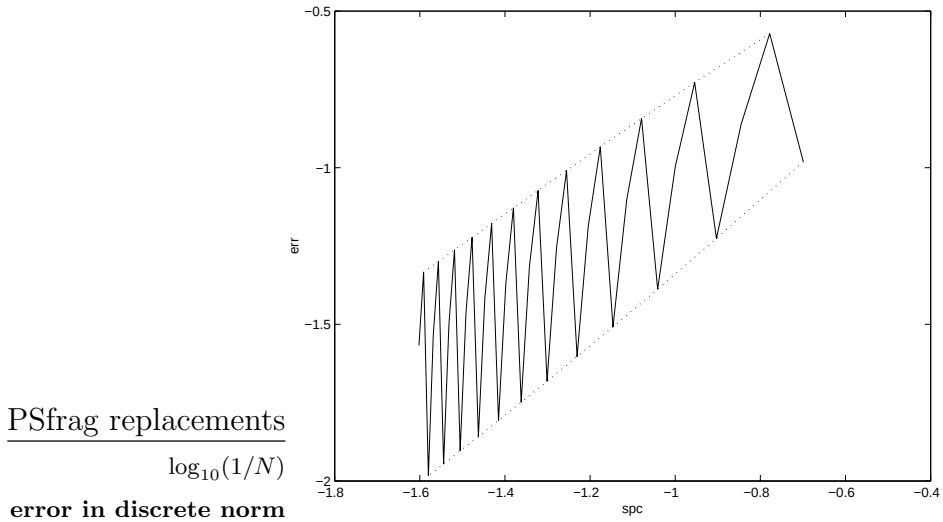


FIGURE 16. Plot of the error as a function of $1/N$ for $t = 5.10^{-4}$ in case of the problem with solution (5.73). The oscillations of the error are of period 3 and the slope of the two dotted lines is about 1.1327.

error as a function of $1/N$ with $t = 5.10^{-4}$. The spatial error is of order one as expected.

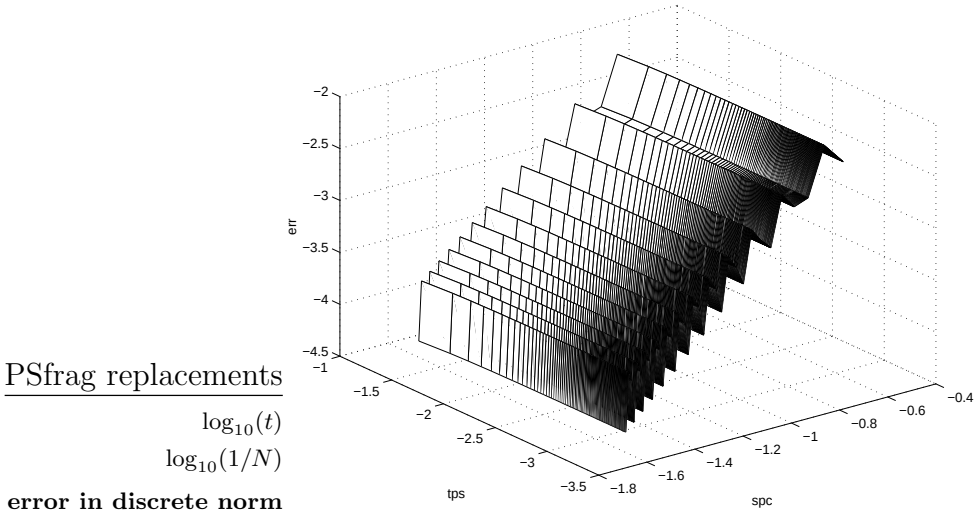


FIGURE 17. Plot of the error as a function of t and $1/N$ in case of the problem with solution (5.74). The space error can be seen in Figure 18. As for the time error, it can be seen for N large enough; for example, for $N = 40$, the slope of the regression line of the logarithm of the error as a function of $\log_{10}(t)$ is equal to 1.0200.

We consider the same problem but with a solution C^2 but not C^3, that is to say:

$$f(x) = \begin{cases} -x^3/3 - x^2 - x/4 + 5/12, & \text{for } x \in [-1, -1/2] \\ -x^2/2 + 11/24, & \text{for } x \in [-1/2, 1/2] \\ x^3/3 - x^2 + x/4 + 5/12, & \text{for } x \in [1/2, 1] \end{cases} \tag{5.74}$$

The plot of the error as a function of t and $1/N$ using RSS scheme can be seen in Figure 17 and Figure 18 plots the error as a function of $1/N$ with $t = 5.10^{-4}$. The time error is negligible and the spatial error is here of order two.

We still consider the same problem solved with RSS but now with a solution C^3 but not C^4, that is to say:

$$f(x) = N_{0,4}(5(x+1)/2) \tag{5.75}$$

where $N_{0,4}$ is a B-spline defined as in [**108**]. The plot of the error as a function of t and $1/N$ is plotted in Figure 19 and Figure 20 shows the error as a function of $1/N$ with $t = 5.10^{-4}$. The oscillations of the error become worse and worse, but the spatial error seems to be more or less of order three.

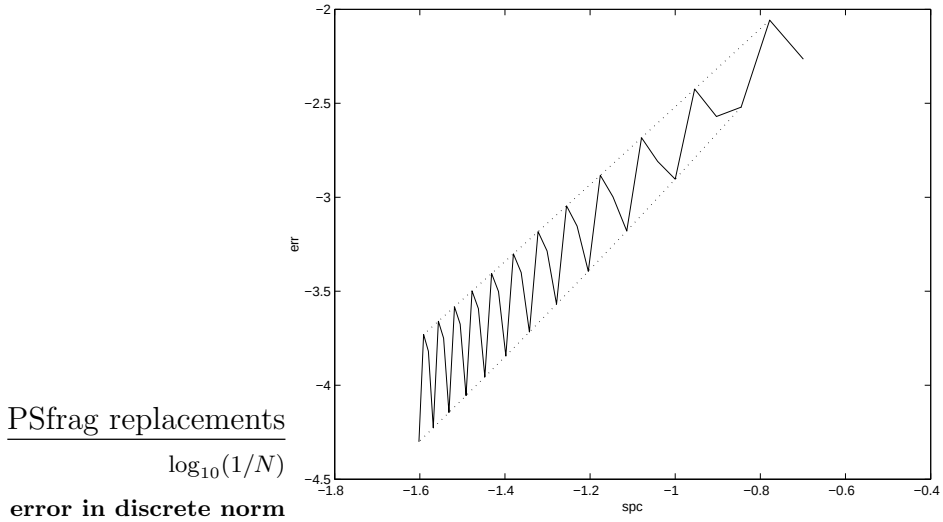


FIGURE 18. Plot of the error as a function of $1/N$ for $t = 5.10^{-4}$ in case of the problem with solution (5.74). The oscillations of the error are still of period 3 and the slope of the two dotted lines is about 2.3434.

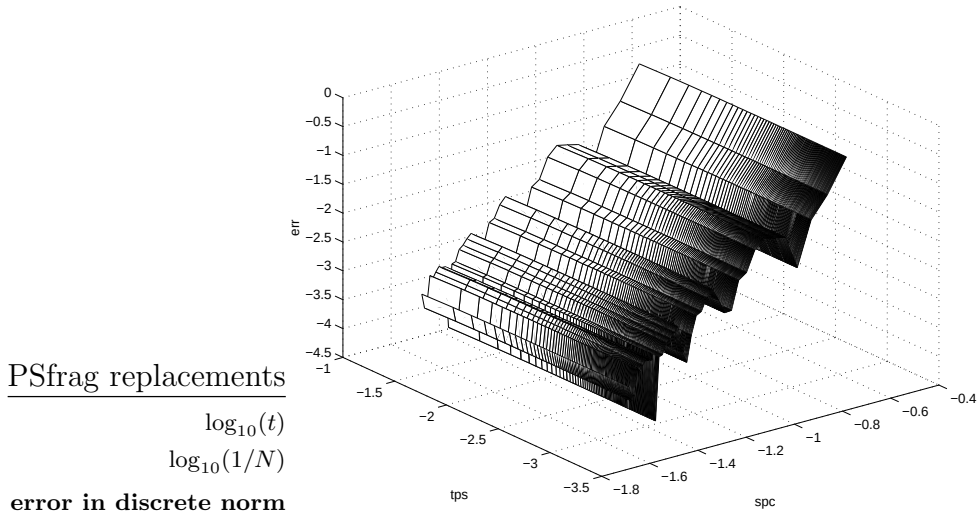


FIGURE 19. Plot of the error as a function of t and $1/N$ in case of the problem with solution (5.75). The space error can be seen in Figure 20. As for the time error, the slope of the regression line of the logarithm of the error as a function of $\log_{10}(t)$ is equal to 0.9487 for $N = 24$ and to 1.1946 for $N = 40$.

All these computations show that, regarding the residual smoothing scheme, the spatial error behaves nearly like the spatial error of the

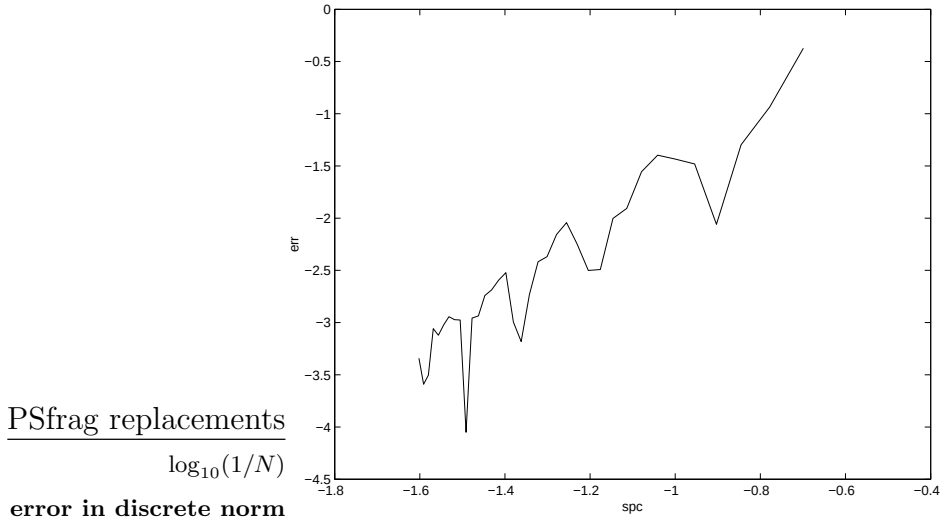


FIGURE 20. Plot of the error as a function of $1/N$ for $t = 5.10^{-4}$ in case of the problem with solution (5.75). The slope of the regression line is about 2.9472.

spectral method, as expected. It does not seem to be too affected by the finite differences preconditioning.

6. Développement asymptotique des polynômes de Legendre et de leurs points critiques

Asymptotic expansion of Legendre polynomials and asymptotics of their extrema

Magali Ribot [1]

Abstract : *Motivated by questions on the preconditioning of spectral methods, and independently of the extensive literature on the approximation of zeroes of orthogonal polynomials, either by the Sturm method, or by the descent method, we develop a stationary phase-like technique for calculating asymptotics of Legendre polynomials. The difference with the classical stationary phase method is that the phase is a nonlinear function of the large parameter and the integration variable, instead of being a product of the large parameter by a function of the integration variable. We then use an implicit functions theorem for approximating the zeroes of the derivatives of Legendre polynomials. This result is used for proving order and consistency of the residual smoothing scheme* [**4**], [**99**].

6.1. Introduction. Spectral methods produce full matrices; therefore, their numerical efficiency depends on the introduction of appropriate preconditioners. In the case of a Laplace — or more generally an elliptic — operator, finite differences or finite elements methods have been proposed for preconditioning spectral methods in Orszag [**87**], Haldenwang et al. [**63**] , Canuto and Quarteroni [**24**] or Deville and Mund [**53, 54**].

Denote by K_S the stiffness matrix associated to a Legendre–Gauss–Lobatto method for $-d^2/dx^2$ with Dirichlet boundary conditions, and by K_F the stiffness matrix associated to the P_1 finite elements method on the nodes of this spectral method.

Let M_S be the mass matrix of the spectral method, and let M_F be the mass-lumped mass matrix of the P_1 finite elements method constructed on the nodes of the spectral method.

Recent results of Parter [**90**] give the following bounds:

$$\frac{1}{C} \leq \frac{\mathrm{Re}\,(K_F M_S M_F^{-1} U, U)}{(K_S U, U)} \leq \frac{|(K_F M_S M_F^{-1} U, U)|}{(K_S U, U)} \leq C.$$

Here (,) denotes the hermitian scalar product. These results are based on [**89**], which itself builds on Gatteschi's results from [**61**]. When M_F is not mass-lumped, Parter proves an analogous result in [**91**].

[1] I would like to thank very warmly Michelle Schatzman for pointing me out this subject and for many helpful discussions. Many thanks are due to Seymour Parter and David Gottlieb for their numerous advice and encouragements.

The main result of [**97**] is the spectral equivalence between

$$M_F^{-1/2} M_S^{1/2} K_F M_S^{1/2} M_F^{-1/2}$$

and K_S. As a consequence of a result of Parter and Rothman [**92**], it suffices to prove the spectral equivalence between K_F and

$$M_F^{-1/2} M_S^{1/2} K_F M_S^{1/2} M_F^{-1/2}.$$

This question is motivated by the analysis of the residual smoothing scheme (see [**4**] and [**97**]), which allows for fast time integration of the spectral approximation of parabolic equation.

It turns out that when we started working on this question, we were not aware of Parter's results, and we did not consult the recent literature on orthogonal polynomials; instead of using a Sturm method or a descent method, as is done by most authors in this field, we took the classical integral representation formula for ultra-spherical polynomials (4.10.3) from Szegő [**118**], and we applied to this formula a stationary phase strategy, in a region where the classical expansions cannot be applied; this method gives an expansion at all orders, with estimates for the error bound. Let us point out that this is not a classical stationary phase method, since the exponential term is a non linear function of the large parameter and of the integration variable.

Though our result on preconditioning can be obtained with Parter's method, we feel that our treatment of the asymptotics is novel.

Let us describe why we need precise asymptotics of the zeroes of the derivatives of Legendre polynomials to prove the equivalence between $M_F^{-1/2} M_S^{1/2} K_F M_S^{1/2} M_F^{-1/2}$ and K_F. Let us also precise our notations.

We denote by $\mathbb{P}_N$ the space of polynomial functions of degree N defined over $[-1, 1]$. Let us denote by L_N the Legendre polynomial of degree N and let $-1 = \xi_0 < \xi_1 < \cdots < \xi_{N-1} < \xi_N = 1$ be the roots of $(1 - x^2)L'_N$; they are the nodes of the spectral method. Let ρ_k, $0 \le k \le N$ be the weights of the quadrature formula associated to the nodes ξ_k; since this is a Gauss-Lobatto formula, we shall have

$$\forall \Phi \in \mathbb{P}_{2N-1}, \int_{-1}^{1} \Phi(x)dx = \sum_{k=0}^{N} \Phi(\xi_k)\rho_k; \tag{6.1}$$

the weights ρ_k are strictly positive.

Bernardi and Maday [**9**] give explicit expressions of the ρ_k's:

$$\begin{aligned} \rho_0 = \rho_N &= \frac{2}{N(N+1)}, \\ \rho_k &= \frac{2}{N(N+1)L_N^2(\xi_k)}, \ 1 \le k \le N-1. \end{aligned} \tag{6.2}$$

We define η_k by

$$\eta_k = \mathrm{Arccos}(\xi_k). \tag{6.3}$$

Since we have

$$-1 = \xi_0 < \xi_1 < \cdots < \xi_{N-1} < \xi_N = 1,$$

we infer that

$$0 = \eta_N < \eta_{N-1} < \cdots < \eta_1 < \eta_0 = \pi.$$

The matrices M_S and M_F are diagonal; we define the diagonal elements of $M_F^{-1} M_S$ as:

$$\sigma_k = \frac{2\rho_k}{\xi_{k+1} - \xi_{k-1}}, \quad \text{for } 1 \le k \le N-1. \tag{6.4}$$

We make the convention that $\sigma_0 = \sigma_N = 0$.

Remark 6.1. *Since L_N is even (resp. odd) when N is even (resp. odd), we see that*

$$\xi_{N-k} = -\xi_k, \quad \textit{for } 1 \le k \le N-1. \tag{6.5}$$

Define the discrete H^1 norm by

$$\|U\|_{\mathbb{H}_N^1} = (U^* K_F U)^{1/2} = \left(\sum_{k=0}^{N-1} \frac{|U_{k+1} - U_k|^2}{\xi_{k+1} - \xi_k} \right)^{1/2};$$

the equivalence between $M_F^{-1/2} M_S^{1/2} K_F M_S^{1/2} M_F^{-1/2}$ and K_F is equivalent to the existence of a constant $C > 0$ such that

$$\|U\|_{\mathbb{H}_N^1}/C \le \|M_F^{-1/2} M_S^{1/2} U\|_{\mathbb{H}_N^1} \le C \|U\|_{\mathbb{H}_N^1}.$$

Here, as is classical, we had to extend the definition of U_k by letting $U_0 = U_N = 0$.

We first decompose $\sqrt{\sigma_{k+1}}\, U_{k+1} - \sqrt{\sigma_k}\, U_k$ as

$$\frac{\sqrt{\sigma_{k+1}} + \sqrt{\sigma_k}}{2} \big(U_{k+1} - U_k\big) + \frac{\sqrt{\sigma_{k+1}} - \sqrt{\sigma_k}}{2} \big(U_{k+1} + U_k\big). \tag{6.6}$$

The discrete H^1 norm of the first term in (6.6) can be estimated simply by

$$\max_k \left| \frac{\sqrt{\sigma_{k+1}} + \sqrt{\sigma_k}}{2} \right| \left(\sum_{k=0}^{N-1} \frac{(U_{k+1} - U_k)^2}{\xi_{k+1} - \xi_k} \right)^{1/2},$$

and we prove in [**97**] that σ_k is bounded independently of k and N.

The bulk of our work consists in proving that the second term of (6.6) can also be estimated in terms of the discrete H^1 norm of U. In a first step, we observe that discrete Hölder continuity estimates give

$$|U_{k+1} + U_k|^2 \le \big(2 - |\xi_k| - |\xi_{k+1}|\big) \|U\|_{\mathbb{H}_N^1}^2.$$

Thus, we are reduced to estimate

$$\sum_{k=0}^{N-1} \frac{2 - |\xi_k| - |\xi_{k+1}|}{\xi_{k+1} - \xi_k} \big|\sqrt{\sigma_{k+1}} - \sqrt{\sigma_k}\big|^2.$$

But σ_k is bounded from above and from below independently of k [**97**]; we define

$$\mu_k = \frac{2 - |\xi_k| - |\xi_{k+1}|}{\xi_{k+1} - \xi_k} \left| \frac{1}{\sigma_{k+1}} - \frac{1}{\sigma_k} \right|^2 \tag{6.7}$$

which is algebraically simpler but analytically equivalent, and it suffices to show [**97**]

$$\Sigma_N = \sum_{k=0}^{N-1} \mu_k \text{ is bounded independently of } N. \tag{6.8}$$

From here, we make the convention $1/\sigma_0 = 1/\sigma_N = 0$.

We deduce from symmetry (6.5), formulas (6.2), (6.4) and (6.7) that

$$\mu_{N-k} = \mu_{k-1}, \quad 1 \le k \le N.$$

Denote by $\lfloor r \rfloor$ the largest integer at most equal to the real r. Define $N' = \left\lfloor \frac{N-1}{2} \right\rfloor$; it suffices to estimate

$$\Sigma'_N = \sum_{k=0}^{N'} \mu_k \tag{6.9}$$

since $\Sigma_N \le 2\Sigma'_N$.

Therefore from the definitions (6.7), (6.4) and (6.2) of μ_k, σ_k and ρ_k, we have to provide asymptotic expansions for L_N and for the zeroes of L'_N; we start from classical integral or asymptotic formulas for Jacobi polynomials that can be found in the literature.

We partition the interval $\{0, \cdots, N'\}$ into three subintervals: $\{0, \cdots, K\}$, $\{K+1, \cdots, \lfloor \Lambda N \rfloor\}$ and $\{\lfloor \Lambda N \rfloor + 1, \cdots, N'\}$ where K is bounded and will be chosen later, and Λ belongs to the open interval $(0, 1/2)$.

Let us begin with the leftmost region $1 \le k \le K$, where, since K is kept finite, it suffices to find the limit of μ_k for N tending to infinity. Asymptotics for the Legendre polynomials and their derivatives in this region are available as follows: if N tends to infinity and z is bounded by πK, then

$$L_N\left(\cos \frac{z}{N}\right) \sim J_0(z)$$

where J_0 is the classical Bessel function; an analogous statement holds for L'_N (formula (8.1.1) of Szegő [**118**]). If z_k denotes the k-th positive zero of the Bessel function J_1, we find for $k \ge 1$:

$$\lim_{N \to +\infty} \mu_k = \frac{z_{k+1}^2 + z_k^2}{64(z_{k+1}^2 - z_k^2)} \left| (z_{k+1}^2 - z_{k-1}^2) J_0^2(z_k) - (z_{k+2}^2 - z_k^2) J_0^2(z_{k+1}) \right|^2.$$

Therefore, we obtain easily the estimate and we do not consider the region $1 \le k \le K$ in this article, since we do not need new asymptotics.

Another result from Szegő's book [**118**], formula (8.21.14), is: if Λ belongs to $(0, 1/2)$ and $\pi\Lambda N \le z \le \pi(1-\Lambda)N$, we have

$$\begin{aligned} P_N^{(1/2)}(\cos(z/N)) &= L_N\left(\cos\frac{z}{N}\right) \\ &= 2\omega_{N,1/2}\sum_{\nu=0}^{p-1}\omega_{\nu,1/2}\frac{1\times 3\cdots\times(2\nu-1)}{(2N-1)(2N-3)\cdots(2N-2\nu+1)} \\ &\quad\times\frac{\cos\bigl((N-\nu+1/2)z/N-(\nu+1/2)\pi/2\bigr)}{(2\sin(z/N))^{\nu+1/2}}+O(N^{-p-1/2}) \end{aligned}$$

where $\omega_{N,1/2}$ is an explicitly known number and the remainder is uniform over the interval $[\pi\Lambda N, \pi(1-\Lambda)N]$. We also have analogous uniform asymptotics for L'_N, L''_N and L'''_N; therefore, in the region $\lfloor\Lambda N\rfloor \le k \le N'$ and thanks to a quantitative implicit function theorem, we can find an expansion in terms of k and N of the zero η_k of $\theta \mapsto L'_N(\cos\theta)$ which lies in a neighborhood of size $O(N^{-2})$ about

$$\eta_{0,k} = \pi - \frac{\pi/4 + k\pi}{N+1/2} = \frac{(N-k)\pi+\pi/4}{N+1/2}; \tag{6.10}$$

this will be done in Theorem 6.3.

There remains to treat the intermediate region, i.e. z between πK and $\pi\Lambda N$; it corresponds to $K \le k \le \lfloor\Lambda N\rfloor$. This case is not treated in the literature, and we had to devise the estimates and their proof, using the stationary phase method.

Denote by $P_N^{(\lambda)}$ the ultra-spherical polynomial of degree N, i.e. the orthogonal polynomial of degree N relatively to the weight $(1-x^2)^{\lambda-1/2}$.

Remark 6.2. *The Legendre polynomial L_N of degree N is precisely equal to $P_N^{(1/2)}$, and as a consequence of (4.7.14) from* [**118**], *L'_N is equal to $P_{N-1}^{(3/2)}$.*

In order to find asymptotics in the intermediate region, we write an integral representation for $P_N^{(\lambda)}$:

$$P_N^{(\lambda)}(x) = \frac{2^{1-2\lambda}}{(\Gamma(\lambda))^2}\frac{\Gamma(N+2\lambda)}{N!}\int_0^\pi\left(x+i\sqrt{1-x^2}\cos\varphi\right)^N\sin^{2\lambda-1}\varphi\,d\varphi.$$

We apply the principle of the stationary phase method as described in Lemma 7.7.3 of Hörmander's book [**66**], but we cannot apply directly the lemma, since the phase is not equal to a large parameter multiplied by a real function of all the other variables: it is a complex function of the large parameter N and all the other variables. We set

$$\chi_N = -iN\sin(z/N)e^{-iz/N} \tag{6.11}$$

and, for λ such that $2\lambda - 1$ is an even integer, we eventually find polynomials $Q_{\nu,\lambda}$ such that

$$\Bigg| P_N^{(\lambda)}(\cos(z/N)) - 2\sqrt{\pi}\frac{2^{1-2\lambda}}{\Gamma(\lambda)^2}\frac{\Gamma(N+2\lambda)}{N!}\mathrm{Re}\left\{ie^{iz}\sum_{\nu=\lambda-1/2}^{\ell-1}\chi_N^{-(\nu+1/2)}Q_{\nu,\lambda}(\chi_N/N)\right\}\Bigg| \leq C(K,\Lambda,\ell,\lambda)\left(N^{-1}+z^{-1}\right)^{\ell-2\lambda+1};$$

here $\chi_N^{-(\nu+1/2)}$ is the principal determination and $C(K,\Lambda,\ell,\lambda)$ depends only on its arguments (Theorem 6.19). We will explain in Remark 6.18 why we could not extend our method to other parameters λ. Finally, we use once again a quantitative implicit function theorem to obtain an asymptotic expansion of the zero of L'_N which lies in a neighborhood of size $O(1/N^2)$ about $\pi(N-k+1/4)/(N+1/2)$, for $K \leq k \leq \lfloor \Lambda N \rfloor$ (Corollary 6.22); hence we obtain in [**97**] an estimate on the sum of the μ_k's for $K \leq k \leq \lfloor \Lambda N \rfloor$.

For the reader's convenience, it is advisable to consult the fourth edition of Szegő's book [**118**], which is the most complete.

The article is organized as follows: in section 6.2, we compute the asymptotics of the zeroes in the rightmost region thanks to an implicit functions theorem. Section 6.3, devoted to the middle region, is split into four subsections: in subsection 6.3.1, we explain the proof strategy; in subsection 6.3.2, we prove a general lemma of stationary or non stationary phase method and we apply it in subsection 6.3.3 to obtain expansions of Legendre polynomials; we finally obtain asymptotics of the zeroes of their derivative in subsection 6.3.4.

6.2. The region $\lfloor \Lambda N \rfloor \leq k \leq N'$. In order to obtain asymptotics for μ_k in the index range $k \in \{\lfloor \Lambda N \rfloor, \cdots, N'\}$ in [**97**] as explained in the introduction, we need first asymptotics for the zeroes of $P_N^{(3/2)}$.

It is more convenient to state the following lemma in an interval which is symmetric about $N/2$:

Theorem 6.3. *Define*

$$\theta_{0,k} = \frac{\pi/4 + k\pi}{N+3/2}.$$

Then for all $\Lambda \in (0,1/2)$, there exist C, C' such that for all $N \geq 2$ and for all integer k in $\{\lfloor \Lambda N \rfloor, \cdots, \lceil (1-\Lambda)N \rceil\}$, there exists a unique zero θ_k of $P_N^{(3/2)}(\cos\theta)$ in a ball of radius C'/N^2 about $\theta_{0,k}$; moreover the following estimate holds

$$\left|\theta_k - \theta_{0,k} + \frac{3}{8N^2\tan\theta_{0,k}} - \frac{9}{8N^3\tan\theta_{0,k}}\right| \leq CN^{-4}. \tag{6.12}$$

PROOF. The idea of the proof is to use the quantitative implicit function theorem given in [**47**]; let us state it here for the reader's convenience:

LEMMA 6.4. *Let X and Z be Banach spaces, and let f be a C^2 function from a neighborhood $\mathcal{U}$ of $x_0 \in X$ to Z. Let $z_0 = f(x_0)$. Assume that $A = Df(x_0)$ has a bounded inverse A^{-1}. Assume that the ball of radius ρ and of center x_0 is included in $\mathcal{U}$. Let*

$$M = \sup_{|\xi|\le\rho} \|A^{-1}D^2 f(x_0+\xi)\|.$$

There exist constants a and K given by

$$a = \min(1, (2\rho M)^{-1}), \quad K = \frac{3a\rho}{4}$$

such that if $|A^{-1}z_0| \le K$, the equation

$$f(x) = 0$$

possesses a unique solution in the ball $\{|x - x_0| \le a\rho\}$; moreover, this solution satisfies

$$|x - x_0| \le 2|A^{-1}z_0| \text{ and } |x - x_0 + A^{-1}z_0| \le 2M|A^{-1}z_0|^2.$$

As $P_N^{(3/2)}$ has the same parity as N, the set of zeroes of $P_N^{(3/2)}$ is invariant by the symmetry $x \mapsto -x$, and therefore, at the index level, θ_k is a zero of $P_N^{(3/2)}$ iff θ_{N-k} is a zero of $P_N^{(3/2)}$, and moreover, $\theta_{N-k} = \pi - \theta_k$. Therefore, it suffices to prove the lemma for $\Lambda N \le k \le N'$.

The definition of the binomial coefficients is extended for all $x \in \mathbb{C}$ and all integer $l \ge 0$ as

$$\binom{x}{l} = \frac{x(x-1)\cdots(x-l+1)}{l!};$$

this expression vanishes if x is set equal to 0 or if l is a negative integer. We use the notation

$$\omega_{N,\lambda} = \binom{N+\lambda-1}{N} = \frac{\Gamma(N+\lambda)}{\Gamma(N+1)\Gamma(\lambda)}, \tag{6.13}$$

and we exploit the asymptotics of $P_N^{(\lambda)}$ given as (8.21.14) of [**118**] for $\lambda = 3/2$, $5/2$ and $7/2$, since we need an estimate of $\partial^j f/\partial\theta^j$ for $j = 0, 1, 2$, in order to apply Lemma 6.4. We write the three term formula

(6.14)

$$\begin{aligned} P_N^{(3/2)}(\cos\theta) = \frac{2\omega_{N,3/2}}{(2\sin\theta)^{3/2}} \Bigg\{ &\cos\big((N+3/2)\theta - 3\pi/4\big) \\ &- \frac{3}{2(2N+1)} \frac{\cos\big((N+1/2)\theta - 5\pi/4\big)}{2\sin\theta} \\ &- \frac{15}{8(2N+1)(2N-1)} \frac{\cos\big((N-1/2)\theta - 7\pi/4\big)}{(2\sin\theta)^2} \Bigg\} + O(N^{-5/2}) \end{aligned}$$

which is uniform in θ in $[\Lambda/2, \pi/2]$ and in N; it is then convenient to define

$$f(\theta, N) = \frac{(2\sin\theta)^{3/2}}{2\omega_{N,3/2}} P_N^{(3/2)}(\cos\theta); \tag{6.15}$$

since we seek the unique root θ_k of f which belongs to a small neighborhood of $\theta_{0,k}$, we will have to calculate $f(\theta_{0,k}, N)$, $\partial f(\theta_{0,k}, N)/\partial\theta$ and to estimate $\partial^2 f(\theta, N)/\partial\theta^2$ in $[\theta_{0,k} - rN^{-2}, \theta_{0,k} + rN^{-2}]$; we will choose r later. We differentiate (6.15) twice, we use formula (4.7.14) from Szegő [**118**], viz.

$$\frac{d}{dx} P_N^{(\lambda)}(x) = 2\lambda P_{N-1}^{(\lambda+1)}(x)$$

and we find

$$\frac{\partial f}{\partial\theta}(\theta, N) = \frac{3}{2}\frac{f(\theta, N)}{\tan\theta} - \frac{3\sqrt{2}}{\omega_{N,3/2}} \sin^{5/2}\theta\, P_{N-1}^{(5/2)}(\cos\theta), \tag{6.16}$$

and

$$\begin{aligned}\frac{\partial^2 f}{\partial\theta^2}(\theta, N) = \frac{3}{4}\left(\frac{1}{\tan^2\theta} - 2\right) f(\theta, N) - \frac{12\sqrt{2}}{\omega_{N,3/2}} \cos\theta \sin^{3/2}\theta\, P_{N-1}^{(5/2)}(\cos\theta) \\ + \frac{15\sqrt{2}}{\omega_{N,3/2}} \sin^{7/2}\theta\, P_{N-2}^{(7/2)}(\cos\theta).\end{aligned} \tag{6.17}$$

We first calculate $f(\theta_{0,k}, N)$ with the help of formula (6.14) and we find

$$\begin{aligned}&f(\theta_{0,k}, N) = \\ &(-1)^k \left\{ \frac{3}{4(2N+1)\tan\theta_{0,k}} - \frac{15}{16(2N+1)(2N-1)\tan\theta_{0,k}} \right\} + O(N^{-3}).\end{aligned} \tag{6.18}$$

We can also evaluate $f(\theta, N)$ for $|\theta - \theta_{0,k}| \le rN^{-2}$: by Taylor expansion,

$$|\cos((N + 3/2)\theta - \pi/4)| \le r(N + 3/2)N^{-2},$$

and therefore

$$|\theta - \theta_{0,k}| \le rN^{-2} \Longrightarrow |f(\theta)| = (r + 1)O(N^{-1}), \tag{6.19}$$

the error term being uniform for k between $\lfloor \Lambda N \rfloor$ and N'.

We calculate now $\partial f(\theta_{0,k}, N)/\partial\theta$: first we substitute the value found at (6.18) into the first term on the right-hand side of (6.16); as $\theta_{0,k}$ is bounded away from 0 and π, this first term is an $O(N^{-1})$, uniformly for $\Lambda N \le k \le N'$. For the second term of the right-hand side of (6.16),

we need a two-term expansion of $P_N^{(5/2)}$, namely

(6.20)
$$P_N^{(5/2)}(\cos\theta) = \frac{2\omega_{N,5/2}}{(2\sin\theta)^{5/2}}\Big(\cos\big((N+5/2)\theta - 5\pi/4\big) - \frac{15}{8(N+3/2)}\frac{\cos\big((N+3/2)\theta - 7\pi/4\big)}{\sin\theta}\Big) + O(N^{-1/2}).$$

The error term is uniform on the interval $[\Lambda/2, \pi/2]$.

We replace N by $N-1$ in (6.20) and we observe that

$$\frac{3\sqrt{2}(\sin\theta)^{5/2}2\omega_{N-1,5/2}}{(2\sin\theta)^{5/2}\omega_{N,3/2}} = \frac{3}{2}\frac{\omega_{N-1,5/2}}{\omega_{N,3/2}} = N, \tag{6.21}$$

according to the definition (6.13) of $\omega_{N,\lambda}$. Furthermore,

$$\cos((N+3/2)\theta_{0,k} - 5\pi/4) = (-1)^{k-1}$$

and

$$\cos((N+1/2)\theta_{0,k} - 7\pi/4) = (-1)^k \sin(\theta_{0,k}).$$

Thus we find the asymptotic

$$A(k,N) = \frac{\partial f}{\partial\theta}(\theta_{0,k}, N) = (-1)^k(N + 15/8) + O(N^{-1}). \tag{6.22}$$

Now, we choose r:

$$r = \frac{4}{3}\sup\{N^2|f(\theta_{0,k}, N)/A(k,N)| : N \geq 1, \Lambda N \leq k \leq N'\};$$

our estimates show that r is indeed bounded.

There remains to give an estimate of $\partial^2 f(\theta, N)/\partial\theta^2$ over the interval $[\theta_{0,k} - rN^{-2}, \theta_{0,k} + rN^{-2}]$. The first term in the right-hand side of (6.17) is an $O(1/N)$, thanks to (6.19); the second term in the right-hand side of (6.17) is an $O(N)$ in virtue of (6.21) and the expansion (6.20); the last term in the right-hand side of (6.17) is estimated with the help of the one-term expansion of $P_N^{(7/2)}$ given by

$$P_N^{(7/2)}(\cos\theta) = \frac{2\omega_{N,7/2}}{(2\sin\theta)^{7/2}}\cos\big((N+7/2)\theta - 7\pi/4\big) + O(N^{3/2});$$

but $\omega_{N-2,7/2}/\omega_{N,3/2} = O(N^2)$ and by a Taylor expansion, $\cos((N + 3/2)\theta - 7\pi/4)$ is an $0(r/N)$ on the relevant interval. Therefore, we obtain the estimate

$$|\theta - \theta_{0,k}| \leq rN^{-2} \Longrightarrow \left|\frac{\partial^2 f}{\partial\theta^2}(\theta, N)\right| = (r+1)O(N), \tag{6.23}$$

and once again, the estimate is uniform with respect to k such that $\Lambda N \leq k \leq N'$, to r, and to N.

We have then $M = O(r+1) = O(1)$ and for all large enough N, $2rMN^{-2}$ is strictly less than 1, so that we may take $a = 1$ in the statement of Lemma 6.4. But then K is equal to $3r/4N^2$, and

by definition of r, $|f(\theta_{0,k},N)/A(k,N)| \leq K$, and the conclusion of the lemma applies. Relation (6.12) is simply the translation to our particular problem of the conclusion of Lemma 6.4. $\square$

6.3. The region $K \leq k \leq \lfloor \Lambda N \rfloor$. Let us prove the asymptotics of the zeroes of the derivatives of Legendre polynomials in the third region, which is the most difficult to handle. We first calculate expansions of Legendre polynomials and we begin by explaining the strategy of the proof.

6.3.1. *The strategy of the proof.* In order to calculate formulas for $L'_N(\cos\theta)$ and for $L_N(\cos\theta)$ and their derivatives, we will use the integral representation given by formula (4.10.3) of Szegő [**118**]: define

$$Z(\lambda,N) = \frac{2^{1-2\lambda}}{(\Gamma(\lambda))^2}\frac{\Gamma(N+2\lambda)}{N!}, \tag{6.24}$$

the following formula holds for $\lambda > 0$ and all $x \in [-1,1]$:

$$P_N^{(\lambda)}(x) = Z(\lambda,N)\int_0^\pi \left(x + i\sqrt{1-x^2}\cos\varphi\right)^N \sin^{2\lambda-1}\varphi\, d\varphi. \tag{6.25}$$

In fact, this formula is also true for all $x \in \mathbb{C}$, provided that we choose the appropriate determination of the square root appearing in the integrand.

We define the two following functions:

$$f_N(z,\varphi) = \big(\cos(z/N) + i\sin(z/N)\cos\varphi\big)^N$$

and g_N such that $f_N = \exp g_N$, i.e.

$$g_N(z,\varphi) = N\ln\big(\cos(z/N) + i\sin(z/N)\cos\varphi\big)$$

where we have taken the principal determination of the logarithm.

We infer from (6.25) the expression of the ultra-spherical polynomials at $x = \cos(z/N)$:

(6.26)

$$\begin{aligned} P_N^{(\lambda)}(\cos(z/N)) &= Z(\lambda,N)\mathrm{Re}\left(\int_0^\pi f_N(z,\varphi)\sin^{2\lambda-1}\varphi\ d\varphi\right) \\ &= Z(\lambda,N)\mathrm{Re}\left(\int_0^\pi \exp g_N(z,\varphi)\sin^{2\lambda-1}\varphi\ d\varphi\right) \end{aligned}$$

In our calculations, we will often need the following useful remark:

Remark 6.5. *The function g_N is an even function of φ and therefore its derivatives of odd order will vanish at $\varphi = 0$.*

We shall seek an asymptotic formula for $\int_0^\pi f_N(z,\varphi)\sin^{2\lambda-1}\varphi\ d\varphi$.

Let δ belong to $[0,\pi/4[$ and ψ be a cut-off function having the following properties

$$\begin{gathered} \psi \text{ is even, } \pi\text{-periodic, of class } C^\infty \text{ with values in } [0,1], \\ \psi \text{ is equal to 1 over } [0,\delta] \text{ and to 0 over } [2\delta,\pi/2]. \end{gathered} \tag{6.27}$$

The function ψ will enable us to localize difficulties.

Therefore, we can write

(6.28)
$$\int_0^\pi f_N(z,\varphi)\sin^{2\lambda-1}\varphi\,d\varphi = \int_0^\pi \psi(\varphi) f_N(z,\varphi)\sin^{2\lambda-1}\varphi\,d\varphi + \int_0^\pi (1-\psi(\varphi)) f_N(z,\varphi)\sin^{2\lambda-1}\varphi\,d\varphi.$$

We will apply a stationary phase strategy, meaning that the second integral in the right hand side of (6.28) is small: this statement is made precise at Corollary 6.10. The main effort is devoted to the estimate of

(6.29)
$$\int_0^\pi \psi(\varphi) f_N(z,\varphi)\sin^{2\lambda-1}\varphi\,d\varphi = \int_0^\pi \psi(\varphi)\exp g_N(z,\varphi)\sin^{2\lambda-1}\varphi\,d\varphi$$
$$= 2\mathrm{Re}\left(\int_0^{\pi/2} \psi(\varphi)\exp g_N(z,\varphi)\sin^{2\lambda-1}\varphi\,d\varphi\right)$$

by the stationary phase method.

We use a homotopy technique as in Hörmander's proof. Let q_N be the quadratic part of Taylor's expansion of $g_N(z,\cdot)$ at 0, i.e.

$$\begin{aligned} q_N(z,\varphi) &= g_N(z,0) + \frac{\varphi^2}{2}\frac{\partial^2 g_N}{\partial\varphi^2}(z,0) \\ &= iz - \frac{iN\varphi^2}{2}\sin(z/N)e^{-iz/N}, \end{aligned} \tag{6.30}$$

and define

$$R_N(z,\varphi) = g_N(z,\varphi) - q_N(z,\varphi). \tag{6.31}$$

The extensions of g_N and f_N as functions over $\mathbb{R}\times[0,\pi]\times[0,1]$ are given by

$$g_N(z,\varphi,s) = sg_N(z,\varphi)+(1-s)q_N(z,\varphi) = q_N(z,\varphi)+sR_N(z,\varphi) \tag{6.32}$$

and

$$f_N(z,\varphi,s) = \exp g_N(z,\varphi,s). \tag{6.33}$$

The double of the real part of the integral

$$\mathcal{I}_{N,\lambda}(z,s) = \int_0^{\pi/2} \psi(\varphi)\exp g_N(z,\varphi,s)\sin^{2\lambda-1}\varphi\,d\varphi \tag{6.34}$$

is equal to (6.29) for $s=1$ and for $s=0$, it can be expanded simply. Therefore, in order to estimate $\mathcal{I}_{N,\lambda}(z,1)$, we use a Taylor expansion at $s=0$, viz.

$$\left|\mathcal{I}_{N,\lambda}(z,1) - \sum_{l=0}^{k-1}\frac{1}{l!}\frac{\partial^l\mathcal{I}_{N,\lambda}}{\partial s^l}(z,0)\right| \le \max_{0\le s\le 1}\left|\frac{1}{k!}\frac{\partial^k\mathcal{I}_{N,\lambda}}{\partial s^k}(z,s)\right|. \tag{6.35}$$

We need explicit approximations of the terms $(\partial^l \mathcal{I}_{N,\lambda}/\partial s^l)(z,0)$ and we have to estimate all the remainders: some remainders come from the difference between $(\partial^l \mathcal{I}_{N,\lambda}/\partial s^l)(z,0)$ and its approximation; another remainder comes from $(\partial^k \mathcal{I}_{N,\lambda}/\partial s^k)(z,s)$.

The derivative $\partial^l \mathcal{I}_{N,\lambda}/\partial s^l$ is given by

$$\frac{\partial^l \mathcal{I}_{N,\lambda}}{\partial s^l}(z,s) = \int_0^{\pi/2} \psi(\varphi) R_N^l(z,\varphi) \exp g_N(z,\varphi,s) \sin^{2\lambda-1}\varphi \, d\varphi. \tag{6.36}$$

Let r_N be the Taylor expansion of $R_N(z,\cdot)$ of order $2(k+1)$ at 0:

$$r_N(z,\varphi) = \sum_{\gamma=2}^{k+1} \frac{\varphi^{2\gamma}}{(2\gamma)!} \frac{\partial^{2\gamma} g_N}{\partial \varphi^{2\gamma}}(z,0); \tag{6.37}$$

observe here that we do not have odd powers of φ, since R_N is even. Corollary 6.15 gives an estimate of

$$\int_0^{\pi/2} \psi(\varphi) \left(R_N^l(z,\varphi) - r_N^l(z,\varphi)\right) \exp q_N(z,\varphi) \, \sin^{2\lambda-1}\varphi d\varphi. \tag{6.38}$$

The usable explicit expressions will include only the first $k-l$ terms of $r_N^l(z,\varphi)$; we have to estimate the remaining $l(k-1)+1-(k-l) = k(l-1)+1$ terms, which is done at Corollary 6.17.

Finally, we estimate $\partial^k \mathcal{I}_{N,\lambda}/\partial s^k$ at Corollary 6.13.

The usable algebraic expressions are given first in general form at Theorem 6.19 for $2\lambda-1$ an even integer, and explicit results for $\lambda = 1/2$, $3/2$, $5/2$ and $7/2$ are given at Corollary 6.20.

6.3.2. *A general lemma of stationary and non-stationary methods.* We show a general lemma to help proving all the estimates explained in subsection 6.3.1.

We need several preliminary technical results. First we estimate $\exp(g_N(z,\varphi,s))$.

Lemma 6.6. *For all $N \geq 2$, for all $\varphi \in [0,\pi]$, for all $z \in \mathbb{R}^+$ and for all $s \in [0,1]$,*

$$|\exp(g_N(z,\varphi,s))| \leq 1.$$

Proof. It suffices to check $\operatorname{Re} g_N(z,\varphi,s) \leq 0$ which is true provided that $\operatorname{Re} g_N(z,\varphi)$ and $\operatorname{Re} q_N(z,\varphi)$ are less than or equal to 0.

The real part of g_N is $N \ln\left(1 - \sin^2(z/N)\sin^2\varphi\right)/2$ which has the required sign. The real part of q_N is $-N\varphi^2 \sin^2(z/N)/2$ which is also less than or equal to 0. □

Differentiating composite functions can be done with the help of Faa di Bruno's formula, see for instance Lemma II.2.8 of Hairer [**62**].

For $m \in \mathbb{N}$, let $\mathcal{C}(m)$ be the set of multi-indices $\gamma = (\gamma_1, \gamma_2, \cdots) \in \mathbb{N}^{\mathbb{N}^*}$ such that $\gamma_1 \geq \gamma_2 \geq \cdots$ and such that $\sum_{i\in\mathbb{N}} \gamma_i = m$. Therefore γ_i vanishes beyond a certain rank; we denote by $l(\gamma)$ the largest

integer such that $\gamma_i \geq 1$ and we observe that $l(\gamma) \leq m$. For instance, if we only write the non zero terms of each γ, $\mathcal{C}(3)$ is equal to $\{(3), (2,1), (1,1,1)\}$.

Faa di Bruno's formula states that there exist integer constants $C(\gamma, m)$ such that

$$\frac{d^m}{dx^m} A \circ B = \sum_{\gamma \in \mathcal{C}(m)} C(\gamma, m)(A^{(l(\gamma))} \circ B) \prod_{j=1}^{l(\gamma)} B^{(\gamma_j)}. \tag{6.39}$$

As a consequence of (6.39), we can calculate for any function B the derivatives of B^k for any integer k:

$$\frac{d^m}{dx^m} B^k = \sum_{\gamma \in \mathcal{C}(m)} C(\gamma, m) \binom{k}{l(\gamma)} l(\gamma)! B^{k-l(\gamma)} \prod_{j=1}^{l(\gamma)} B^{(\gamma_j)}. \tag{6.40}$$

Let us estimate now the derivatives of $(\partial g_N/\partial\varphi)^{-1}$, which will arise later when we will perform several integrations by part, and the derivatives of g_N.

Lemma 6.7. *For all $k \in \mathbb{N}$, for all $\alpha > 0$, there exists $C > 0$ such that for all $N \geq 2$, for all $\varphi \in (0, \pi - \alpha]$ and for all $z \in [\pi K, \pi \Lambda N]$, the following estimates hold*

$$\left| \frac{\partial^k}{\partial \varphi^k} \left(\frac{1}{\partial g_N/\partial\varphi} \right)(z, \varphi) \right| \leq \frac{C}{\varphi^{k+1}} (N^{-1} + z^{-1}) \tag{6.41}$$

and

$$\left| \frac{\partial^{k+1} g_N}{\partial \varphi^{k+1}}(z, \varphi) \right| \leq Cz. \tag{6.42}$$

Proof. Write

$$\begin{aligned} &\nu(\varphi) = \sin\varphi \text{ (as in numerator) and} \\ &d_N(z, \varphi) = \big(\cos(z/N) + i \sin(z/N) \cos\varphi\big) \text{ (as in denominator).} \end{aligned} \tag{6.43}$$

Then the first derivative of g_N and its inverse are

$$\frac{\partial g_N}{\partial \varphi}(z, \varphi) = -iN \sin(z/N) \frac{\nu(\varphi)}{d_N(z, \varphi)} \tag{6.44}$$

and

$$\frac{1}{\partial g_N/\partial\varphi}(z, \varphi) = \frac{i}{N \sin(z/N)} \frac{d_N(z, \varphi)}{\nu(\varphi)}. \tag{6.45}$$

Leibniz' formula gives

$$\frac{\partial^k}{\partial \varphi^k} \left(\frac{1}{\partial g_N/\partial\varphi} \right) = \frac{i}{N \sin(z/N)} \sum_{m=0}^{k} \binom{k}{m} \frac{\partial^{k-m} d_N}{\partial \varphi^{k-m}} \frac{\partial^m}{\partial \varphi^m} \left(\frac{1}{\nu} \right). \tag{6.46}$$

The successive derivatives of $1/\nu$ are computed using (6.40) for $k=-1$; up to arithmetic constants, the terms we find in (6.46) are of the form

$$\frac{i}{N\sin(z/N)}\frac{\partial^{k-m}d_N}{\partial\varphi^{k-m}}\nu^{-1-l(\gamma)}\prod_{j=1}^{l(\gamma)}\frac{\partial^{\gamma_j}\nu}{\partial\varphi^{\gamma_j}}; \tag{6.47}$$

we substitute the expressions of the derivatives

$$\frac{\partial^j d_N}{\partial\varphi^j}(z,\varphi)=i\sin(z/N)\cos\left(\varphi+j\pi/2\right),\ \text{for all } j\geq 1 \tag{6.48}$$

and

$$\frac{\partial^n\nu}{\partial\varphi^n}(\varphi)=\sin\left(\varphi+n\pi/2\right) \tag{6.49}$$

into (6.47): for $m=k$, the expressions (6.47) are equal to

$$\frac{i(\cos(z/N)+i\sin(z/N)\cos\varphi)}{N\sin(z/N)}\frac{1}{\sin^{1+l(\gamma)}\varphi}\prod_{j=1}^{l(\gamma)}\sin(\varphi+\gamma_j\pi/2)$$

which can be estimated by $C/(z\varphi^{k+1})$. For $m\leq k-1$, the terms (6.47) are of the form

$$-\frac{1}{N}\cos(\varphi+(k-m)\pi/2)\frac{1}{\sin^{1+l(\gamma)}\varphi}\prod_{j=1}^{l(\gamma)}\sin(\varphi+\gamma_j\pi/2)$$

which can be estimated by $C/(N\varphi^{k+1})$, proving thus (6.41).

Similarly, we write a Leibniz formula for $\partial^{k+1}g_N/\partial\varphi^{k+1}$:

$$\frac{\partial^{k+1}g_N}{\partial\varphi^{k+1}}(z,\varphi)=-iN\sin(z/N)\sum_{m=0}^{k}\binom{k}{m}\frac{\partial^{k-m}\nu}{\partial\varphi^{k-m}}(\varphi)\frac{\partial^m}{\partial\varphi^m}\left(\frac{1}{d_N}\right)(z,\varphi). \tag{6.50}$$

We use formula (6.40) with $k=-1$, i.e.

$$\begin{aligned}\frac{\partial^m}{\partial\varphi^m}\left(\frac{1}{d_N}\right)(z,\varphi)=&\sum_{\gamma\in\mathcal{C}(m)}C(\gamma,m)(-1)^{l(\gamma)}l(\gamma)!d_N^{-1-l(\gamma)}(z,\varphi)\\&\times\prod_{j=1}^{l(\gamma)}\frac{\partial^{\gamma_j}d_N}{\partial\varphi^{\gamma_j}}(z,\varphi);\end{aligned} \tag{6.51}$$

up to arithmetic constants, we substitute the values (6.48) and (6.49) of the derivatives of ν and d_N and for $k\geq 1$, Leibniz formula yields

$$\begin{aligned}&-iN\sin(z/N)\sin(\varphi+(k-m)\pi/2)\frac{(i\sin z/N)^{l(\gamma)}}{(\cos z/N+i\sin z/N\ \cos\varphi)^{1+l(\gamma)}}\\&\quad\times\prod_{j=1}^{l(\gamma)}\cos(\varphi+\gamma_j\pi/2).\end{aligned} \tag{6.52}$$

It is plain that the modulus of (6.52) is at most equal to $N|\sin z/N|$ and the conclusion of the lemma is clear. $\square$

The technical lemma 6.9 will be used many times in the foregoing estimates; it depends on the preliminary Lemma 6.8.

Let $p \in \mathbb{N}$ and $b \in [0, \pi)$. Let u be a function of class C^p over $[\pi K, +\infty) \times [0, b]$; assume that there exist a real $c \geq 2p$ and a real $l \geq 0$ such that the following norm

$$\|u\|_{p,c,l} = \max_{0 \leq i \leq p} \max_{N \in \mathbb{N}} \max_{\substack{\varphi \in (0,b] \\ z \in [\pi K, \pi \Lambda N]}} z^{-l} \varphi^{-c+i} \left| \frac{\partial^i u}{\partial \varphi^i}(z, \varphi) \right| \tag{6.53}$$

is finite. We define by induction

$$\begin{aligned} &U_0 = u \text{ and} \\ &U_{m+1} = \frac{\partial}{\partial \varphi} \left(\frac{U_m}{\partial g_N / \partial \varphi} \right) \text{ for all } m \in \{0, \cdots, p-1\}. \end{aligned} \tag{6.54}$$

We need to estimate the derivatives of the functions (6.54), since they will appear in the integration by parts which will be performed in the stationary and non stationary phase methods.

Lemma 6.8. *Let u be a function of class C^p over $[\pi K, +\infty) \times [0, b]$; assume that there exist $c \geq 2p$ and $l \geq 0$ such that $\|u\|_{p,c,l} < +\infty$. Then, there exists $C > 0$ such that for all $N \geq 2$, for all $m \in \{0, \cdots, p\}$, for all $q \in \{0, \cdots, p-m\}$, for all $\varphi \in [0, b]$ and for all $z \in [\pi K, \pi \Lambda N]$,*

$$\left| \frac{\partial^q}{\partial \varphi^q} U_m(z, \varphi) \right| \leq C \|u\|_{q+m,c,l} (N^{-1} + z^{-1})^{m-l} \varphi^{c-q-2m}. \tag{6.55}$$

Proof. Let us prove this lemma by induction on m. We have

$$\frac{\partial^q U_0}{\partial \varphi^q}(z, \varphi) = \frac{\partial^q u}{\partial \varphi^q}(z, \varphi)$$

and thus using the hypothesis made on $\|u\|_{p,c,l}$, we infer that

$$\left| \frac{\partial^q U_0}{\partial \varphi^q}(z, \varphi) \right| \leq \|u\|_{q,c,l} z^l \varphi^{c-q} \leq C \|u\|_{q,c,l} (N^{-1} + z^{-1})^{-l} \varphi^{c-q}.$$

Assuming that estimate (6.55) is proved for m, we use definition (6.54) and Leibniz formula:

$$\frac{\partial^q U_{m+1}}{\partial \varphi^q} = \sum_{s=0}^{q+1} \binom{q+1}{s} \frac{\partial^{q+1-s} U_m}{\partial \varphi^{q+1-s}} \frac{\partial^s}{\partial \varphi^s} \left(\frac{1}{\partial g_N / \partial \varphi} \right).$$

Using the induction hypothesis and Lemma 6.7,

$$\begin{aligned}\left|\frac{\partial^{q+1-s}U_m}{\partial\varphi^{q+1-s}}(z,\varphi)\frac{\partial^s}{\partial\varphi^s}\left(\frac{1}{\partial g_N/\partial\varphi}\right)(z,\varphi)\right| \\ &\leq C\,\|u\|_{q+m+1-s,c,l}\,(N^{-1}+z^{-1})^{m+1-l}\varphi^{c-q-2-2m} \\ &\leq C\,\|u\|_{q+m+1,c,l}\,(N^{-1}+z^{-1})^{m+1-l}\varphi^{c-q-2(m+1)},\end{aligned}$$

and the proof of Lemma 6.8 is complete. □

Here is our general lemma:

LEMMA 6.9. *Let $k\in\mathbb{N}^*$ and $b\in[0,\pi)$. Take u in $C_0^\infty([\pi K,+\infty)\times[0,b])$; assume that there exist $l\geq 0$ and $c\geq 2(k+l)$ such that $\|u\|_{k+l,c,l}$ is finite. Then there exists C such that for all $N\geq 2$ and all $z\in[\pi K,\pi\Lambda N]$,*

(6.56)
$$\max_{s\in[0,1]}\left|\int_0^b u(z,\varphi)\exp g_N(z,\varphi,s)\,d\varphi\right|\leq C\,\|u\|_{k+l,c,l}\,(N^{-1}+z^{-1})^k.$$

PROOF. Thanks to several integrations by part and using definition (6.54), we can write the integral appearing in the left hand side of (6.56) as

$$\begin{aligned}&\int_0^b u(z,\varphi)\exp g_N(z,\varphi,s)\,d\varphi= \\ &\sum_{m=0}^{k+l-1}\left[\frac{(-1)^m}{\partial g_N/\partial\varphi}U_m(z,\varphi)\exp g_N(z,\varphi,s)\right]_0^b \\ &\qquad+(-1)^{k+l}\int_0^b U_{k+l}(z,\varphi)\exp g_N(z,\varphi,s)\,d\varphi.\end{aligned}\tag{6.57}$$

Since u is equal to 0 in a neighborhood of $\varphi=b$, for all $m\in\{0,\cdots,k+l-1\}$, for all z in $[\pi K,\pi\Lambda N]$, $U_m(z,b)$ vanishes and thus all the integrated terms at $\varphi=b$ disappear:

(6.58)
$$\begin{aligned}\int_0^b u(z,\varphi)\exp g_N(z,\varphi,s)\,d\varphi&=\sum_{m=0}^{k+l-1}(-1)^m\frac{U_m}{\partial g_N/\partial\varphi}(z,0)\exp g_N(z,0,s) \\ &+(-1)^{k+l}\int_0^b U_{k+l}(z,\varphi)\exp g_N(z,\varphi,s)\,d\varphi.\end{aligned}$$

Thanks to Lemmas 6.7 and 6.8, we can estimate all these terms.

Lemma 6.7 with $k=0$ and Lemma 6.8 with $q=0$ give for m in $\{0,\cdots,k+l-1\}$, in the neighborhood of $\varphi=0$,

$$\frac{U_m}{\partial g_N/\partial\varphi}(z,\varphi)=O(1)(N^{-1}+z^{-1})^{m-l+1}\varphi^{c-2m-1}$$

where $O(1)$ is bounded independently of $\varphi \in [0,b]$, $z \in [\pi K, \pi\Lambda N]$, $N \geq 2$ and finite l and m. Since $c \geq 2(k+l) > 2m+1$, we obtain $\big[U_m/(\partial g_N/\partial\varphi)\big](z,0) = 0$. Moreover, $\exp g_N(z,0,s) = \exp(iz)$ and thus equation (6.58) becomes

$$\int_0^b u(z,\varphi)\exp g_N(z,\varphi,s)\,d\varphi$$
$$= (-1)^{k+l}\int_0^b U_{k+l}(z,\varphi)\exp g_N(z,\varphi,s)\,d\varphi.$$

Thanks to Lemma 6.6 and Lemma 6.8 with $m = k+l$ and $q = 0$, we obtain estimate (6.56). □

6.3.3. *Asymptotics of Legendre polynomials.* Now that Lemma 6.9 is proved, we can estimate the integrals displayed in subsection 6.3.1.

First, a straightforward corollary of Lemma 6.9 shows that the second integral of the right hand side of (6.28) is small.

Corollary 6.10. *Let ψ satisfy conditions* (6.27). *For all positive integer k and for all $\lambda > 0$, there exists C such that for all $N \geq 2$ and for all z in $[\pi K, \pi\Lambda N]$, the following estimate holds:*

$$\left|\int_0^\pi (1-\psi(\varphi))\exp g_N(z,\varphi)\sin^{2\lambda-1}\varphi\,d\varphi\right| \leq C(N^{-1}+z^{-1})^k. \tag{6.59}$$

Proof. We use Lemma 6.9 with $u(z,\varphi) = (1-\psi(\varphi))\sin^{2\lambda-1}\varphi$ and $b = \pi-\delta/2$. The function u and its derivatives vanish in a neighborhood of $\varphi = b$ and in the neighborhood $[-\delta,\delta]$ of 0; if we set $l = 0$ and $c = 2k$ and $\|u\|_{k,2k,0}$ is finite. We infer from Lemma 6.9 that

$$\left|\int_0^\pi (1-\psi(\varphi))\exp g_N(z,\varphi,1)\sin^{2\lambda-1}\varphi\,d\varphi\right|$$
$$= \left|\int_0^{\pi-\frac{\delta}{2}} (1-\psi(\varphi))\exp g_N(z,\varphi,1)\sin^{2\lambda-1}\varphi\,d\varphi\right|$$
$$\leq C\,\|u\|_{k,2k,0}\,(N^{-1}+z^{-1})^k,$$

where C depends only on k, which is estimate (6.59). □

In order to apply Lemma 6.9 to the remainder defined by equation (6.35), we need to estimate the derivatives of the powers of R_N, defined at equation (6.31).

Lemma 6.11. *For all $k \in \mathbb{N}^*$ and $m \in \mathbb{N}$, there exists $C > 0$ such that for all $N \geq 2$, for all $z \in [\pi K, \pi\Lambda N]$ and for all $\varphi \in [0,\pi/2]$,*

$$\left|\frac{\partial^m R_N^k}{\partial\varphi^m}(z,\varphi)\right| \leq Cz^k\min(1,\varphi^{4k-m}). \tag{6.60}$$

PROOF. For $k = 1$ and $m \leq 3$, Taylor's integral formula gives

$$\frac{\partial^m R_N(z,\varphi)}{\partial \varphi^m} = \int_0^\varphi \frac{\partial^4 g_N}{\partial \varphi^4}(z,\varphi') \frac{(\varphi-\varphi')^{3-m}}{(3-m)!}\, d\varphi',$$

and for $m \geq 4$,

$$\frac{\partial^m R_N}{\partial \varphi^m} = \frac{\partial^m g_N}{\partial \varphi^m}.$$

We infer immediately from these relations and the parity of R_N with respect to φ the estimates

$$\begin{aligned}
&|R_N(z,\varphi)| \leq C\varphi^4 z,\\
&\left|\frac{\partial R_N}{\partial \varphi}(z,\varphi)\right| \leq C\varphi^3 z,\\
&\left|\frac{\partial^2 R_N}{\partial \varphi^2}(z,\varphi)\right| \leq C\varphi^2 z,\\
&\left|\frac{\partial^m R_N}{\partial \varphi^m}(z,\varphi)\right| \leq C\varphi z \text{ for } m \geq 3,\ m \text{ odd},\\
&\left|\frac{\partial^m R_N}{\partial \varphi^m}(z,\varphi)\right| \leq Cz \text{ for } m \geq 4,\ m \text{ even}.
\end{aligned} \tag{6.61}$$

Therefore, using Faa di Bruno's formula (6.40), we obtain

$$\frac{\partial^m R_N^k}{\partial \varphi^m}(z,\varphi) = \sum_{\gamma \in \mathcal{C}(m)} C(\gamma,m) \binom{k}{l(\gamma)} l(\gamma)! R_N^{k-l(\gamma)}(z,\varphi) \prod_{j=1}^{l(\gamma)} \frac{\partial^{\gamma_j} R_N}{\partial \varphi^{\gamma_j}}(z,\varphi). \tag{6.62}$$

Let us denote by ν_1 the number of indices $j \in \{1, \cdots, l(\gamma)\}$ such that $\gamma_j = 1$, by ν_2 the number of indices $j \in \{1, \cdots, l(\gamma)\}$ such that $\gamma_j = 2$; ν_o is the number of indices j such that $\gamma_j \geq 3$, is odd and ν_e is the number of indices such that $\gamma_j \geq 4$ is even.

Thus, we have the two following relations:

$$\nu_1 + \nu_2 + \nu_o + \nu_e = l(\gamma) \tag{6.63}$$

and

$$m = \gamma_1 + \cdots + \gamma_l(\gamma) \geq \nu_1 + 2\nu_2 + 3\nu_o + 4\nu_e. \tag{6.64}$$

We infer from equation (6.61) the estimate

$$\left| R_N^{k-l(\gamma)}(z,\varphi) \prod_{j=1}^{l(\gamma)} \frac{\partial^{\gamma_j} R_N}{\partial \varphi^{\gamma_j}}(z,\varphi) \right| \leq C z^{\alpha} \varphi^{4k-4l(\gamma)+3\nu_1+2\nu_2+\nu_o}$$

where $\alpha = k - l(\gamma) + \nu_1 + \nu_2 + \nu_o + \nu_e$ and from equation (6.63), the estimate

$$\left| R_N^{k-l(\gamma)}(z,\varphi) \prod_{j=1}^{l(\gamma)} \frac{\partial^{\gamma_j} R_N}{\partial \varphi^{\gamma_j}}(z,\varphi) \right| \leq C z^{k} \varphi^{4k-4l(\gamma)+3\nu_1+2\nu_2+\nu_o}.$$

Equations (6.63) and (6.64) lead to

$$4k - 4l(\gamma) + 3\nu_1 + 2\nu_2 + \nu_o = 4k - \nu_1 - 2\nu_2 - 3\nu_o - 4\nu_e \geq 4k - m$$

and the expression $4k - 4l(\gamma) + 3\nu_1 + 2\nu_2 + \nu_o$ is also non negative since $l(\gamma)$ belongs to $\{0, \cdots, k\}$; this completes the proof of estimate (6.60). □

We deduce easily an analogous lemma for the derivatives of the powers of r_N, defined at equation (6.37).

Lemma 6.12. *For all $k \in \mathbb{N}^*$ and $m \in \mathbb{N}$, there exists $C > 0$ such that for all $N \geq 2$, for all $z \in [\pi K, \pi \Lambda N]$ and for all $\varphi \in [0, \pi/2]$,*

$$\left| \frac{\partial^m r_N^k}{\partial \varphi^m}(z,\varphi) \right| \leq C z^k \min(1, \varphi^{4k-m}). \tag{6.65}$$

Proof. The estimates for $\partial^m r_N / \partial \varphi^m$ are analogous to the estimates (6.61) and consequently the estimate for $\partial^m r_N^k / \partial \varphi^m$ is the same as estimate (6.60) for $\partial^m R_N^k / \partial \varphi^m$. □

Recall that $\mathcal{I}_{N,\lambda}$ has been defined at equation (6.34). The following corollary gives estimates of its derivatives.

Corollary 6.13. *For all integer $k \geq 1$ and all $\lambda \geq 1/2$, there exists C such that for all $N \geq 2$ and for all z in $[\pi K, \pi \Lambda N]$,*

$$\max_{s \in [0,1]} \left| \frac{\partial^k \mathcal{I}_{N,\lambda}}{\partial s^k}(z,s) \right| \leq C (N^{-1} + z^{-1})^k. \tag{6.66}$$

Proof. We use formula (6.36) and Lemma 6.9 with $u(z,\varphi) = \psi(\varphi) R_N^k(z,\varphi) \sin^{2\lambda-1}\varphi$ and $b = \pi/2$; since, in virtue of Lemma 6.11, $\|u\|_{2k,4k,k}$ is finite, we obtain from Lemma 6.9

$$\max_{s \in [0,1]} \left| \frac{\partial^k \mathcal{I}_{N,\lambda}}{\partial s^k}(z,s) \right| \leq C \, \|u\|_{2k,4k,k} \, (N^{-1} + z^{-1})^k,$$

that is estimate (6.66). □

We estimate in next lemma the derivatives of the difference between R_N^l and r_N^l, where r_N is defined at (6.37).

LEMMA 6.14. *For all $l \in \mathbb{N}^*$ and $m \in \mathbb{N}$, there exists $C > 0$ such that for all $N \geq 2$, for all $z \in [\pi K, \pi\Lambda N]$ and for all $\varphi \in [0, \pi/2]$*

$$\left|\frac{\partial^m}{\partial\varphi^m}(R_N^l - r_N^l)(z,\varphi)\right| \leq Cz^l \min(1, \varphi^{2k+4l-m}). \tag{6.67}$$

PROOF. First, as in Lemma 6.11, we consider the case $l = 1$ and we estimate the successive derivatives of $R_N - r_N$. We observe that the derivative of order m of r_N vanishes for $m \geq 2k+3$. We calculate the derivatives of $R_N - r_N$ in terms of the derivatives of g_N and using Taylor's formula and Lemma 6.7 we find the inequalities

(6.68)

$$\begin{aligned}
&\left|\frac{\partial^m (R_N - r_N)}{\partial\varphi^m}(z,\varphi)\right| \leq C\varphi^{2k+4-m}z, \text{ for } m \in \{0, \cdots, 2k+2\},\\
&\left|\frac{\partial^m (R_N - r_N)}{\partial\varphi^m}(z,\varphi)\right| \leq C\varphi z, \text{ for } m \geq 2k+3,\ m \text{ odd},\\
&\left|\frac{\partial^m (R_N - r_N)}{\partial\varphi^m}(z,\varphi)\right| \leq Cz, \text{ for } m \geq 2k+4,\ m \text{ even}.
\end{aligned}$$

We factorize $R_N^l - r_N^l$ as

$$R_N^l - r_N^l = (R_N - r_N)\left(R_N^{l-1} + R_N^{l-2}r_N + \cdots + r_N^{l-1}\right)$$

and a Leibniz formula gives

$$\begin{aligned}
\frac{\partial^m \left(R_N^l - r_N^l\right)}{\partial\varphi^m}(z,\varphi) = \sum_{\gamma=0}^{m}\sum_{\beta=0}^{\gamma}\sum_{\nu=0}^{l-1}\binom{m}{\gamma}\binom{\gamma}{\beta}\frac{\partial^{m-\gamma}(R_N - r_N)}{\partial\varphi^{m-\gamma}}(z,\varphi)\\
\times \frac{\partial^{\gamma-\beta}R_N^{l-1-\nu}}{\partial\varphi^{\gamma-\beta}}(z,\varphi)\frac{\partial^\beta r_N^\nu}{\partial\varphi^\beta}(z,\varphi).
\end{aligned}$$

Let us write

$$T_{\beta,\gamma,\nu} = \frac{\partial^{m-\gamma}(R_N - r_N)}{\partial\varphi^{m-\gamma}}(z,\varphi)\frac{\partial^{\gamma-\beta}R_N^{l-1-\nu}}{\partial\varphi^{\gamma-\beta}}(z,\varphi)\frac{\partial^\beta r_N^\nu}{\partial\varphi^\beta}(z,\varphi). \tag{6.69}$$

Thanks to estimate (6.68), Lemmas 6.11 and 6.12, we can estimate $T_{\beta,\gamma,\nu}$ as follows:

$$\begin{aligned}
|T_{\beta,\gamma,\nu}| &\leq Cz^l \min\left(1, \varphi^{2k+4-m+\gamma}\right)\min\left(1, \varphi^{4l-4-4\nu-\gamma+\beta}\right)\min\left(1, \varphi^{4\nu-\beta}\right)\\
&\leq Cz^l \min(1, \varphi^{2k+4l-m})
\end{aligned}$$

which proves estimate (6.67). □

We can now infer from Lemma 6.14 an estimate of the remainder (6.38):

COROLLARY 6.15. *For k in $\mathbb{N}^*$, $l \in \{0, \cdots, k-1\}$ and $\lambda \geq 1/2$, there exists $C > 0$ such that for all $N \geq 2$ and for all $z \in [\pi K, \pi \Lambda N]$*

$$\left| \int_0^{\pi/2} \psi(\varphi) \left(R_N^l(z,\varphi) - r_N^l(z,\varphi) \right) \exp q_N(z,\varphi) \sin^{2\lambda-1} \varphi \, d\varphi \right| \leq C(N^{-1} + z^{-1})^k. \tag{6.70}$$

PROOF. We set $u(z,\varphi) = \psi(\varphi) \left(R_N^l(z,\varphi) - r_N^l(z,\varphi) \right) \sin^{2\lambda-1} \varphi$ and $b = \pi/2$. We deduce from Lemma 6.14 that $\|u\|_{k+l,2k+4l,l}$ is finite and Lemma 6.9 yields equation (6.70). □

We state for the reader's convenience the one-dimensional version of Lemma 7.7.3 of Hörmander [**66**]:

LEMMA 6.16. *Assume $a \neq 0$ with $\Im(a) \geq 0$ and $u \in \mathcal{S}$, the Schwartz space over $\mathbb{R}$. Then for every $p \in \mathbb{N}^*$, there exists $C > 0$ such that*

$$\left| \int u(x) e^{iax^2/2} \, dx - \left(\frac{a}{2\pi i} \right)^{-1/2} T_p(u,a) \right| \leq C \left(\frac{1}{|a|} \right)^{p+1/2} \|u\|_{H^{2p+1}},$$

with

$$T_p(u,a) = \sum_{j=0}^{p-1} \frac{(2ia)^{-j}}{j!} \frac{\partial^{2j} u}{\partial \varphi^{2j}}(0). \tag{6.71}$$

Here, the principal determination of the fractional power is chosen.

We estimate the last remainder; the number χ_N is defined at equation (6.11). Let $\mathbf{1}_{[a,b]}$ be the characteristic function of $[a,b]$.

COROLLARY 6.17. *Let k in $\mathbb{N}^*$, $l \in \{0, \cdots, k-1\}$ and λ such that $2\lambda - 1$ is an even integer, there exists C such that for all $N \geq 2$, for all $z \in [\pi K, \pi \Lambda N]$,*

$$\begin{aligned} &\left| \int_0^{\pi/2} \psi(\varphi) r_N^l(z,\varphi) \sin^{2\lambda-1} \varphi e^{\chi_N \varphi^2/2} \, d\varphi - \frac{i}{2} \sqrt{\frac{2\pi}{\chi_N}} \sum_{j=0}^{k+l-1} \frac{1}{j!} \frac{1}{(2\chi_N)^j} \right. \\ &\left. \times \frac{\partial^{2j}}{\partial \varphi^{2j}} \left(\psi(\varphi) r_N^l(z,\varphi) \sin^{2\lambda-1} \varphi \mathbf{1}_{[-\pi/2,\pi/2]}(\varphi) \right)(z,0) \right| \\ &\leq C(N^{-1} + z^{-1})^{k+1/2}. \end{aligned} \tag{6.72}$$

Here, the principal determination of the square root has been chosen.

PROOF. We use Lemma 6.16 with

$u(z,\varphi) = \mathbf{1}_{[-\pi/2,\pi/2]}(\varphi) \psi(\varphi) r_N^l(z,\varphi) \sin^{2\lambda-1} \varphi$, $p = k+l$ and $a = -i\chi_N$,

the remainder is equal to $C \left| \chi_N \right|^{-(k+l+1/2)} \|u\|_{H^{2p+1}}$.

In virtue of Lemma 6.12, the norm $\|u\|_{2k+2l+1,4l,l}$ is finite and the remainder is bounded by

$$C\,\|u\|_{2k+2l+1,4l,l}\, z^l (N^{-1}+z^{-1})^{k+l+1/2}$$

which completes the proof. □

REMARK 6.18. *The parameter $2\lambda-1$ must be an integer here; otherwise the function u we use would not belong to the Schwartz space as required in Lemma 6.16. Moreover, $2\lambda-1$ must be even; otherwise, the integral we consider over $[-\pi/2,\pi/2]$ would vanish and would not be the double of the integral over $[0,\pi/2]$.*

Now that Lemmas 6.6 to 6.17 are proved, we can apply the strategy of proof described at the beginning of the present subsection to find an asymptotic formula for $P_N^{(\lambda)}$.

THEOREM 6.19. *Let $\lambda = 1/2,\, 3/2,\, 5/2,\, 7/2,\, \cdots$. Then, there exist real polynomials $Q_{\nu,\lambda}$ of degree ν for all $\nu\in\mathbb{N}$ such that, for all $k\in\mathbb{N}^*$, for all $K\in\mathbb{N}$ and for all $\Lambda\in(0,1/2)$, the following estimate holds for all $N\geq 2$ and for all $z\in[\pi K,\pi\Lambda N]$:*

$$\begin{aligned}\Bigg| & P_N^{(\lambda)}(\cos(z/N)) \\ & - 2\sqrt{\pi}Z(\lambda,N)\mathrm{Re}\left\{ie^{iz}\sum_{\nu=\lambda-1/2}^{k-1}\chi_N^{-(\nu+1/2)}Q_{\nu,\lambda}(\chi_N/N)\right\}\Bigg| \\ & \leq C(K,\Lambda,k,\lambda)\left(N^{-1}+z^{-1}\right)^{k-2\lambda+1},\end{aligned} \tag{6.73}$$

where $C(K,\Lambda,k,\lambda)$ depends only on the displayed arguments and the constant $Z(\lambda,N)$ is defined at equation (6.24).

PROOF. We split (6.26) as in (6.28). Corollary 6.10 implies that the second integral of the right hand side of (6.28) is an $O(N^{-1}+z^{-1})^k$.

We deduce from equation (6.35) and Corollary 6.13 that

$$\mathcal{I}_{N,\lambda}(z,1) = \sum_{l=0}^{k-1}\frac{1}{l!}\frac{\partial^l\mathcal{I}_{N,\lambda}}{\partial s^l}(z,0) + O\left((N^{-1}+z^{-1})^k\right). \tag{6.74}$$

Let us obtain an expression for

$$\frac{\partial^l\mathcal{I}_{N,\lambda}}{\partial s^l}(z,0) = \int_0^{\pi/2}\psi(\varphi)R_N^l(z,\varphi)\exp q_N(z,\varphi)\sin^{2\lambda-1}\varphi\,d\varphi.$$

We replace R_N by its Taylor expansion r_N defined at equation (6.37). We set

$$\mathcal{J}_{N,l,\lambda}(z) = \int_0^{\pi/2}\psi(\varphi)r_N^l(z,\varphi)\exp(\chi_N\varphi^2/2)\sin^{2\lambda-1}\varphi\,d\varphi.$$

Corollary 6.15 implies that

$$\frac{\partial^l \mathcal{I}_{N,\lambda}}{\partial s^l}(z,0) = e^{iz}\mathcal{J}_{N,l,\lambda}(z) + O\left((N^{-1}+z^{-1})^k\right).$$

We now use Corollary 6.17 to obtain an algebraic expression for $\mathcal{J}_{N,l,\lambda}$. Equation (6.72) yields

(6.75)

$$\mathcal{J}_{N,l,\lambda}(z) = i\sqrt{\pi}\sum_{j=0}^{k+l-1}\frac{1}{j!}\frac{1}{(2\chi_N)^{j+1/2}}\frac{\partial^{2j}(r_N^l(z,\varphi)\sin^{2\lambda-1}\varphi)}{\partial\varphi^{2j}}(z,0) + O\left((N^{-1}+z^{-1})^{k+1/2}\right).$$

We differentiate $r_N^l(z,\varphi)\sin^{2\lambda-1}\varphi$ with respect to φ up to order $2j$ and we take its value at $\varphi = 0$.

Define

$$s_{n,\lambda} = \frac{\partial^n \sin^{2\lambda-1}}{\partial\varphi^n}(0).$$

We first remark that $s_{n,\lambda}$ vanishes when n is odd or $n \le 2\lambda-3$. Indeed, since $2\lambda-1$ is even, $x \mapsto \sin^{2\lambda-1}x$ is an even function and its derivatives of odd order at $\varphi = 0$ vanish. Moreover, Faa di Bruno's formula (6.40) yields

$$\frac{\partial^n \sin^{2\lambda-1}}{\partial\varphi^n}(0) = \sum_{\gamma\in\mathcal{C}(n)} C(\gamma,n)\binom{2\lambda-1}{l(\gamma)}l(\gamma)!\sin^{2\lambda-1-l(\gamma)}(0) \times\prod_{j=1}^{l(\gamma)}\sin(\gamma_j\frac{\pi}{2}).$$

Consequently, when $n \le 2\lambda-3$, $2\lambda-1-l(\gamma)$ is positive since $l(\gamma)\le n$ and thus for all $\gamma\in\mathcal{C}(n)$, $\sin^{2\lambda-1-l(\gamma)}(0)$ vanishes.

Therefore, for $l = 0$, we infer that

$$\mathcal{J}_{N,0,\lambda}(z) = i\sqrt{\pi}\sum_{j=\lambda-1/2}^{k-1}\frac{1}{j!}\frac{1}{(2\chi_N)^{j+1/2}}s_{2j,\lambda} + O\left((N^{-1}+z^{-1})^{k+1/2}\right).$$

Consider next the case $l \ge 1$. We need first to calculate the successive even derivatives of $r_N^l(z,\varphi)$ at $\varphi = 0$. We deduce from the definition (6.37) of r_N that for j in $\{0,\cdots,2l-1\}$ and for $j \ge l(k+1)+1$, $\partial^{2j}r_N^l/\partial\varphi^{2j}(z,0)$ vanishes.

Using version (6.40) of Faa di Bruno's formula and observing that for γ in $\mathcal{C}(j)$, $r_N^{l-l(\gamma)}(z,0) = \delta_{l,l(\gamma)}$ we find that for j in $\{2l,\cdots,(k+1)l\}$:

$$\frac{\partial^{2j}r_N^l}{\partial\varphi^{2j}}(z,0) = \sum_{\substack{\gamma\in\mathcal{C}(j)\\ l(\gamma)=l}} C(\gamma,j,l)\prod_{1\le i\le l}\frac{\partial^{2\gamma_i}r_N}{\partial\varphi^{2\gamma_i}}(z,0) \tag{6.76}$$

and in virtue of definition (6.37),

$$= \sum_{\substack{\gamma\in\mathcal{C}(j)\\ l(\gamma)=l}} C(\gamma,j,l) \prod_{1\le i\le l} \frac{\partial^{2\gamma_i} g_N}{\partial\varphi^{2\gamma_i}}(z,0). \tag{6.77}$$

Thanks to equations (6.50), (6.49), (6.51) and (6.48), we obtain

$$\begin{aligned}\frac{\partial^{2\gamma_i} g_N}{\partial\varphi^{2\gamma_i}}(z,0) &= (-1)^{\gamma_i-1}\chi_N \\ &\times \sum_{p=0}^{\gamma_i-1}\binom{2\gamma_i-1}{2p}\sum_{\alpha\in\mathcal{C}(p)} C(2\alpha,2p)l(\alpha)!\left(\frac{\chi_N}{N}\right)^{l(\alpha)},\end{aligned} \tag{6.78}$$

that is to say there exists a real polynomial T_{γ_i} of degree γ_i-1 such that

$$\frac{\partial^{2\gamma_i} g_N}{\partial\varphi^{2\gamma_i}}(z,0) = \chi_N T_{\gamma_i}(\chi_N/N).$$

Therefore we deduce that for j in $\{2l,\cdots,(k+1)l\}$,

$$\frac{\partial^{2j} r_N^l}{\partial\varphi^{2j}}(z,0) = \chi_N^l S_{l,j}(\chi_N/N), \tag{6.79}$$

where $S_{l,j}$ is a real polynomial of degree $j-l$.

Hence, using Leibniz' formula, we infer for $l\ge 1$ that for j in $\{0,\cdots,k+l-1\}$,

$$\begin{aligned}\frac{\partial^{2j}\left(r_N^l(z,\varphi)\sin^{2\lambda-1}\varphi\right)}{\partial\varphi^{2j}}(z,0) &= \chi_N^l \\ &\times \sum_{m=2l}^{\min(j-\lambda+1/2,(k+1)l)}\binom{2j}{2m} s_{2j-2m,\lambda} S_{l,m}(\chi_N/N).\end{aligned}$$

Therefore, we deduce that for j in $\{0,\cdots,2l+\lambda-3/2\}$,

$$\frac{\partial^{2j}\left(r_N^l(z,\varphi)\sin^{2\lambda-1}\varphi\right)}{\partial\varphi^{2j}}(z,0) = 0 \tag{6.80}$$

and for $j\ge 2l+\lambda-1/2$

$$\frac{\partial^{2j}\left(r_N^l(z,\varphi)\sin^{2\lambda-1}\varphi\right)}{\partial\varphi^{2j}}(z,0) = \chi_N^l \widetilde{S}_{l,j,\lambda}(\chi_N/N), \tag{6.81}$$

where $\widetilde{S}_{l,j,\lambda}$ is a real polynomial of degree $j-l-\lambda+1/2$.

Eventually, formulas (6.75), (6.80) and (6.81) yield

$$\mathcal{J}_{N,l,\lambda}(z) = i\sqrt{\pi}\sum_{j=2l+\lambda-1/2}^{k+l-1}\frac{\chi_N^{l-j-1/2}}{j!\,2^{j+1/2}}\widetilde{S}_{l,j,\lambda}(\chi_N/N) + O\left((N^{-1}+z^{-1})^{k+1/2}\right)$$

and henceforth to

$$\mathcal{I}_{N,\lambda}(z,1) = i\sqrt{\pi}e^{iz}\sum_{l=0}^{k-1}\sum_{j=2l+\lambda-1/2}^{k+l-1}\frac{\chi_N^{l-j-1/2}}{l!j!2^{j+1/2}}\widetilde{S}_{l,j,\lambda}(\chi_N/N) + O\left((N^{-1}+z^{-1})^k\right), \tag{6.82}$$

where $\widetilde{S}_{0,j,\lambda}$ is a constant and for $l \geq 1$, $\widetilde{S}_{l,j,\lambda}$ is a polynomial of degree $j-l-\lambda+1/2$.

Finally, we obtain with $\nu = j-l$ in formula (6.82) that

$$P_N^{(\lambda)}(\cos(z/N)) = Z(\lambda,N)2\sqrt{\pi}\mathrm{Re}\left\{ie^{iz}\sum_{\nu=\lambda-1/2}^{k-1}\chi_N^{-(\nu+1/2)}Q_{\nu,\lambda}(\chi_N/N)\right\} + O\left(\left(N^{-1}+z^{-1}\right)^{k-2\lambda+1}\right),$$

with $Q_{\nu,\lambda}$ of degree $\nu-\lambda+1/2$, which completes the proof. □

Corollary 6.20 gives explicit values of the asymptotic for the cases $\lambda = 1/2$, $3/2$, $5/2$ and $7/2$.

Corollary 6.20. *Let $\zeta_N = ie^{iz/N}\chi_N = N\sin(z/N)$.*
For $\lambda = 1/2$ and $k = 3$, Theorem 6.19 yields

$$\begin{aligned} P_N^{(1/2)}(\cos(z/N)) &= \sqrt{\frac{2}{\pi}}\frac{1}{\zeta_N^{1/2}}\left[\cos\left(z+\frac{z}{2N}+\frac{3\pi}{4}\right)\right. \\ &\times\left(1-\frac{3}{8N}+\frac{185}{128N^2}\right)+\frac{1}{\zeta_N}\sin\left(z+\frac{3z}{2N}+\frac{3\pi}{4}\right)\left(\frac{1}{8}-\frac{55}{64N}\right) \\ &\left.-\frac{43}{384}\frac{1}{\zeta_N^2}\cos\left(z+\frac{5z}{2N}+\frac{3\pi}{4}\right)\right] \\ &+O\left((N^{-1}+z^{-1})^3\right). \end{aligned} \tag{6.83}$$

For $\lambda = 3/2$ and $k = 4$, we obtain

$$\begin{aligned} P_N^{(3/2)}(\cos(z/N)) &= -(N+2)(N+1)\sqrt{\frac{2}{\pi}}\frac{1}{\zeta_N^{3/2}}\left[\cos\left(z+\frac{3z}{2N}+\frac{\pi}{4}\right)\right. \\ &\times\left(1-\frac{15}{8N}+\frac{1505}{128N^2}\right)+\frac{1}{\zeta_N}\sin\left(z+\frac{5z}{2N}+\frac{\pi}{4}\right)\left(\frac{13}{8}-\frac{735}{64N}\right) \\ &\left.-\frac{1187}{384\zeta_N^2}\cos\left(z+\frac{7z}{2N}+\frac{\pi}{4}\right)\right]+O\left((N^{-1}+z^{-1})^2\right). \end{aligned} \tag{6.84}$$

For $\lambda = 5/2$ and $k = 4$, Theorem 6.19 implies

$$P_N^{(5/2)}(\cos(z/N)) = \frac{1}{18}\sqrt{\frac{2}{\pi}}\frac{(N+4)!}{N!}\frac{1}{\zeta_N^{5/2}}\left[\cos\left(z+\frac{5z}{2N}+\frac{3\pi}{4}\right)\right.$$
$$\left.\times\left(6-\frac{105}{4N}\right)+\frac{115}{4\zeta_N}\sin\left(z+\frac{7z}{2N}+\frac{3\pi}{4}\right)\right]+O(1). \tag{6.85}$$

For $\lambda = 7/2$ and $k = 4$, we find

$$P_N^{(7/2)}(\cos(z/N)) = -\frac{4}{15}\sqrt{\frac{2}{\pi}}\frac{(N+6)!}{N!}\frac{1}{\zeta_N^{7/2}}\cos\left(z+\frac{7z}{2N}+\frac{\pi}{4}\right)$$
$$+O(N+z)^2. \tag{6.86}$$

Proof. We follow the proof of Theorem 6.19 and we find that for $\lambda = 1/2$,

$$Q_{0,1/2}(\chi_N/N) = 1/\sqrt{2}$$

and, for $\nu \geq 1$,

$$Q_{\nu,1/2}(\chi_N/N) = \sum_{\substack{1\leq l\leq k-1\\ 2l\leq j\leq k+l-1\\ j-l=\nu}}\ \sum_{\substack{\gamma_1+\cdots+\gamma_l=j\\ \gamma_i\geq 2}}\frac{(2j)!}{j!l!}\frac{1}{2^{j+1/2}}$$
$$\times\prod_{1\leq i\leq l}\frac{1}{(2\gamma_i)!}\left(\frac{1}{\chi_N}\frac{\partial^{2\gamma_i}g_N}{\partial\varphi^{2\gamma_i}}(z,0)\right). \tag{6.87}$$

Let us calculate $Q_{1,1/2}$ and $Q_{2,1/2}$. We infer from equation (6.78) the following derivatives of g_N with respect to φ at $\varphi = 0$:

$$\frac{\partial^4 g_N}{\partial\varphi^4}(z,0) = -\chi_N\left(1+3\frac{\chi_N}{N}\right)$$

and

$$\frac{\partial^6 g_N}{\partial\varphi^6}(z,0) = \chi_N\left(1+15\frac{\chi_N}{N}+30\frac{\chi_N^2}{N^2}\right).$$

We deduce from these derivatives that

$$Q_{1,1/2}(\chi_N/N) = -\frac{1}{8\sqrt{2}}\left(1+3\frac{\chi_N}{N}\right)$$

and

$$Q_{2,1/2}(\chi_N/N) = \frac{1}{2\sqrt{2}}\left(\frac{43}{192}+\frac{55\chi_N}{32N}+\frac{185\chi_N^2}{64N^2}\right),$$

which give the asymptotic formula (6.83).

We use for $\lambda = 3/2$ the successive derivatives of the square of the sine function at $\varphi = 0$ which are

$$\frac{\partial^n \sin^2 \varphi}{\partial \varphi^n}(0) = \begin{cases} (-1)^{n/2+1} 2^{n-1} & \text{when } n \text{ is even, } n \geq 2, \\ 0 & \text{when } n \text{ is odd or } n = 0. \end{cases} \tag{6.88}$$

Therefore, we obtain

(6.89)

$$\begin{aligned} Q_{\nu,3/2}(\chi_N/N) = \frac{1}{2}\Bigg[& \frac{(-1)^{\nu+1}}{\nu!} 2^{\nu-1/2} \\ & + \sum_{\substack{1 \leq l \leq k-1 \\ 2l+1 \leq j \leq k+l-1 \\ j-l=\nu}} \sum_{\substack{2l \leq m \leq j-1 \\ \gamma_1+\cdots+\gamma_l=m \\ \gamma_i \geq 2}} \binom{2j}{2m} \frac{(-1)^{j+m+1}}{2^{2m+1/2-j}} \frac{(2m)!}{j!l!} \\ & \times \prod_{1 \leq i \leq l} \frac{1}{(2\gamma_i)!} \left(\frac{1}{\chi_N} \frac{\partial^{2\gamma_i} g_N}{\partial \varphi^{2\gamma_i}}(z,0) \right) \Bigg], \end{aligned}$$

and more precisely, we have the following values:

$$\begin{aligned} Q_{1,3/2}(\chi_N/N) &= 1/\sqrt{2}, \\ Q_{2,3/2}(\chi_N/N) &= -\frac{1}{2\sqrt{2}} \left(\frac{13}{4} + \frac{15\chi_N}{4N} \right), \end{aligned}$$

and

$$Q_{3,3/2}(\chi_N/N) = \frac{1}{2\sqrt{2}} \left(\frac{1187}{192} + \frac{735\chi_N}{32N} + \frac{1505\chi_N^2}{64N^2} \right),$$

which lead to equation (6.84).

We calculate the successive derivatives of the sine function to the power 4 at $\varphi = 0$ and we find

$$\frac{\partial^n \sin^4 \varphi}{\partial \varphi^n}(0) = \begin{cases} (-1)^{n/2}(2^{2n-3} - 2^{n-1}) & \text{when } n \text{ is even, } n \geq 4, \\ 0 & \text{when } n \text{ is odd or} \\ & \quad n = 0,\ n = 2. \end{cases}$$

These derivatives enable us to calculate:

$$Q_{2,5/2}(\chi_N/N) = 6\sqrt{2},$$

and

$$Q_{3,5/2}(\chi_N/N) = -\sqrt{2} \left(\frac{115}{4} + \frac{105}{4} \frac{\chi_N}{N} \right)$$

and this yields formula (6.85).

Eventually, the successive derivatives of the sine function to the power 6 at $\varphi = 0$ are

$$\frac{\partial^n \sin^6 \varphi}{\partial \varphi^n}(0) = \begin{cases} (-1)^{n/2+1} 2^{n-5} 3(3^{n-1} - 2^{n+1} + 5) \\ \qquad \text{when } n \text{ is even, } n \geq 6, \\ 0 \text{ when } n \text{ is odd or } n = 0,\ n = 2,\ n = 4. \end{cases}$$

Therefore we find that

$$Q_{3,7/2}(\chi_N/N) = 30\sqrt{2}$$

and the calculation of formula (6.86) completes the proof. □

6.3.4. *Asymptotics of the zeroes of the Legendre polynomials.* Now that the formulas for L_N and its derivatives have been computed in Corollary 6.20, we can find an asymptotic formula for the zeroes of L'_N in the region $K \leq k \leq \Lambda N$.

THEOREM 6.21. *Define*

$$z_{0,k} = \frac{\pi/4 + k\pi}{1 + 3/2N}.$$

Then for all Λ in $(0, 1/2)$ and for all $K \in \mathbb{N}$, there exist C, C' such that for all $N \geq 2$ and for all integer k in $\{K, \cdots, \lfloor \Lambda N \rfloor\}$, there exists a unique zero z_k of $P_N^{(3/2)}(\cos(z/N))$ in a ball of radius C'/N about $z_{0,k}$ and moreover the following estimate holds

(6.90)
$$\left| z_k - z_{0,k} - \frac{13}{8N\tan(z_{0,k}/N)} + \frac{22}{3N^2\tan(z_{0,k}/N)} \right| \leq C(N^{-1} + K^{-1})^3.$$

PROOF. We use the same method as in the proof of Theorem 6.3 and we use again Lemma 6.4 to calculate an asymptotic formula for the zero z_k of $P_N^{(3/2)}(\cos(z/N))$; this function is given by formula (6.84) of Corollary 6.20.

It is equivalent to calculate the zero z_k of

$$f(z, N) = -\sqrt{\frac{\pi}{2}} \frac{(N\sin(z/N))^{3/2}}{(N+2)(N+1)} P_N^{(3/2)}(\cos(z/N)). \tag{6.91}$$

We are searching this zero in the neighborhood of

$$z_{0,k} = \frac{\pi/4 + k\pi}{1 + 3/2N}.$$

We calculate $f(z_{0,k}, N)$ thanks to formula (6.84) of Corollary 6.20 and we obtain

(6.92)
$$f(z_{0,k}, N) = \frac{(-1)^k}{\tan(z_{0,k}/N)} \left(\frac{13}{8N} - \frac{509}{96N^2} \right) + O\left((N^{-1} + K^{-1})^{5/2} \right).$$

We differentiate formula (6.91) to obtain:

$$\begin{aligned}\frac{\partial f}{\partial z}(z,N) &= \frac{3}{2N}\frac{f(z,N)}{\tan(z/N)} \\ &\quad + 3\sqrt{\frac{\pi}{2}}\frac{\sqrt{N}}{(N+2)(N+1)}\sin^{5/2}(z/N)P_{N-1}^{(5/2)}(\cos(z/N))\end{aligned} \tag{6.93}$$

and using formula (6.92) and equation (6.85) of Corollary 6.20, we find

$$A(k,N) = \frac{\partial f}{\partial z}(z_{0,k},N) = (-1)^{k-1} + O(N^{-1}+K^{-1}).$$

We calculate now the derivative of second order of the function f using formula (6.93):

$$\begin{aligned}\frac{\partial^2 f}{\partial z^2}(z,N) &= \frac{3}{4N^2}\left(\frac{1}{\tan^2(z/N)} - 2\right) f(z,N) \\ &\quad + 12\sqrt{\frac{\pi}{2N}}\frac{N!}{(N+2)!}\cos(z/N)\sin^{3/2}(z/N)P_{N-1}^{(5/2)}(\cos(z/N)) \\ &\quad - 15\sqrt{\frac{\pi}{2N}}\frac{N!}{(N+2)!}\sin^{7/2}(z/N)P_{N-2}^{(7/2)}(\cos(z/N)).\end{aligned}$$

Let C be a positive real such that $|A^{-1}(k,N)f(z_{0,k},N)| \le CN^{-1}$. Let z belong to the ball of center $z_{0,k}$ and of radius $2CN^{-1}$. We still use formula (6.92) and equations (6.85) and (6.86) of Corollary 6.20 to compute

$$\frac{\partial^2 f}{\partial z^2}(z,N) = O(N^{-1}+K^{-1}).$$

Therefore the number M of Lemma 6.4 is finite and the number a is equal to 1. The radius $2CN^{-2}$ has been chosen so that the hypothesis $|A^{-1}(k,N)f(z_{0,k},N)| \le K$ is satisfied. Therefore, we have the following asymptotic formula:

$$z_k = z_{0,k} + \frac{13}{8N\tan(z_{0,k}/N)} + O\left((N^{-1}+K^{-1})^2\right).$$

In order to have a more precise asymptotic formula, we use once more Lemma 6.4 with the same function f but in the neighborhood of

$$z_{1,k} = z_{0,k} + \frac{13}{8N\tan(z_{0,k}/N)}.$$

We compute the values of f and its derivatives at $z = z_{1,k}$ and we find

$$\begin{aligned}f(z_{1,k},N) &= (-1)^{k+1}\frac{22}{3N^2\tan(z_{0,k}/N)} + O\left((N^{-1}+K^{-1})^3\right), \\ \frac{\partial f}{\partial z}(z_{1,k},N) &= (-1)^{k+1} + O(N^{-1}+K^{-1})\end{aligned}$$

and if z belongs to the ball of center $z_{1,k}$ and of radius $2CN^{-1}$, the following estimate holds:

$$\frac{\partial^2 f}{\partial z^2}(z,N) = O(N^{-1} + K^{-1}).$$

Eventually, we obtain the following asymptotic formula

$$z_k = z_{0,k} + \frac{13}{8N\tan(z_{0,k}/N)} - \frac{22}{3N^2\tan(z_{0,k}/N)} + O\left((N^{-1}+K^{-1})^3\right).$$

□

We then have the straightforward corollary:

Corollary 6.22. *Define*

$$\theta_{0,k} = \frac{\pi(N-k+1/4)}{N+1/2}.$$

Then for all $K > 0$ and for all $\Lambda \in (0,1/2)$, there exist C, C' such that for all $N \geq 2$ and for all integer k in $\{K, \cdots, \lfloor \Lambda N \rfloor\}$, there exists a unique zero θ_k of $L'_N(\cos\theta)$ in a ball of radius C'/N^2 about $\theta_{0,k}$; moreover the following estimate holds

(6.94)
$$\left|\theta_k - \theta_{0,k} - \frac{13}{8N^2\tan\theta_{0,k}} + \frac{49}{12N^3\tan\theta_{0,k}}\right| \leq C\left((N^{-1}+K^{-1})^4\right).$$

Remark 6.23. *Observe that* (6.94) *is compatible with* (6.12), *because the error term in* (6.94) *is large with respect to the error term in* (6.12).

CHAPITRE 2

Approximation de Peaceman – Rachford

1. Rappels sur les opérateurs sectoriels et leur lien avec les semi-groupes analytiques

Dans les estimations de la section 2.3 de ce chapitre et dans la démonstration du Lemme 2.8 de la section 2.4, nous utilisons le fait qu'un opérateur est sectoriel et le fait qu'il engendre un semi-groupe analytique. Nous faisons donc ici un rappel de ces notions.

1.1. Opérateurs sectoriels. Rappelons tout d'abord la définition d'un opérateur sectoriel donnée par D. Henry [**65**] :

Définition 1.1. *Un opérateur linéaire A sur un espace de Banach X est appelé sectoriel si A est un opérateur fermé à domaine dense tel qu'il existe un réel ϕ dans $]0, \pi/2[$, un réel $M \geq 1$ et un réel a tels que le secteur*

$$S_{a,\phi} = \{\lambda : \phi \leq |\arg(\lambda - a)| \leq \pi, \lambda \neq a\}$$

soit inclus dans l'ensemble résolvant de A et tels que

$$\left\|(\lambda - A)^{-1}\right\| \leq \frac{M}{|\lambda - a|} \text{ pour tout } \lambda \in S_{a,\phi}.$$

Un exemple de secteur $S_{a,\phi}$ est représenté à la figure 1.

Par exemple, un opérateur linéaire borné sur X est sectoriel ; un opérateur auto – adjoint à domaine dense borné inférieurement est également sectoriel.

Le Théorème 1.3.2 du livre de D. Henry [**65**] que l'on trouve également dans le livre de A. Pazy [**93**], Théorème 2.1 de la section 3.2, montre que la perturbation d'un opérateur sectoriel reste sectorielle sous certaines conditions sur la perturbation :

Théorème 1.2. *Soit A un opérateur sectoriel et soit B un opérateur linéaire fermé tel que $D(B) \supset D(A)$ et tel que*

$$\|Bx\| \leq c\,\|Ax\| + b\,\|x\| \text{ pour } x \in D(A).$$

Alors, il existe un nombre strictement positif δ tel que si $0 \leq c \leq \delta$, $A + B$ est sectoriel.

Le corollaire de ce théorème est que si A est un opérateur sectoriel et B est un opérateur borné, alors la somme $A + B$ est un opérateur sectoriel. On utilisera ce résultat à la section 2.3 de l'article.

1.2. Semi-groupes analytiques et lien avec les opérateurs sectoriels. Rappelons maintenant la définition d'un semi-groupe analytique :

DÉFINITION 1.3. *Un semi-groupe analytique sur un espace de Banach X est une famille d'opérateurs linéaires continus sur X, $\{T(t)\}_{t\geq 0}$ telle que*

– $T(0) = \mathbf{1}$, $T(t)T(s) = T(t+s)$ *pour tout* $t \geq 0$, $s \geq 0$,

– $T(t)x \to x$ *quand* $t \to 0^+$ *pour tout* $x \in X$,

– $t \to T(t)x$ *est réel et analytique sur* $0 < t < \infty$ *pour tout* $x \in X$.

Un générateur de ce semi-groupe est alors défini par

$$Lx = \lim_{t\to 0^+} \frac{1}{t}(T(t)x - x)$$

et le domaine de L est composé des éléments de X tels que cette limite existe.

Nous avons alors le résultat suivant du livre de D. Henry [**65**] qui fait le lien entre opérateurs sectoriels et semi-groupes analytiques :

THÉORÈME 1.4. *Si A est sectoriel, alors $-A$ est le générateur infinitésimal d'un semi-groupe analytique $\{e^{-tA}\}_{t\geq 0}$ défini par*

$$e^{-tA} = \frac{1}{2\pi i}\int_\Gamma (\lambda + A)^{-1} e^{\lambda t} d\lambda, \tag{1.1}$$

où Γ est un contour dans l'ensemble résolvant de $-A$ tel qu'il existe θ dans $]\pi/2, \pi[$ tel que $\arg\lambda \to \pm\theta$ quand $|\lambda| \to \infty$.

De plus, e^{-tA} peut être prolongé analytiquement sur un secteur $\{t \neq 0 : |\arg t| < \varepsilon\}$ qui contient l'axe réel positif et si pour tout λ appartenant au spectre de A, $\operatorname{Re}\lambda > a$, alors pour tout $t > 0$,

$$\left\|e^{-tA}\right\| \leq Ce^{-at},\ \left\|Ae^{-tA}\right\| \leq \frac{C}{t}e^{-at}.$$

Finalement, on a également pour $t > 0$,

$$\frac{d}{dt}e^{-tA} = -Ae^{-tA}.$$

Un exemple de contour Γ est représenté à la figure 1.

On remarque que la réciproque de ce théorème est également vraie : si $-A$ engendre un semi-groupe analytique, alors A est sectoriel.

Lors de la démonstration du théorème précédent, on montre l'estimation suivante :

$$\left\|\mathrm{e}^{-tA}x - x\right\| \leq Ct\left\|Ax\right\|,$$

dont nous allons nous servir plus loin à la section 2.4 de l'article.

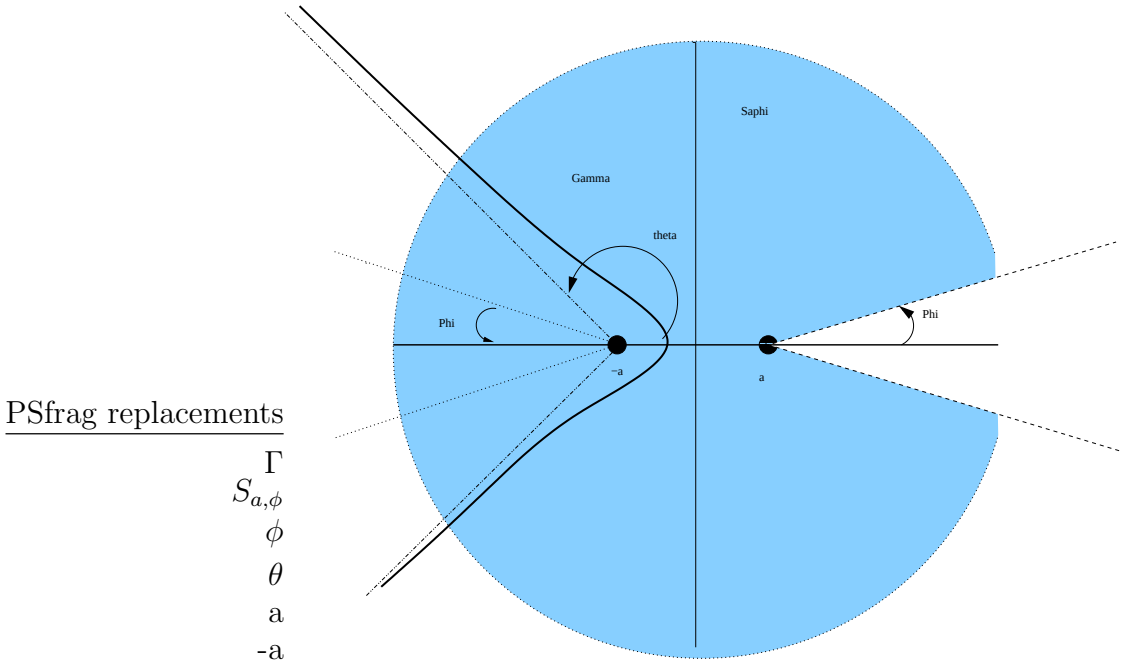


FIG. 1. Un exemple de secteur $S_{a,\phi}$ de la définition 1.1 et de contour Γ du Théorème 1.4.

2. Convergence du schéma de Peaceman-Rachford pour des systèmes de réaction – diffusion

Cet article est publié dans le journal Numerische Mathematik en 2003.

Convergence of the Peaceman-Rachford approximation for reaction-diffusion systems

Stéphane Descombes and Magali Ribot [1]

Abstract : *Consider a reaction-diffusion system of the form* $u_t - M\Delta u + F(u) = 0$, *where* M *is a* $m \times m$ *matrix whose spectrum is included in* $\{\mathrm{Re} z > 0\}$. *We approximate its solution by the Peaceman-Rachford approximation defined by*

$$P(t) = (\mathbf{1} + tF/2)^{-1}(\mathbf{1} + tM\Delta/2)(\mathbf{1} - tM\Delta/2)^{-1}(\mathbf{1} - tF/2).$$

We prove convergence of this scheme and show that it is of order two.

2.1. Introduction. Reaction-diffusion systems appear in many different situations, in chemistry, in biology, and we are specifically interested in pattern formation. For example, in [**86**] page 438, J. D. Murray has given the non dimensional system, with parameters $a > 0$, $b > 0$, $\alpha > 0$, $\gamma > 0$, $\rho > 0$, $K > 0$ and $d > 1$,

$$\begin{cases} \dfrac{\partial u}{\partial t} = \Delta u + \gamma \left(a - u - \dfrac{\rho u v}{1 + u + K u^2} \right), \\ \dfrac{\partial v}{\partial t} = d\Delta v + \gamma \left(\alpha(b - v) - \dfrac{\rho u v}{1 + u + K u^2} \right), \end{cases} \tag{2.1}$$

as a mathematical model for pattern formation. This system could be included in reaction-diffusion systems of the general form with d, m two integers, $1 \leq d \leq 3$ and M an $m \times m$ matrix:

$$\begin{cases} \dfrac{\partial u}{\partial t} - M\Delta u + F(u) = 0, & x \in \mathbb{R}^d,\ t > 0, \\ u(0, x) = u_0(x), & x \in \mathbb{R}^d. \end{cases} \tag{2.2}$$

The simplest numerical scheme for solving this kind of system is the forward Euler scheme, but in that case, the time step Δt is limited by $O(\Delta x^2)$, a drastic condition. Another strategy is to use the backward Euler scheme; while its stability is unconditional, one has to solve a large system of nonlinear equations. These two examples show that we need to consider other methods. A possibility for integrating (2.2) is the implicit-explicit multistep methods studied by G. Akrivis, M. Crouzeix and C. Makridakis in [**1**]. They use two multistep formulae:

[1]We would like to thank Prof. Michelle Schatzman for reading carefully the manuscript.

an implicit one for the linear part and an explicit one for the non linear part. However, they are limited by the condition M self – adjoint.

Approximations based on flow splitting are often proposed; they are defined as follows: let $X^t v_0$ be the solution of

$$\begin{cases} \dot{v} - M\Delta v = 0 \\ v(0) = v_0 \end{cases}$$

and $Y^t w_0$ be the solution of

$$\begin{cases} \dot{w} + F(w) = 0 \\ w(0) = w_0; \end{cases}$$

then the Lie formula is defined by

$$L^t u_0 = X^t Y^t u_0$$

and is of order one (at least formally). The Strang approximation formula ([**116, 117**]) is defined by

$$S^t u_0 = X^{t/2} Y^t X^{t/2} u_0$$

and is of order two (at least formally). We then have to discretize these approximations in time and in space to solve (2.2). The approximation in time could be defined, for example, by this formula:

$$Q^t u_0 = (\mathbf{1} + tF/2)^{-1} (\mathbf{1} - tM\Delta/2)^{-1} u_0$$

which is of order one; this scheme can be studied as in [**42**], where C. Chiu and N. Walkington study a splitting scheme for solving an hysteretic reaction-diffusion system.

In this article we consider only the time discretization and we concentrate on the Peaceman-Rachford formula given by:

(2.3)
$$P^t u_0 = (\mathbf{1} + tF/2)^{-1} (\mathbf{1} + tM\Delta/2) (\mathbf{1} - tM\Delta/2)^{-1} (\mathbf{1} - tF/2) u_0.$$

Originally introduced in [**94**] to solve the heat equation (see also [**58**]), and also called ADI (Alternate Directions Implicit), this method is seldom analyzed in non commutative and/or nonlinear cases. In the linear case, M. Schatzman has given in [**107**] a sufficient condition for the stability of

$$P(t) = (\mathbf{1} + tA/2)^{-1} (\mathbf{1} - tB/2) (\mathbf{1} + tB/2)^{-1} (\mathbf{1} - tA/2),$$

when A and B are unbounded positive linear self adjoint operators and the commutator of $\sqrt{A}$ and $\sqrt{B}$ is dominated in a very precise way by $\sqrt{A} + \sqrt{B}$, making A and B abstractly of order 2. This situation is realized if for instance x and y are periodic variables, a and b are positive smooth functions of x and y on $\mathbb{R}^2/\mathbb{Z}^2$ and

$$A = -\frac{\partial}{\partial x} a \frac{\partial}{\partial x}, \quad B = -\frac{\partial}{\partial y} b \frac{\partial}{\partial y}.$$

B. O. Dia in [**56**] has shown some numerical results for this kind of operators which prove a good behavior of this scheme.

Let us now introduce the assumptions on the problem (2.2) and on the solution u that we wish to approximate. We assume that F is a Lipschitz continuous function of class $\mathrm{C}^{\mathbf{7}}$ from $\mathbb{R}^m$ to itself satisfying

$$F(0) = 0, \tag{2.4}$$

and that all its derivatives of order at most 7 are bounded. Observe that (2.4) is necessary for the existence of solutions in $L^2(\mathbb{R}^d)^m$. M is an $m \times m$ matrix whose spectrum is included in $\{\mathrm{Re} z > 0\}$. We assume that the initial condition u_0 belongs to $L^2(\mathbb{R}^d)^m \cap L^\infty(\mathbb{R}^d)^m$; then u exists, is unique and satisfies, for all $\tau > 0$,

$$u \in \mathrm{C}([0,\tau], L^2(\mathbb{R}^d)^m \cap L^\infty(\mathbb{R}^d)^m).$$

This assertion is proved in D. Henry [**65**], section 3.3 p. 52 when u_0 belongs only to $L^2(\mathbb{R}^d)^m$, the extension when u_0 belongs to $L^2(\mathbb{R}^d)^m \cap L^\infty(\mathbb{R}^d)^m$ is straightforward using the results of the first section of P. de Mottoni and M. Schatzman [**46**].

We write $u(t,.) = T^t u_0$, that is to say T^t is the flow associated to (2.2). For simplicity, we shall write henceforth

$$L(t) = (\mathbf{1} + tM\Delta)(\mathbf{1} - tM\Delta)^{-1}. \tag{2.5}$$

We introduce now some functional spaces; we need an adapted scalar product for some of our operators to be contractions. As the spectrum of M is included in $\{\mathrm{Re} z > 0\}$, the matrix

$$S = \int_0^{+\infty} e^{-sM^*} e^{-sM} ds \tag{2.6}$$

is well defined, symmetric, positive definite and satisfies

$$M^* S + SM = \mathbf{1}, \tag{2.7}$$

as can be verified by a simple integration by parts. The corresponding inner product in $\mathbb{R}^m$ is

$$\langle x, y\rangle = (Sx, y)$$

and the corresponding norm is $\big|x\big| = \langle x, x\rangle^{1/2}$. The space $H = L^2(\mathbb{R}^d)^m$ is equipped with the inner product defined by

$$(f,g)_0 = \int_{\mathbb{R}^d} \langle f, g\rangle \, dx,$$

the associated norm is denoted by $|\cdot|_0$ and the corresponding operator norm by $|\cdot|_{\mathcal{L}(L^2)}$. We also introduce $H^s(\mathbb{R}^d)^m$, the Sobolev space of order s, equipped with the norm $|\cdot|_s$ defined by

$$|f|_s^2 = \sum_{|\alpha| \le s} |\partial^\alpha f|_0^2;$$

$L^\infty(\mathbb{R}^d)^m$, equipped with the norm $|\cdot|_\infty$ and $W^{s,\infty}(\mathbb{R}^d)^m$, equipped with the norm $|\cdot|_{s,\infty}$. Finally, $\mathcal{G}$ is the subspace of $H^6(\mathbb{R}^d)^m$ made out

of functions which belong to $C^6(\mathbb{R}^d)^m$ whose first sixth derivatives are bounded and, for u_0 belonging to $\mathcal{G}$, we denote by $C(|u_0|_{6,\infty})$ a constant depending on $\max_{0\le|\alpha|\le 6}|\partial^\alpha u_0|_\infty$.

The main result of the article is Theorem 2.14 which can be stated as follows:

THEOREM 2.14. *For all u_0 in $\mathcal{G}$ and for all $\tau > 0$, there exist $C(|u_0|_{6,\infty})$ and h_0 such that for all $h \in (0, h_0]$, for all n such that $nh \le \tau$*

$$\left|\left(P^h\right)^n u_0 - T^{nh}u_0\right|_0 \le C(|u_0|_{6,\infty})h^2|u_0|_6.$$

The article is organized as follows: in Section 2.2, we consider the linear case, and we obtain an algebraic expression of the difference $e^{-2t(A+B)} - P(2t)$. The method follows an idea of [**57**].

In Section 2.3, we validate from the analytical point of view the result of Section 2.2 when A is the multiplication by a bounded potential V and $B = -M\Delta$ and deduce an estimate on the difference between $e^{t(M\Delta - V)}$ and $P(t)$ in operator norms.

In Section 2.4, we prove Theorem 2.14 with the help of a comparison with the linear case.

In Section 2.5, we show some numerical results and explain how we approximate the operator $(\mathbf{1} + tF/2)^{-1}$ without solving a nonlinear system.

2.2. Error formula in the linear case. In this section, as in [**52**], we perform algebraic computations when A and B are two linear operators. We are interested in the difference between $P(2t) = (\mathbf{1}+tA)^{-1}(\mathbf{1}-tB)(\mathbf{1}+tB)^{-1}(\mathbf{1}-tA)$, denoted $P_2(t)$, and $E(2t) = e^{-2t(A+B)}$.

We introduce the following notation: $[A, B] = AB - BA$ is the commutator of A and B. We recall the identity:

$$[A, BC] = [A, B]C + B[A, C].$$

We also recall Duhamel's formula: if V satisfies

$$\dot{V} + AV = f,$$

then

$$V(t) = e^{-tA}\,V(0) + \int_0^t e^{-(t-s)A}\,f(s)\,ds. \tag{2.8}$$

Finally, we write $\alpha = (\mathbf{1} + tA)^{-1}$ and $\beta = (\mathbf{1} + tB)^{-1}$ and we recall the identities:

$$(\mathbf{1} - tA)\alpha = \alpha(\mathbf{1} - tA) = 2\alpha - \mathbf{1}, \tag{2.9}$$

$$\mathbf{1} - \alpha = t\alpha A, \tag{2.10}$$

$$[B, \alpha] = t\alpha[A, B]\alpha. \tag{2.11}$$

Analogous identities hold when we exchange A and B, and α and β. With these notations, we see that $P_2(t) = P(2t) = \alpha(1-tB)\beta(1-tA)$.

LEMMA 2.1. *Define*

(2.12)
$$\begin{aligned} S(t) =& \big(\alpha\beta\big[B,[A,B]\big]\beta^2 - \alpha[A,B]\alpha(A\beta^2+\beta^2A) - \alpha^2A^3(2\beta-\mathbf{1}) \\ &+ A\alpha(\alpha-2)\beta[A,B]\beta - B\alpha B^2\beta^2(\mathbf{1}-tA)\big). \end{aligned}$$

We have the following identity:

$$P(2t) - E(2t) = 2\int_0^t s^2 e^{-2(t-s)(A+B)} S(s)\,ds. \tag{2.13}$$

PROOF. Relation (2.9), or rather its analogue for B and β, lets us rewrite

$$P_2(t) = \alpha(2\beta-\mathbf{1})(\mathbf{1}-tA).$$

In order to use Duhamel's formula, we differentiate $P_2(t)$, we add $2(A+B)P_2(t)$ and we obtain

(2.14)
$$\begin{aligned} \dot{P}_2(t) + 2(A+B)P_2(t) =& -A\alpha^2(2\beta-\mathbf{1})(\mathbf{1}-tA) - 2\alpha B\beta^2(\mathbf{1}-tA) \\ &- \alpha(2\beta-\mathbf{1})A + 2A\alpha(2\beta-\mathbf{1})(\mathbf{1}-tA) \\ &+ 2B\alpha(2\beta-\mathbf{1})(\mathbf{1}-tA). \end{aligned}$$

The game in this computation is to rearrange the right hand side of (2.14), in order to show that it is the product of t^2 by a term which we will control. But we have to be careful, since limited expansions are not permissible; and therefore we drag big remainder terms coming from the rational nature of the approximation.

We regroup the terms of (2.14) by separating those which give a multiple of A for $t=0$,

$$\begin{aligned} Q_1(t) =& -A\alpha^2(2\beta-\mathbf{1})(\mathbf{1}-tA) - \alpha(2\beta-\mathbf{1})A \\ &+ 2A\alpha(2\beta-\mathbf{1})(\mathbf{1}-tA) \end{aligned} \tag{2.15}$$

and those which give a multiple of B for $t=0$

$$Q_2(t) = -2\alpha B\beta^2(\mathbf{1}-tA) + 2B\alpha(2\beta-\mathbf{1})(\mathbf{1}-tA),$$

so that

$$\dot{P}_2(t) + 2(A+B)P_2(t) = Q_1(t) + Q_2(t).$$

Since $Q_1(0)$ vanishes, we will try to factor t in $Q_1(t)$; for this purpose, we commute A and β in the second term of $Q_1(t)$, obtaining thus

$$\alpha(2\beta-\mathbf{1})A = \alpha A(2\beta-\mathbf{1}) + 2\alpha[\beta,A]; \tag{2.16}$$

and the commutator $[\beta,A]$ is equal to $t\beta[A,B]\beta$ using formula (2.11). We substitute (2.16) into (2.15),getting

$$\begin{aligned} Q_1(t) =& -A\alpha^2(2\beta-\mathbf{1})(\mathbf{1}-tA) + 2t\alpha\beta[B,A]\beta - \alpha A(2\beta-\mathbf{1}) \\ &+ 2A\alpha(2\beta-\mathbf{1})(\mathbf{1}-tA). \end{aligned}$$

We regroup the terms of degree 0 in t on one hand and the terms of degree 1 in t on the other to obtain, after simple factorizations,

$$Q_1(t) = (1-\alpha)\alpha A(2\beta - \mathbf{1}) + tA\alpha(\alpha-2)(2\beta-\mathbf{1})A + 2t\alpha\beta[B,A]\beta.$$

The interesting term is the term of degree 0 and using equation (2.10), we obtain

$$Q_1(t) = t\alpha^2 A^2(2\beta-\mathbf{1}) + tA\alpha(\alpha-2)(2\beta-\mathbf{1})A + 2t\alpha\beta[B,A]\beta. \tag{2.17}$$

Let us turn to $Q_2(t)$.

The first term of $Q_2(t)$ can be rewritten after commuting B and α

$$-2\alpha B\beta^2(\mathbf{1}-tA) = 2[B,\alpha]\beta^2(\mathbf{1}-tA) - 2B\alpha\beta^2(\mathbf{1}-tA),$$

and using formula (2.11),

$$= 2t\alpha[A,B]\alpha\beta^2(\mathbf{1}-tA) - 2B\alpha\beta^2(\mathbf{1}-tA).$$

We regroup the second term of the above expression with the last term of $Q_2(t)$, and the sum simplifies as follows:

$$Q_2(t) = 2t\alpha[A,B]\alpha\beta^2(\mathbf{1}-tA) - 2B\alpha(\beta-\mathbf{1})^2(\mathbf{1}-tA);$$

using formula (2.10), we obtain

$$Q_2(t) = 2t\alpha[A,B]\alpha\beta^2(\mathbf{1}-tA) - 2t^2B\alpha B^2\beta^2(\mathbf{1}-tA). \tag{2.18}$$

Finally, thanks to computations (2.17) and (2.18) we find that

$$\begin{aligned}\dot{P}_2(t) + 2(A+B)P_2(t) = t\Big(&2\alpha\beta[B,A]\beta + \alpha^2A^2(2\beta-\mathbf{1})\\ &+ A\alpha(\alpha-2)(2\beta-\mathbf{1})A\\ &+ 2\alpha[A,B]\alpha\beta^2(\mathbf{1}-tA)\\ &- 2tB\alpha B^2\beta^2(\mathbf{1}-tA)\Big).\end{aligned}$$

The factor of t in the above expression vanishes at 0; we rewrite this factor as a sum of two expressions and one term: the terms of the first expression contain $[A,B]$, and the terms of the second expression do not, and the last term has already the factor t. Thus we define

$$R_1(t) = 2\alpha\beta[B,A]\beta + 2\alpha[A,B]\alpha\beta^2(\mathbf{1}-tA)$$

and

$$R_2(t) = \alpha^2A^2(2\beta-\mathbf{1}) + A\alpha(\alpha-2)(2\beta-\mathbf{1})A.$$

Then

$$\dot{P}_2(t) + 2(A+B)P_2(t) = t\left(R_1(t) + R_2(t) - 2tB\alpha B^2\beta^2(\mathbf{1}-tA)\right).$$

We compute $R_1(t)$: in the same fashion as above, in the first term of R_1, we commute β and $[B,A]$, we regroup the terms with a t factor on one hand, and those without on the other hand. Among the terms without a factor t, there is a $[\beta,[B,A]]$, that we rewrite $t\beta\,[[B,A],B]\,\beta$,

yielding thus a factor t; for the last term without t, we use (2.10), obtaining thus

$$R_1(t) = 2t\left(\alpha\beta\,[B,[A,B]]\,\beta^2 - \alpha[A,B]\alpha(A\beta^2+\beta^2A)\right).$$

We proceed analogously for $R_2(t)$; this time the commutation takes place in the second term and we obtain:

$$R_2(t) = 2t\left(-\alpha^2A^3(2\beta-\mathbf{1}) + A\alpha(\alpha-2)\beta[A,B]\beta\right).$$

Eventually, we find

$$\begin{aligned}\dot{P}_2(t) + 2(A+B)P_2(t) = 2t^2\big(&\alpha\beta\big[B,[A,B]\big]\beta^2\\ &- \alpha[A,B]\alpha(A\beta^2+\beta^2A) - \alpha^2A^3(2\beta-\mathbf{1})\\ &+ A\alpha(\alpha-2)\beta[A,B]\beta - B\alpha B^2\beta^2(\mathbf{1}-tA)\big),\end{aligned}$$

and the proof of (2.13) is an immediate consequence of Duhamel's formula (2.8). □

Remark 2.2. *If A and B are bounded, formula* (2.13) *shows that the approximation is exactly of order two.*

2.3. Estimates on Matrix Schrödinger operators. Let now A be the multiplication by an $m\times m$ matrix V depending on $x\in\mathbb{R}^d$, and let B be $-M\Delta$; we assume that V is of class C^6 and that it is bounded as well as all its derivatives of order at most 6. We will estimate in that case the difference between $P(t)$ and $E(t)$ in operator norm, using the expression found in section 2.2. Thus, we calculate $[A,B]$ in that case and we have the following very simple lemma, whose proof is left to the reader.

Lemma 2.3. *There exist bounded functions L_0, $L_{1,i}$, $1\le i\le d$, and L_2 depending on the first two derivatives of V and on M such that*

$$[A,B] = [M\Delta,V] = L_0 + \sum_{i=1}^d L_{1,i}\partial_i + L_2\Delta; \tag{2.19}$$

there exist also bounded functions N_0, $N_{1,i}$, $1\le i\le d$, N_2, $N_{3,i}$, $1\le i\le d$, and N_4 depending on the first four derivatives of V and on M, such that

$$\begin{aligned}[[A,B],B] &= [M\Delta,[M\Delta,V]]\\ &= N_0 + \sum_{i=1}^d N_{1,i}\partial_i + N_2\Delta + \sum_{i=1}^d N_{3,i}\partial_i\Delta + N_4\Delta^2.\end{aligned} \tag{2.20}$$

Now, let us estimate $\|P(t)-E(t)\|$. We recall the following estimates, that we will need throughout this section: since A is a bounded operator, there exists a constant $C>0$ such that, for $s=0,\cdots,6$

$$\|A\|_{\mathcal{L}(H^s)} \le C \tag{2.21}$$

and

$$\|\alpha\|_{\mathcal{L}(H^s)} \leq 1,\ \|\beta\|_{\mathcal{L}(H^s)} \leq 1; \tag{2.22}$$

by definition, for $s = 0, \cdots, 4$,

$$\|B\|_{\mathcal{L}(H^{s+2},H^s)} \leq C. \tag{2.23}$$

Finally, it has been shown in [**49**] that the operator B is sectorial in $L^2(\mathbb{R}^d)^m$, and since A is a bounded operator, there exists a constant $C > 0$ such that, for $t \leq 1$,

$$\left\|e^{-t(A+B)}\right\|_{\mathcal{L}(L^2)} \leq C. \tag{2.24}$$

LEMMA 2.4. *We have the following estimates for $t \in [0,1]$:*

$$\|P(t) - E(t)\|_{\mathcal{L}(H^2,L^2)} = O\left(t\right), \tag{2.25}$$

$$\|P(t) - E(t)\|_{\mathcal{L}(H^4,L^2)} = O\left(t^2\right), \tag{2.26}$$

and

$$\|P(t) - E(t)\|_{\mathcal{L}(H^6,L^2)} = O\left(t^3\right). \tag{2.27}$$

PROOF. We decompose $s^2S(s)$ as $T_1 + T_2 + T_3 + T_4 + T_5$ with

$$\begin{aligned}
T_1 &= s^2\alpha\beta\left[B,[A,B]\right]\beta^2,\\
T_2 &= -s^2\alpha[A,B]\alpha(A\beta^2 + \beta^2 A),\\
T_3 &= -s^2\alpha^2A^3(2\beta - \mathbf{1}),\\
T_4 &= s^2A\alpha(\alpha-2)\beta[A,B]\beta,\\
T_5 &= -s^2B\alpha B^2\beta^2(\mathbf{1} - sA).
\end{aligned}$$

Then, we can rewrite the integrand in (2.13) as

$$T = e^{-2(t-s)(A+B)}(T_1 + T_2 + T_3 + T_4 + T_5).$$

We now estimate these five terms, using estimates (2.21), (2.23), (2.22) and (2.24). The largest of these terms is T_5 since it involves six differentiations.

First, for the first term, using equation (2.20), we find for $u \in H^6$,

$$|T_1u|_{L^2} \leq C_1s^2|u|_6.$$

Consider now T_2; relation (2.19), together with (2.21), implies, if $u \in H^6$,

$$|T_2u|_{L^2} \leq C_2s^2|u|_6.$$

The term T_3u is trivially estimated for $u \in H^6$ by

$$|T_3u|_{L^2} \leq C_3s^2|u|_6.$$

The fourth term is treated almost as T_2, using (2.19) and (2.21), so that if $u \in H^6$,

$$|T_4u|_{L^2} \leq C_4s^2|u|_6.$$

Finally, for the fifth term, if $u \in H^6$,

$$|T_5 u|_{L^2} \leq C_5 s^2 |u|_6.$$

In conclusion, adding all those estimates, we find that, if $u \in H^6$,

$$|Tu|_{L^2} \leq C s^2 |u|_6.$$

Integrating the previous estimate, we obtain equation (2.27).

Moreover, if $u \in L^2$, we have thanks to estimates (2.21), (2.22) and (2.24),

$$\|P(t) - E(t)\|_{\mathcal{L}(L^2,L^2)} = O(1). \tag{2.28}$$

Thus, from equations (2.27) and (2.28) and with the help of interpolation theory, see for example J. Bergh and J. Löfström [**8**], we obtain for all $s \in [0,6]$,

$$\|P(t) - E(t)\|_{\mathcal{L}(H^s,L^2)} = O\left(t^{s/2}\right),$$

and especially equations (2.25) and (2.26). □

2.4. The nonlinear case. In this section, we consider the nonlinear case and we compare the exact solution $T^t u_0$ to its approximation $P^t u_0$.

2.4.1. *Some preliminary and useful lemmas.* We recall first a Gronwall type lemma proved in [**10**].

LEMMA 2.5 (Gronwall). *Let Q be a polynomial with positive coefficients satisfying $Q(0) = 0$. We assume that function ϕ is such that there exists a constant $C \geq 0$ such that for all $t \geq 0$*

$$0 \leq \phi(t) \leq \phi(0) + Q(t) + C \int_0^t \phi(s) ds.$$

Then for all $\alpha > 1$ there exists $t_0 > 0$ such that for all $0 \leq t \leq t_0$

$$\phi(t) \leq \phi(0) e^{Ct} + \alpha Q(t).$$

The next useful result is

LEMMA 2.6. *For $t \geq 0$, we have*

$$|L(t)|_{\mathcal{L}(L^2)} \leq 1.$$

PROOF. Let u belong to H and let $\mathcal{F}(u)$ be its Fourier transform. The Fourier transform of $L(t)u$ is given by

$$\mathcal{F}(L(t)u)(\xi) = (1 - tM|\xi|^2)(1 + tM|\xi|^2)^{-1} \mathcal{F}(u)(\xi).$$

Following the definition of H and denoting

$$v(\xi) = (1 + tM|\xi|^2)^{-1} \mathcal{F}(u)(\xi), \tag{2.29}$$

we have

$$\begin{aligned}|L(t)\,u|_0^2 &= |\mathcal{F}(L(t)u)|_0^2\\ &= \int_{\mathbb{R}^d} \langle (1-tM|\xi|^2)v(\xi), (1-tM|\xi|^2)v(\xi)\rangle\, d\xi,\\ &= \int_{\mathbb{R}^d} v^*(\xi)(1-tM^*|\xi|^2)S(1-tM|\xi|^2)v(\xi)\, d\xi.\end{aligned}$$

Finally, with the help of (2.7), we obtain

$$\begin{aligned}|L(t)\,u|_0^2 - |u|_0^2 &= -2t\int_{\mathbb{R}^d} v^*(\xi)(M^*S+SM)|\xi|^2 v(\xi)\, d\xi\\ &= -2t\int_{\mathbb{R}^d} |\xi|^2 v^*(\xi)v(\xi)\, d\xi \le 0;\end{aligned}$$

this concludes the proof of Lemma 2.6. □

The previous lemma allows us to prove the stability of Peaceman-Rachford scheme in next corollary, which is the first non linear result in the article:

Corollary 2.7. *There exists a constant $C_0 > 0$ such that for t small enough and all u_0 and v_0 in $L^2(\mathbb{R}^d)^m$*

$$\left|P^t u_0 - P^t v_0\right|_0 \le (1+C_0 t)|u_0 - v_0|_0. \tag{2.30}$$

Proof. Let us recall that $P^t u_0$ is defined by

$$P^t u_0 = (\mathbf{1}+tF/2)^{-1}\, L(t/2)\,(\mathbf{1}-tF/2)\, u_0.$$

Denote $u_1 = L(t/2)\,(\mathbf{1}-tF/2)\,u_0$ and $v_1 = L(t/2)\,(\mathbf{1}-tF/2)\,v_0$, we have

$$|u_1 - v_1|_0 \le |L(t/2)|_{\mathcal{L}(H)}\, |u_0 - v_0 - tF(u_0)/2 + tF(v_0)/2|_0,$$

since F is Lipschitz continuous and using Lemma 2.6, we deduce that

$$|u_1 - v_1|_0 \le (1+Ct)|u_0 - v_0|_0.$$

Next

$$\begin{aligned}|P^t u_0 - P^t v_0|_0 &\le |u_1 - v_1|_0 + t/2|F(P^t u_0) - F(P^t v_0)|_0\\ &\le |u_1 - v_1|_0 + Ct/2|P^t u_0 - P^t v_0|_0\\ &\le (1+Ct)|u_0 - v_0|_0 + Ct/2|P^t u_0 - P^t v_0|_0.\end{aligned}$$

Thus, for t small, we obtain that

$$\left|P^t u_0 - P^t v_0\right|_0 \le (1+C_0 t)|u_0 - v_0|_0.$$

□

We also need the following linear result:

LEMMA 2.8. *There exists a constant $C > 0$ such that for t small and all u_0 in $H^2(\mathbb{R}^d)^m$*

$$\left|\left(e^{tM\Delta} - \mathbf{1}\right) u_0\right|_0 \leq Ct|u_0|_2, \tag{2.31}$$

and such that for t small and all u_0 in $W^{2,\infty}(\mathbb{R}^d)^m$

$$\left|\left(e^{tM\Delta} - \mathbf{1}\right) u_0\right|_\infty \leq Ct|u_0|_{2,\infty}. \tag{2.32}$$

PROOF. Since it has been shown in [**49**] that the operator $-M\Delta$ is sectorial in $L^2(\mathbb{R}^d)^m$, estimate (2.31) is a consequence of Henry [**65**] p. 22. Concerning estimate (2.32), if we consider the case $m = 1$, the fundamental solution of the heat equation is given explicitly by

$$G(t,x) = (4\pi t)^{-d/2} \exp(-|x|^2/4\pi t)\chi_{\{t\geq 0\}}\mathbf{1}$$

and (2.32) is an immediate consequence of a Taylor's formula. The matrix generalization is straightforward using the generalized Green function

$$G(t,x) = (4\pi Mt)^{-d/2} \exp(-M^{-1}|x|^2/4t)\chi_{\{t\geq 0\}}\mathbf{1},$$

and the fact that

$$\int \left\|(4\pi M)^{-d/2} \exp(-M^{-1}|y|^2/4)\right\| |y|^2 dy \leq C.$$

□

2.4.2. *The difference between $P^t u_0$ and $T^t u_0$.* We recall that $\mathcal{G}$ is the subspace of $H^6(\mathbb{R}^d)^m$ made out of functions which belong to $C^6(\mathbb{R}^d)^m$ whose first sixth derivatives are bounded and, for u_0 belonging to $\mathcal{G}$, we denote by $C(|u_0|_{6,\infty})$ a constant depending on $\max_{0\leq|\alpha|\leq 6} |\partial^\alpha u_0|_\infty$. We have the following result:

THEOREM 2.9. *For u_0 in $\mathcal{G}$ and t small, the following estimate holds:*

$$|P^t u_0 - T^t u_0|_0 \leq C(|u_0|_{6,\infty})t^3|u_0|_6. \tag{2.33}$$

The proof of Theorem 2.9 depends on three lemmas and we begin by introducing two auxiliary functions. Define

$$F_1(v) = F(u_0) + DF(u_0)(v - u_0),$$

$v = T^t_{\text{aff}} u_0$ is the solution of the following system

$$\begin{cases} \dfrac{\partial v}{\partial t} - M\Delta v + F_1(v) = 0 & x \in \mathbb{R}^d,\ t > 0, \\ v(0,x) = u_0(x), & x \in \mathbb{R}^d. \end{cases} \tag{2.34}$$

The system (2.34) is not the linearized system at u_0 but the expansion up to order one and aff stands for affine. Using Duhamel's formula, the solution $v = T^t_{\text{aff}} u_0$ of (2.34) is given explicitly by

$$v = e^{t(M\Delta - DF(u_0))}u_0 + \int_0^t e^{(t-s)(M\Delta - DF(u_0))}G(u_0)ds. \tag{2.35}$$

For simplicity we also denote:

$$G(u_0) = DF(u_0)u_0 - F(u_0),$$

the constant part of F_1, so that

$$F_1(v) = DF(u_0)v - G(u_0).$$

We expect T^t_{aff} to be a second order approximation of T^t, since there is one time integration; that this expectation holds is proved in next lemma.

LEMMA 2.10. *For u_0 in $\mathcal{G}$ and t small, the following estimate holds:*

$$|T^t_{\text{aff}}u_0 - T^t u_0|_0 \leq C(|u_0|_{6,\infty})t^3|u_0|_2. \tag{2.36}$$

PROOF. Let us define the function $y = v - u = T^t_{\text{aff}}u_0 - T^t u_0$. If we denote $R(u, u_0) = F(u) - F_1(u)$, the function y verifies the system

$$\begin{cases} \dfrac{\partial y}{\partial t} - M\Delta y + DF(u_0)y = R(u, u_0) & x \in \mathbb{R}^d, t > 0, \\ y(0, x) = 0, & x \in \mathbb{R}^d. \end{cases} \tag{2.37}$$

We start by an estimate of $|R(u, u_0)|_0$. Taylor's formula is written

$$R(u, u_0) = \int_0^1 (1-t)D^2F\left(u_0 + t(u - u_0)\right)(u - u_0)^{\otimes 2}dt \tag{2.38}$$

and therefore

$$|R(u, u_0)|_0 \leq \left|\int_0^1 (1-t)D^2F\left(u_0 + t(u - u_0)\right)(u - u_0)^{\otimes 2}dt\right|_0. \tag{2.39}$$

Since D^2F is bounded, it follows that there exists a constant $C > 0$ such that

$$|R(u, u_0)|_0 \leq C|u - u_0|_0|u - u_0|_{L^\infty}. \tag{2.40}$$

Now, let us estimate the two norms of the right hand side of the above inequality; $u - u_0$ verifies

$$u(t) - u_0 = \left(e^{tM\Delta} - \mathbf{1}\right)u_0 - \int_0^t e^{(t-s)M\Delta}F(u)ds.$$

Since F is a Lipschitz function, and $F(0) = 0$, we deduce that for $t \in [0, 1]$

$$\begin{aligned} |u(t) - u_0|_0 &\leq \left|\left(e^{tM\Delta} - \mathbf{1}\right)u_0\right|_0 + C\int_0^t |u|_0 ds \\ &\leq \left|\left(e^{tM\Delta} - \mathbf{1}\right)u_0\right|_0 + C\int_0^t |u - u_0|_0 ds + Ct|u_0|_0. \end{aligned}$$

Using the estimate (2.31) of the norm $\left|\left(e^{tM\Delta} - \mathbf{1}\right)u_0\right|_0$, we obtain

$$|u - u_0|_0 \leq Ct|u_0|_2 + C\int_0^t |u - u_0|_0 ds, \tag{2.41}$$

therefore using Lemma 2.5 we obtain that for t small

$$|u - u_0|_0 \leq Ct|u_0|_2.$$

Using (2.32) and arguing as for (2.41), we have

$$|u - u_0|_{L^\infty} \leq Ct|u_0|_{W^{2,\infty}} + C\int_0^t |u - u_0|_{L^\infty} ds,$$

and using Lemma 2.5 again we get

$$|u - u_0|_{L^\infty} \leq Ct|u_0|_{W^{2,\infty}} \leq C(|u_0|_{6,\infty})t.$$

Since the function $y = v - u$ is solution of (2.37), using Duhamel's formula and estimate (2.40) on $R(u, u_0)$, we obtain that

$$\begin{aligned} |y|_0 &\leq C\int_0^t |u - u_0|_0 |u - u_0|_{L^\infty} ds \\ &\leq C(|u_0|_{6,\infty})t^3|u_0|_2. \end{aligned}$$

This concludes the proof of Lemma 2.10. □

For the Peaceman-Rachford approximation of (2.34), the object equivalent to T^t_{aff} is

$$P^t_{\text{aff}}v = (\mathbf{1} + tF_1/2)^{-1} L(t/2)(\mathbf{1} - tF_1/2)v. \tag{2.42}$$

The value of $(\mathbf{1} + tF_1/2)^{-1}v$ is obtained by solving

$$z - tG(u_0)/2 + tDF(u_0)z/2 = v$$

and consequently

$$(\mathbf{1} + tF_1/2)^{-1} v = (\mathbf{1} + tDF(u_0)/2)^{-1}(v + tG(u_0)/2); \tag{2.43}$$

moreover

$$(\mathbf{1} - tF_1/2)v = (\mathbf{1} - tDF(u_0)/2)v + tG(u_0)/2.$$

These identities enable us to write $P^t_{\text{aff}}v$ as the sum of a constant term and a linear term in v:

$$\begin{aligned} P^t_{\text{aff}}v = &(\mathbf{1} + tDF(u_0)/2)^{-1} L(t/2)(\mathbf{1} - tDF(u_0)/2)v \\ &+ t/2(\mathbf{1} + tDF(u_0)/2)^{-1}(\mathbf{1} + L(t/2))G(u_0). \end{aligned} \tag{2.44}$$

In the two next lemmas, we prove that the difference between T^t_{aff} and P^t_{aff} is of order two:

LEMMA 2.11. *For u_0 in $\mathcal{G}$ and t small, the following estimate holds:*

$$|P^t_{\text{aff}}u_0 - T^t_{\text{aff}}u_0|_0 \leq C(|u_0|_{6,\infty})t^3|u_0|_6. \tag{2.45}$$

PROOF. Since v is given by (2.35), we can write the difference $v - w = T^t_{\text{aff}}u_0 - P^t_{\text{aff}}u_0$ as

$$v - w = E_1 + E_2$$

with E_1 the difference between linear terms

$$\begin{aligned} E_1 =& e^{t(M\Delta - DF(u_0))}u_0 \\ & - (\mathbf{1} + tDF(u_0)/2)^{-1} L(t/2) (\mathbf{1} - tDF(u_0)/2) u_0 \end{aligned}$$

and E_2 the difference between constant terms

$$\begin{aligned} E_2 = & - t/2 (\mathbf{1} + tDF(u_0)/2)^{-1} (\mathbf{1} + L(t/2)) G(u_0) \\ & + \int_0^t e^{(t-s)(M\Delta - DF(u_0))} G(u_0) ds. \end{aligned}$$

Remarking that the first term on the right hand side of (2.44) is the Peaceman-Rachford approximation of $e^{t(M\Delta - DF(u_0))}u_0$ and using (2.27) we have

$$|E_1(t)|_{\mathcal{L}(H^6, L^2)} \leq C(|u_0|_{6,\infty}) t^3.$$

E_2 can be rewritten

$$E_2 = \int_0^t \Gamma(t,s) ds$$

with

$$\begin{aligned} \Gamma(t,s) =& e^{(t-s)(M\Delta - DF(u_0))} G(u_0) \\ & - (\mathbf{1} + tDF(u_0)/2)^{-1} (\mathbf{1} + L(t/2)) G(u_0)/2. \end{aligned}$$

To conclude the proof of Lemma 2.11, it suffices to show that

$$\left| \int_0^t \Gamma(t,s) ds \right|_0 \leq C(|u_0|_{6,\infty}) t^3 |u_0|_6,$$

which is proved in next lemma. □

This is a complicated lemma: as to estimate the three terms of Γ (one exponential of a sectorial operator and two rational approximations), we need to introduce two new exponential terms, to study their difference with their rational approximation (using methods similar as those of sections 2 and 3), and finally to estimate the sum of the three exponential terms.

Lemma 2.12. *We have the following estimate*

$$\left| \int_0^t \Gamma(t,s) ds \right|_0 \leq C(|u_0|_{6,\infty}) t^3 |u_0|_6. \tag{2.46}$$

Proof. We rewrite the term $\Gamma(t,s)$ under the form

$$\begin{aligned} \Gamma(t,s) = & \Big(e^{(t-s)(M\Delta - DF(u_0))} - e^{t(M\Delta - DF(u_0)/2)}/2 \\ & - e^{-tDF(u_0)/2}/2 \Big) G(u_0) \\ & + \left(e^{t(M\Delta - DF(u_0)/2)} - (\mathbf{1} + tDF(u_0)/2)^{-1} L(t/2) \right) G(u_0)/2 \\ & + \left(e^{-tDF(u_0)/2} - (\mathbf{1} + tDF(u_0)/2)^{-1} \right) G(u_0)/2. \end{aligned} \tag{2.47}$$

We have chosen two new exponential terms so that their limited expansions coincides up to order one with the rational expressions involved in $\Gamma(t,s)$. The three subexpressions in Γ are called Γ_j:

$$\Gamma(t,s) = \Gamma_1(t,s) + \Gamma_2(t,s) + \Gamma_3(t,s).$$

We begin with the last term of the right hand side of (2.47), i.e. Γ_3. Since $DF(u_0)$ is a bounded operator, it is clear that for small t

$$|\Gamma_3(t,s)|_0 \leq C(|u_0|_{6,\infty})t^2|u_0|_0. \tag{2.48}$$

To simplify the notations we write

$$\begin{aligned} A &= DF(u_0)/2, & B &= -M\Delta/2, \\ \alpha &= (\mathbf{1} + tDF(u_0)/2)^{-1}, & \beta &= (\mathbf{1} - tM\Delta/2)^{-1}. \end{aligned}$$

We can use the method of Lemma 2.1 for estimating the second term of the right hand side of (2.47):

$$\Gamma_2(t,s) = \left(e^{-t(A+2B)} - \alpha(2\beta - \mathbf{1})\right) G(u_0)/2.$$

This calculation is different from the previous ones, since a term is missing from the Peaceman-Rachford approximation. A simple computation (we omit here the details) shows that if we write

$$Q_1(t) = \alpha^2 A^2(2\beta - \mathbf{1}) + 2\alpha[A,B]\alpha\beta^2 - 2tB\alpha\beta^2 B^2,$$

we have formally, assuming $DF(u_0)$ to be smooth enough,

$$\Gamma_2(t,s) = -\int_0^t s e^{-(t-s)(A+2B)} Q_1(s)ds.$$

Thus, as in the section 2.3, it is easy to show that the term $\Gamma_2(t,s)$ satisfies

$$|\Gamma_2(t,s)|_0 \leq C(|u_0|_{6,\infty})t^2|u_0|_6. \tag{2.49}$$

Finally, we have to consider the first term of the right hand side of (2.47)

$$\Gamma_1(t,s) = \left(e^{-2(t-s)(B+A)} - e^{-t(2B+A)}/2 - e^{-tA}/2\right)G(u_0).$$

It follows from Kato [**71**] that for $w \in D(S^2)$, the following formula holds:

$$e^{-tS}w = w - tSw + \int_0^t (t-s)e^{-sS}S^2 w ds = W(S,t)w,$$

with

$$|W(S,t)w|_0 \leq Ct^2 \left|S^2 w\right|_0. \tag{2.50}$$

Applying this estimate for the three terms

$$e^{-2(t-s)(B+A)}G(u_0)\ ,\ e^{-t(2B+A)}G(u_0) \text{ and } e^{-tA}G(u_0)$$

we obtain

$$\Gamma_1(t,s) = (2s-t)(A+B)G(u_0) + O(t^2)G(u_0).$$

Integrating this equality between 0 and t, the following estimate holds

$$\left|\int_0^t \Gamma_1(t,s)ds\right|_0 \leq C(|u_0|_{6,\infty})t^3|u_0|_4. \tag{2.51}$$

Summarizing (2.48), (2.49) and (2.51), estimate (2.46) is now clear. □

We now reach the last lemma concerning the difference between $P^t_{\text{aff}}u_0$ and $P^t u_0$.

LEMMA 2.13. *For u_0 in $\mathcal{G}$ and t small, the following estimate holds:*

$$|P^t_{\text{aff}}u_0 - P^t u_0|_0 \leq C(|u_0|_{6,\infty})t^3|u_0|_2. \tag{2.52}$$

PROOF. Formula (2.42) gives

$$P^t_{\text{aff}}u_0 = (\mathbf{1} + tF_1/2)^{-1}\, L\,(t/2)\,(\mathbf{1} - tF_1/2)\, u_0.$$

We observe that $F_1(u_0) = F(u_0)$; therefore $w = P^t_{\text{aff}}u_0$ can be rewritten with the help of (2.43) as

$$\begin{aligned} w = &(\mathbf{1} + tDF(u_0)/2)^{-1}\, L\,(t/2)\,(\mathbf{1} - tF/2)\, u_0 \\ &+ t/2\,(\mathbf{1} + tDF(u_0)/2)^{-1}\, G(u_0); \end{aligned}$$

and thus w satisfies

$$w + tDF(u_0)w/2 = L\,(t/2)\,(\mathbf{1} - tF/2)\, u_0 + tG(u_0)/2. \tag{2.53}$$

On the other hand, $v = P^t u_0$ satisfies

$$v + tF(v)/2 = L\,(t/2)\,(\mathbf{1} - tF/2)\, u_0. \tag{2.54}$$

We subtract (2.53) from (2.54)

$$v - w + tF(v)/2 - tDF(u_0)w/2 = -tG(u_0)/2. \tag{2.55}$$

In (2.55), we replace $F(v)$ by $F_1(v) + R(v,u_0)$, which yields

$$\begin{aligned} v - w &+ tDF(u_0)v/2 - tG(u_0)/2 + tR(v,u_0)/2 - tDF(u_0)w/2 \\ &= -tG(u_0)/2, \end{aligned}$$

and therefore

$$v - w = -t/2\,(\mathbf{1} + tDF(u_0)/2)^{-1}\, R(v,u_0). \tag{2.56}$$

It remains to estimate $R(v,u_0)$ and we will show

$$|v - u_0|_0 \leq C(|u_0|_{6,\infty})t|u_0|_2.$$

Relation (2.54) is equivalent to

$$v - u_0 = (\mathbf{1} + tF/2)^{-1}\, L\,(t/2)\,(\mathbf{1} - tF/2)\, u_0 - u_0.$$

The aim of the following calculation is to write $v - u_0$ as the product of t by a term which can be estimated. Another way of writing (2.9) is

$$L\,(t/2) = \mathbf{1} + tM\Delta\,(\mathbf{1} - tM\Delta/2)^{-1}\,,$$

then

$$\begin{aligned} v - u_0 &= (\mathbf{1} + tF/2)^{-1} (\mathbf{1} - tF/2)\, u_0 - u_0 \\ &\quad + (\mathbf{1} + tF/2)^{-1}\, tM\Delta\, (\mathbf{1} - tM\Delta/2)^{-1} (\mathbf{1} - tF/2)\, u_0, \\ &= (\mathbf{1} + tF/2)^{-1}\, tM\Delta\, (\mathbf{1} - tM\Delta/2)^{-1} (\mathbf{1} - tF/2)\, u_0 \\ &\quad - (\mathbf{1} + tF/2)^{-1} (tF(u_0)) \end{aligned}$$

and the estimate on $v - u_0$ is now clear. This leads to an estimate on $v - w$ by the product of t^3 and of an adequate term. □

2.4.3. *Convergence.*

THEOREM 2.14. *For all u_0 in $\mathcal{G}$ and for all $\tau > 0$, there exist $C(|u_0|_{6,\infty})$ and h_0 such that for all $h \in (0, h_0]$, for all n such that $nh \leq \tau$*

$$\left|\left(P^h\right)^n u_0 - T^{nh} u_0\right|_0 \leq C(|u_0|_{6,\infty}) h^2 |u_0|_6.$$

PROOF. We fix $\tau > 0$, the triangle inequality gives

$$\begin{aligned} \left|\left(P^h\right)^n u_0 - T^{nh} u_0\right|_0 \leq \sum_{j=0}^{n-1} \Big| &\left(P^h\right)^{n-j-1} P^h T^{jh} u_0 \\ &- \left(P^h\right)^{n-j-1} T^{(j+1)h} u_0 \Big|_0 \end{aligned}$$

and we infer from (2.30) that

(2.57)

$$\left|\left(P^h\right)^n u_0 - T^{nh} u_0\right|_0 \leq \sum_{j=0}^{n-1} (1 + C_0 h)^{n-j-1} \left|P^h T^{jh} u_0 - T^h T^{jh} u_0\right|_0 .$$

For the case $j = 0$, it follows from Theorem 2.9 that $|P^h u_0 - T^h u_0|_0 \leq C(|u_0|_{6,\infty}) h^3 |u_0|_6$. For $j \geq 1$, we notice that $T^{jh} u_0$ belongs to H^6 and we still use Theorem 2.9 to obtain

$$|P^h T^{jh} u_0 - T^h T^{jh} u_0|_0 \leq C(|u_0|_{6,\infty}) h^3 |T^{jh} u_0|_6,$$

this yields with (2.57)

$$\left|\left(P^h\right)^n u_0 - T^{nh} u_0\right|_0 \leq C(|u_0|_{6,\infty}) n h^3 |u_0|_6.$$

Thus we obtain

$$\left|\left(P^h\right)^n u_0 - T^{nh} u_0\right|_0 \leq C(|u_0|_{6,\infty}) h^2 |u_0|_6.$$

This concludes the proof of Theorem 2.14. □

2.5. Numerical Implementation. We use the Peaceman-Rachford scheme to solve the system (2.1) presented in the introduction and a Ginzburg-Landau equation.

2.5.1. *Leopard spots system.* We recall the equations of the system (2.1)

$$\begin{cases} \dfrac{\partial u}{\partial t} = \Delta u + \gamma \left(a - u - \dfrac{\rho u v}{1 + u + K u^2} \right), \\ \dfrac{\partial v}{\partial t} = d \Delta v + \gamma \left(\alpha (b - v) - \dfrac{\rho u v}{1 + u + K u^2} \right). \end{cases}$$

This has been proposed as a model of Turing system where spatial structures are created by interaction of non linear phenomena and spectral properties. The numerical parameters are

$$\begin{array}{llll} a = 92, & b = 64, & \alpha = 1.5, & \rho = 18.5, \\ K = 0.1, & d = 10, & \gamma = 1. & \end{array}$$

For one time step, we have to do an Euler-forward scheme on the non-linear part, a Crank-Nicolson scheme on the linear part and an Euler-backward scheme on the non-linear part. It is useful to program several time steps altogether, as for n time steps the formula can be written:

$$\begin{aligned} P^{nt} u_0 =& \left((\mathbf{1} + tF/2)^{-1} L(t/2) (\mathbf{1} - tF/2) \right)^n u_0 \\ =& (\mathbf{1} + tF/2)^{-1} L(t/2) \\ & \left((\mathbf{1} - tF/2) (\mathbf{1} + tF/2)^{-1} L(t/2) \right)^{n-1} (\mathbf{1} - tF/2) u_0 \\ =& (\mathbf{1} + tF/2)^{-1} L(t/2) \\ & \left(\left(2 (\mathbf{1} + tF/2)^{-1} - \mathbf{1} \right) L(t/2) \right)^{n-1} (\mathbf{1} - tF/2) u_0. \end{aligned}$$

In practice, we replace the non linear implicit step by its approximation of order 2:

$$(\mathbf{1} + tF/2)^{-1} \simeq \mathbf{1} - tF/2 (\mathbf{1} - tF/2).$$

and to compute more easily the Crank-Nicolson scheme and to avoid programming an Euler-forward scheme, we notice that

$$\begin{aligned} L(t/2) =& (\mathbf{1} + tM\Delta/2) (\mathbf{1} - tM\Delta/2)^{-1} \\ =& 2 (\mathbf{1} - tM\Delta/2)^{-1} - \mathbf{1}. \end{aligned}$$

The spatial discretization is performed with the finite element method and all computations are made using Freefem+. The evolution of the solution of (2.1), starting from four peaks is shown in Figure 2.

2.5.2. *Ginzburg-Landau equation.* We introduce the quintic Ginzburg – Landau equation:

$$\frac{\partial u}{\partial t} = m_0 \Delta u + m_1 u + m_2 |u|^2 u + m_3 |u|^4 u$$

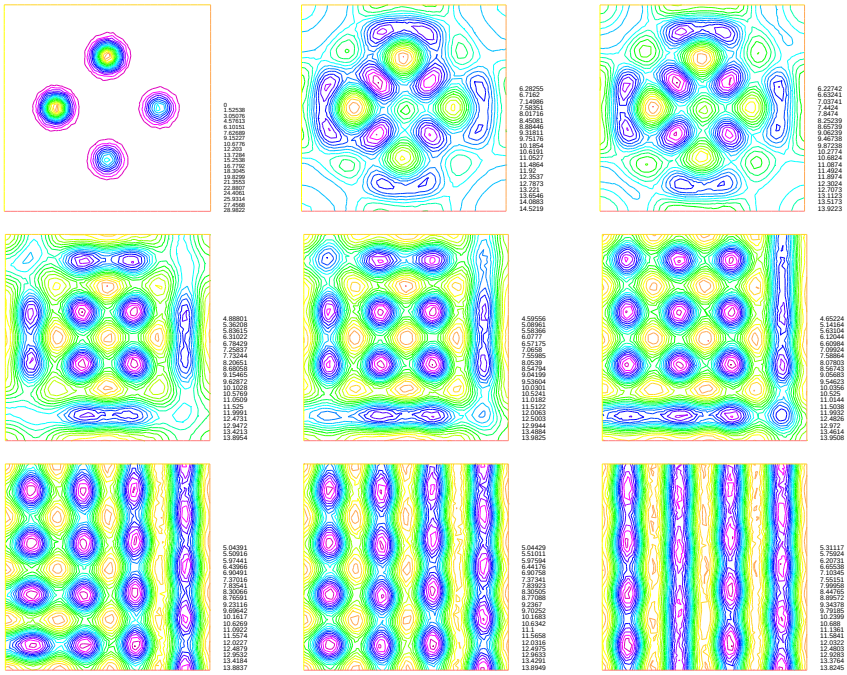


FIGURE 2. The evolution of the solution of (2.1), starting from four peaks.

where the coefficients m_0, m_1, m_2 and m_3 belong to the complex plane and $u : \mathbb{C} \to \mathbb{C}$.

This equation leads to interesting stable pulse-like solutions; these phenomena have been studied in [**121**].

As this equation is a complex one, we split it into a system of two equations, one for the real part, one for the imaginary part and we use exactly the same scheme as above to solve it with these numerical parameters:

$$\begin{array}{llll} \mathrm{Re}m_0 = 1, & \mathrm{Re}m_1 = -0.1, & \mathrm{Re}m_2 = 4, & \mathrm{Re}m_3 = -2.75, \\ \mathrm{Im}m_0 = 0, & \mathrm{Im}m_1 = 0, & \mathrm{Im}m_2 = 1, & \mathrm{Im}m_3 = 1. \end{array}$$

The numerical results, exhibited in Figure 3 show collapse between two pulses.

FIGURE 3. The evolution of the solution of the quintic Ginzburg-Landau, starting from two pulses.

CHAPITRE 3

Un système d'Allen – Cahn avec transition triple ; application numérique à la croissance de grains

1. Rappels sur les méthodes multipas explicites - implicites

G. Akrivis, M. Crouzeix et C. Makridakis [**1**] ont proposé un schéma multipas explicite-implicite ; le principe est d'utiliser deux formules multipas : une formule explicite pour la partie non-linéaire et une formule implicite pour la partie linéaire. Écrivons ce schéma pour l'équation suivante

$$\begin{cases} \dfrac{\partial u}{\partial t}(t) - Au(t) = F(u(t)),\ 0 < t < T \\ u(0) = u_0, \end{cases}$$

où A est un opérateur linéaire. Cette formulation est un peu plus générale que celle du système de réaction-diffusion (1.1).

Soit V_h l'espace de discrétisation spatiale et A_h et F_h les approximations dans cet espace de A et F. On suppose alors que A et A_h sont auto – adjoints.

Soit Δt le pas de temps ; on se donne les q premières approximations $U^0, \ldots, U^{q-1} \in V_h$ de $u(t_0), \ldots, u(t_{q-1})$. Alors le schéma considéré par G. Akrivis, M. Crouzeix et C. Makridakis est le schéma multipas suivant :

$$\sum_{i=0}^{q} \alpha_i U^{n+i} + \Delta t \sum_{i=0}^{q} \beta_i A_h U^{n+i} = \Delta t \sum_{i=0}^{q-1} \gamma_i F_h(U^{n+i}).$$

Sous une hypothèse de stabilité, qui implique $q \leq 6$, cette méthode est consistante, convergente et son ordre en temps est l'ordre des schémas explicite (α, γ) et implicite (α, β). Cependant, pour $A = -M\Delta$, ces résultats ne sont valables que lorsque la matrice de diffusion M est auto – adjointe, puisque A doit être auto – adjoint.

2. Une modélisation de la croissance de grains par un système de réaction – diffusion

A reaction – diffusion approach for grain growth

Magali Ribot

Abstract : *Grain growth is a problem of great interest in material science. Several deterministic or statistic methods have been introduced to simulate it but it would be interesting to have an evolution equation to compute it efficiently. We present here a method based on the evolution of developed interfaces where the evolution equation is a reaction – diffusion equation.*

2.1. Introduction. The grain growth is a major problem of materials science to understand polycrystalline materials. Here we try to simulate two–dimensional grain growth by using a reaction–diffusion equation with a developed interface.

The evolution of grains has been studied theoretically by Kinderlehrer and Liu [**73**] using a network of grain boundaries subject to curvature driven growth. Let us explain briefly on which principles relies this method.

We consider a grain boundary as a curve

$$\Gamma : x = \xi(s,t),\ 0 \leq s \leq 1$$

with the tangent l and the normal n defined as

$$l = \frac{\partial \xi}{\partial s} \text{ and } n = Rl,$$

where R is the rotation through $\pi/2$. We introduce the surface energy density or surface tension $\sigma = \sigma(\theta, \alpha)$ which depends on the angle θ between the normal n and the x-axis and on the misorientation between two grains, α. Mullins' equation expresses the balance between the rate of growth of area and the work done through deforming the curve, where v_n is the normal velocity of the grain boundary, μ is the mobility, T is the line stress and its derivative dT/ds is the line force per unit length and κ is the curvature:

$$v_n = \mu \frac{dT}{ds} = \mu \left(\frac{d^2\sigma}{d\theta^2} + \sigma \right) \kappa. \tag{2.1}$$

We can define the grain boundary energy by

$$E(t) = \int_0^1 \sigma(\theta(s,t), \alpha) |l(s,t)| ds.$$

Let us consider now a triple junction, a point where three boundaries meet. The total rate of dissipation at this point is computed by differentiating the energy with respect to time [**73**]; to ensure that the

whole system is dissipative, we impose the Herring condition at the triple junction point:

$$\sum_{j=1\cdots 3} T^j = 0, \tag{2.2}$$

where T^j is the line stress associated to the jth-boundary, whose endpoint is the triple junction.

Kinderlehrer and Liu [**73**] show that this equation (2.1) of curvature driven growth with the Herring condition (2.2) can be solved for all times provided the initial condition is closed to an equilibrium state. Moreover for any sequence of states, there exists a subsequence that converges to a stationary configuration, not necessarily the one close to the initial condition.

This curvature driven growth method also gives a numerical scheme to simulate the grain growth [**74, 75**]. We call here triple junctions the endpoints of the edges and interior nodes the other discretization points of the boundaries. First, the parabolic partial differential equation (2.1) is discretized in the interior nodes and then the Herring condition (2.2) is solved by a fixed point method to find the new triple junctions. This algorithm is not so simple since simulations must take into account two critical events which are grain boundary flipping and grain disappearance. Grain boundary flipping consists in the removal of too short edges; this leads to a quadruple junction which is straight away split into two new triple junctions. Another problem is that this algorithm is time explicit and therefore to ensure stability, we must have a time step restriction $\Delta t = O(h^2)$, where h is the parameter describing the grain boundary discretization.

A review of some methods to simulate grain growth can be found in [**119**], among them the previous algorithm and also some statistical simulations. One is based on Monte-Carlo model; the microstructure is based on a lattice and each lattice site i is assigned a spin number S_i. We chose randomly a site i and a new orientation S_i^* with a transition probability depending on the change of energy. The grain boundary is then defined to lie between sites of different orientations. This is a simple model since, in particular, there is no need to care of critical events; however, this scheme needs high memory and CPU time to discretize the whole network, including the grain interiors.

Another statistical method is a stochastic model [**119**]; this model is based on a deterministic evolution based on the $(n-6)$ rule interrupted by random critical events. The Von Neumann–Mullins $(n-6)$ rule says that the temporal variation of the area $A(t)$ of a grain is proportional to $n-6$, where n is the number of edges of the grain, namely

$$\frac{dA}{dt}(t) = \frac{\pi}{3}\sigma\mu(n-6),$$

where σ and μ are the constant energy and mobility introduced before. The drawback of this method is that we need first to compute the probabilities of grain disappearance and of edge flipping by the driven growth model.

We need essentially to compute the evolution of a large system of grains during a large time since we are faced to a metastable system. Therefore, it would be interesting to have an evolution equation to simulate grain growth; this would allow to have implicit, unconditionally stable and efficient method. An idea is to use the theory of the evolution of developed interfaces modelised by a reaction–diffusion equation.

2.2. Evolution of developed interfaces. In [**46**] and [**45**], de Mottoni and Schatzman study the evolution of developed interfaces. Let us give here their main results.

Let Φ be an even smooth potential with two wells at -1 and 1 of equal depth and let $\phi = \Phi'$. Let $h > 0$ be a fixed small parameter and let $u : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ be the solution of the semilinear parabolic Allen–Cahn equation

$$\frac{\partial u}{\partial t} - h^2 \Delta u + \phi(u) = 0 \tag{2.3}$$

with initial condition

$$u(x, 0) = u_0(x),\ x \in \mathbb{R}^N. \tag{2.4}$$

If u_0 is in $L^\infty(\mathbb{R}^N)$, the system (2.3)– (2.4) possesses a unique solution which is also in $L^\infty(\mathbb{R}^N)$ uniformly in t [**46**].

For small times, the solution is given by the reaction part and in particular it tends to -1 or 1 according to the sign of the initial data u_0. When the gradient $|\nabla u|$ becomes large enough, the diffusive effect appears and the solution develops transitions between the regions $u \sim 1$ and $u \sim -1$.

The properties of the solution will be given in terms of Θ which is the unique increasing solution of

$$-\Theta'' + \phi(\Theta) = 0$$

such that $\Theta(0) = 0$. Let $\Gamma^0(h^2 t)$ be an hypersurface and $\Lambda^0(x, h^2 t)$ be the algebraic distance to the hypersurface. The evolution of the hypersurface Γ^0 is governed by two equations:

$$|\nabla \Lambda^0(x, h^2 t)| = 1 \tag{2.5}$$

in a certain tubular neighborhood of Γ^0 and

$$\left(\frac{\partial \Lambda^0}{\partial s} - \Delta \Lambda^0\right)(x, h^2 t) = 0,\ x \in \Gamma^0(h^2 t). \tag{2.6}$$

Moreover, the velocity of the hypersurface is normal and proportional to the mean curvature. This problem (2.5)- (2.6) has a unique local solution up to $t < h^{-2} T$.

With an initial condition of the form

$$u_0(x) = \Theta\left(\frac{\Lambda^0(x,0)}{h}\right)$$

in the neighborhood of $\Gamma_0 = \{x : \lambda^0(x,0) = 0\}$, we can find an expansion of u and of the distance to the interface Λ in terms of the powers of h; these expansions are valid for time $t \leq h^{-2}T^*$, where T^* is less than T [**46**].

At last, for all $\alpha > 0$ and $b > 0$, there exist constant c_1 and c_2 such that for $c_1 \log(1/h) \leq t \leq c_2/\sqrt{h}$ we have the following estimates [**45**]:

$$\left|u(x,t) - \Theta\left(\frac{\Lambda(x)}{h}\right)\right| \leq \alpha$$

when x belongs to a tubular neighborhood of the interface Γ and

$$|u(x,t) - \operatorname{sgn}(u_0(x))| \leq h^b$$

outside this neighborhood. This result gives us the time scale we have to consider.

Let us generalize this study to the case of a two-dimensional potential. Let a, b and c be the vertices of an equilateral triangle in $\mathbb{C}$, for example $a = 1$, $b = \mathrm{e}^{2i\pi/3}$ and $c = \mathrm{e}^{4i\pi/3}$. Let W be the potential

$$W(u) = |u-a|^2|u-b|^2|u-c|^2 \tag{2.7}$$

and DW its derivative

$$\begin{aligned} DW(u) = 2\big((u-a)|u-b|^2|u-c|^2 + (u-b)|u-a|^2|u-c|^2 \\ + (u-c)|u-a|^2|u-b|^2\big). \end{aligned} \tag{2.8}$$

Let Ω be the square $[-L/2, L/2] \times [-L/2, L/2]$ and $\Gamma = \partial\Omega$ be its boundary. We consider the following system where ε is a positive parameter:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \varepsilon\Delta u + \dfrac{1}{\varepsilon}DW(u)^T = 0, & x \in \Omega,\ t > 0, \\ \dfrac{\partial u}{\partial n} = 0 & x \in \Gamma,\ t > 0 \end{cases} \tag{2.9}$$

and we expect the solution to develop an interface between the three values a, b and c, that is to say some sort of grains around these three values.

2.3. Numerical implementation. To compute the approximation of the solution of the system (2.9), we use a code for reaction – diffusion systems developed by T. Dumont. There are two possibilities: either we use an explicit-implicit multistep method, such as the ones introduced by Akrivis, Crouzeix and Makridakis [**1**], either a splitting method very similar to the Peaceman-Rachford approximation studied in [**51**]. Those methods are both of order two.

The idea of the explicit-implicit multistep method is to use two multistep formulas: an explicit one for the non-linear reactive part and an implicit one for the linear diffusive part. We first discretize spatially the Laplacian thanks to finite differences approximation and Δ_h stands for the discretized Laplacian. Then, we discretize the equation (2.9) in time; we denote by Δt the time step and by $t_n = n\Delta t$ the n-th discretization time. We determine some approximations U_0 of $u(.,t_0)$ and U_1 of $u(.,t_1)$ and we denote by U_n the approximation of $u(.,t_n)$; the approximation U_{n+1} of $u(.,t_{n+1})$ can be computed by the 2-steps scheme as follows:

$$\begin{aligned} U_{n+1} - 2U_n + \frac{1}{2}U_{n-1} - \varepsilon\Delta t\Delta_h U_{n+1} \\ - \frac{\Delta t}{\varepsilon}\left(2DW(U_n)^T - DW(U_{n-1})^T\right) = 0. \end{aligned}$$

The other approximation we can use is a splitting method, also called alternate directions method. Let us explain what is a splitting method. First we consider the diffusive part

$$\left\{ \begin{array}{l} \dfrac{\partial u}{\partial t} - \varepsilon\Delta u = 0, \\ u(0,.) = v_0 \end{array} \right. \tag{2.10}$$

and we denote by $D^t v_0$ its solution at time t. We consider separately the reactive part

$$\left\{ \begin{array}{l} \dfrac{\partial u}{\partial t} + \dfrac{1}{\varepsilon}DW(u)^T = 0, \\ u(0,.) = w_0 \end{array} \right. \tag{2.11}$$

whose solution is denoted by $R^t w_0$. The Lie formulas

$$L_1^t u_0 = D^t R^t u_0 \text{ and } L_2^t u_0 = R^t D^t u_0$$

are formally some approximations of order one of the solution of system (2.9). As for the Strang formulas

$$S_1^t u_0 = D^{t/2} R^t D^{t/2} u_0 \text{ and } S_2^t u_0 = R^{t/2} D^t R^{t/2} u_0,$$

they are formally of order 2. But, according to S. Descombes and M. Massot calculations [**50**], in case of stiff problems, formula S_2^t is the one with the greatest order.

We have now to discretize in time and space the two systems (2.10) and (2.11) to compute a solution of (2.9).

We first discretize the diffusive part using a Crank-Nicolson scheme for the time discretization and finite differences approximation for the space discretization. Using the same notations as above, the scheme can be written:

$$U_{n+1} - U_n = \frac{\varepsilon\Delta t}{2}\left(\Delta_h U_{n+1} + \Delta_h U_n\right).$$

We then solve the linear system we obtain, thanks to the Fishpack library, which uses the Fast Fourier Transform.

We then discretize the reactive part still using a Crank-Nicolson scheme for the time discretization and finite differences approximation for the space discretization and we obtain the following scheme:

$$U_{n+1} - U_n = -\frac{\Delta t}{2\varepsilon}\left(DW(U_{n+1})^T + DW(U_n)^T\right).$$

To solve this non-linear system, we use Newton method since the derivative of DW is easy to compute.

Once the approximation of u is computed at time t, we visualize the function v defined by

(2.12)

$$v(x,t) = \begin{cases} a & \text{if } |u(x,t)-a| < \min\left(|u(x,t)-b|, |u(x,t)-c|\right), \\ b & \text{if } |u(x,t)-b| < \min\left(|u(x,t)-a|, |u(x,t)-c|\right), \\ c & \text{if } |u(x,t)-c| < \min\left(|u(x,t)-a|, |u(x,t)-b|\right). \end{cases}$$

2.4. Numerical results. In this section, we present the numerical results we obtain. The initial condition $u_0 : \mathbb{C} \to \mathbb{C}$ is a random data: we take randomly a complex value of u_0 at each point of the domain Ω.

When ε is big enough, we use either the multistep explicit-implicit method, either the splitting method since the results are equivalent. When ε becomes smaller, the problem becomes stiff and the multistep method develops instabilities. We therefore use the splitting method; moreover, in this case, the time for which interfaces appear increases. Thus, we have to compute further in time and we use a high order scheme, which enables us to take larger time steps. For example, in the case $\varepsilon = 0.01$, we make the computations with a Richardson extrapolation of Peaceman – Rachford which is of order four.

In figure 1, we plot the result of a simulation on a square of length $L = 100$ with a space step $h = 0.1$ and a time step $\Delta t = 0.001$ until time $T = 3$ for a random initial data; we take the parameter ε equal to 1. In figure 2, we take the same parameters except the final time which is equal to $T = 10$.

We also test the evolution of the system with parameter ε equal to $\varepsilon = 0.1$; we can see the results we obtain in figure 3 for $T = 10$ and figure 4 for $T = 30$.

We first remark that the evolution for $\varepsilon = 1$ and for $\varepsilon = 0.1$ is nearly the same, except that the time scale is different.

We notice that we do not exactly obtain the grains we expected, since they are not polygonal; the problem is maybe that we do not take into account here the Herring condition (2.2), which is important in the process of grain growth. The problem can also come from the initial data which is generally taken close to a stationary state in theoretical proofs and in numerical simulations; we have taken here random initial

datas. However, we are satisfied of the scale change which is clear in the simulations we obtain.

Eventually, we compute the evolution of the system with $\varepsilon = 0.01$ and $T = 900$; the results are displayed in figure 5. We remark a deep improvement since the grains are nearly polygonal; we expect to get better and better simulations with a smaller ε.

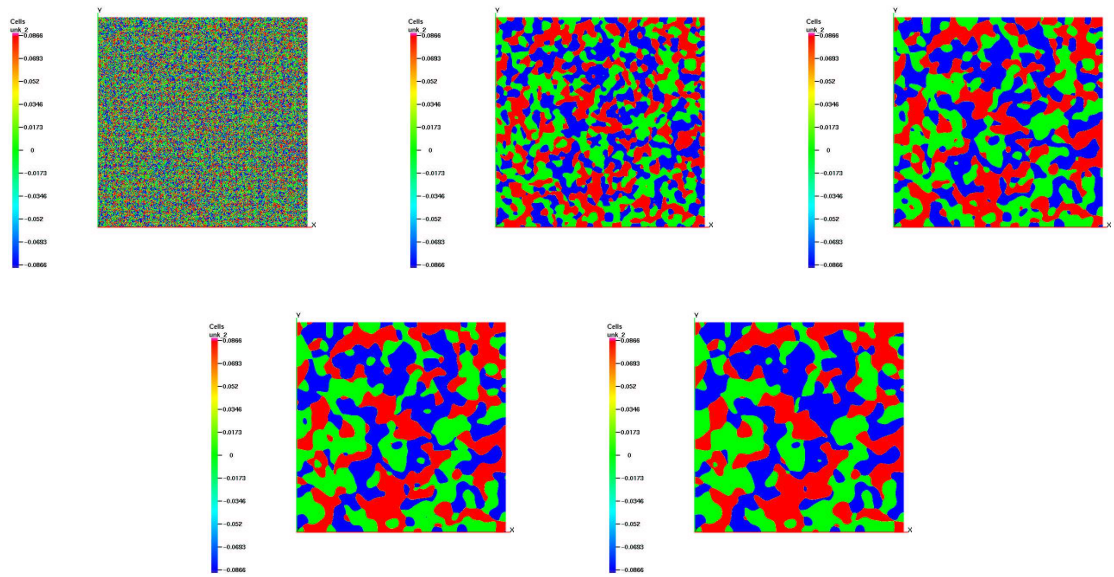


FIGURE 1. Case $\varepsilon = 1$. Evolution of the system at times $t = 0.01$, $t = 0.75$, $t = 1.5$, $t = 2.2$ and $t = 3$.

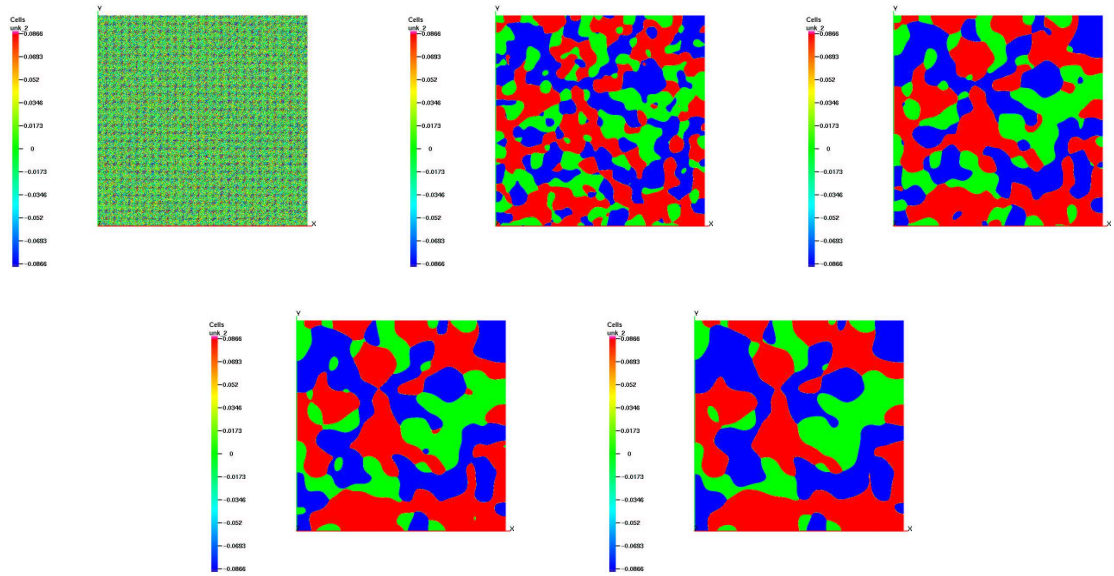


FIGURE 2. Case $\varepsilon = 1$. Evolution of the system at times $t = 0.1$, $t = 2.5$, $t = 5$, $t = 7.5$ and $t = 10$.

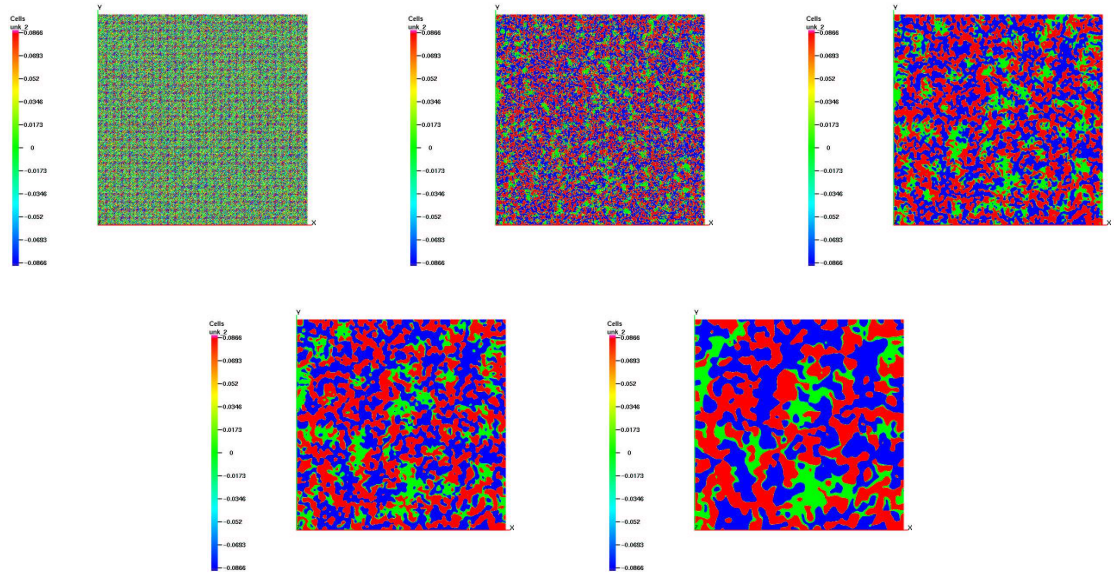


FIGURE 3. Case $\varepsilon = 0.1$. Evolution of the system at times $t = 0.01$, $t = 0.25$, $t = 1.5$, $t = 3$ and $t = 10$.

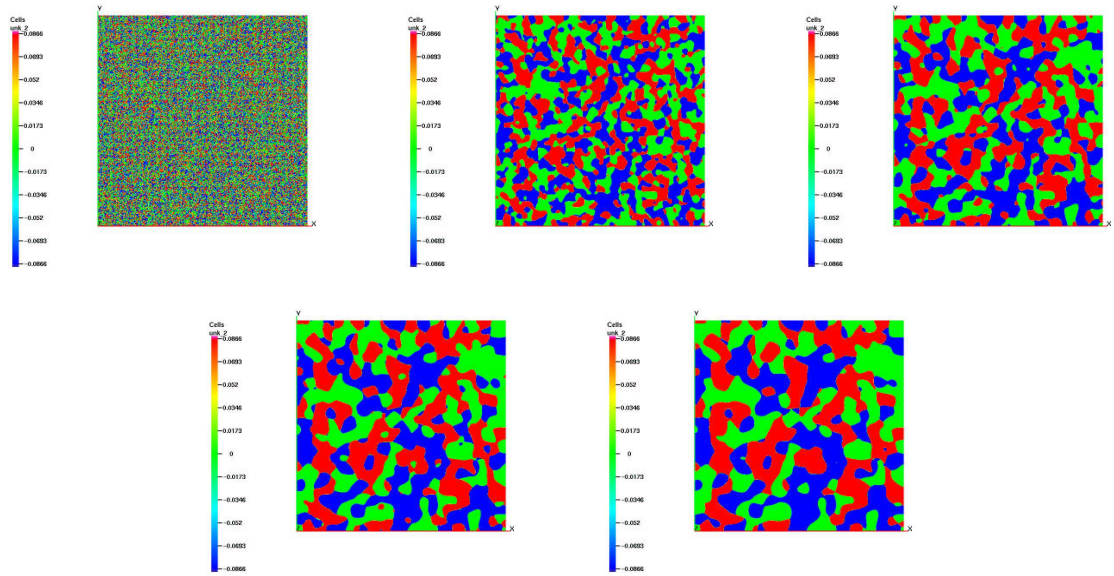


FIGURE 4. Case $\varepsilon = 0.1$. Evolution of the system at times $t = 0.1$, $t = 7.5$, $t = 15$, $t = 22$ and $t = 30$.

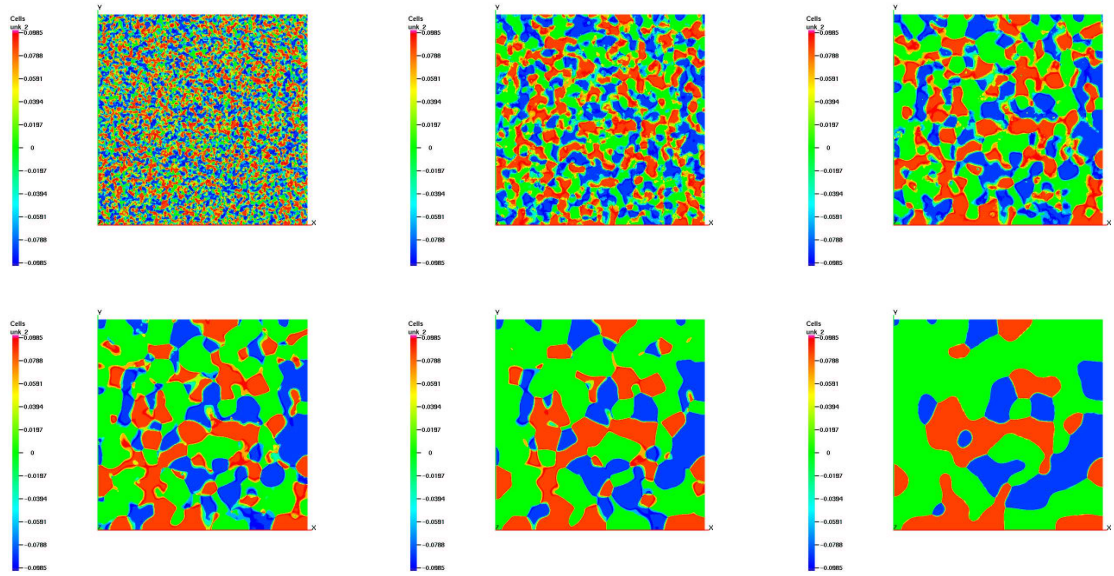


FIGURE 5. Cas $\varepsilon = 0.01$. Évolution du système aux temps $t = 10$, $t = 90$, $t = 200$, $t = 300$, $t = 400$ et $t = 900$.

CHAPITRE 4

Mécanique statistique des systèmes auto – gravitants de fermions

1. Description de l'algorithme spectral de Gauss – Lobatto – Legendre pour le système de Smoluchowski – Poisson

Nous décrivons ici en détails la méthode spectrale utilisée dans les simulations numériques.

Nous rappelons tout d'abord la définition de l'intégrale de Fermi – Dirac donnée à l'équation (5.17) de l'introduction :

$$I_n(t) = \int_0^{+\infty} \frac{x^n}{1+te^x} dx.$$

Par souci de simplicité pour la programmation, nous la notons ici J_n avec le changement de variables

$$J_n(\eta) = I_n(\mathrm{e}^{-\eta});$$

nous remarquons alors que

$$\frac{dJ_n}{d\eta} = nJ_{n-1}(\eta). \tag{1.1}$$

Nous considérons donc le système de Smoluchowski-Poisson (5.19)-(5.21) décrit dans l'introduction et que nous rappelons ici pour $x \in [0,1]$:

$$\frac{\partial n}{\partial t} = \frac{1}{x^2}\frac{\partial}{\partial x}\left\{x^2\left(\frac{\partial P}{\partial x} + n\frac{\partial \psi}{\partial x}\right)\right\}, \tag{1.2a}$$

$$\Delta\psi = 4\pi n, \tag{1.2b}$$

avec les équations d'état

$$n = \frac{\mu}{4\pi}\theta^{3/2}J_{1/2}(\eta), \tag{1.3a}$$

$$P = \frac{\mu}{6\pi}\theta^{5/2}J_{3/2}(\eta), \tag{1.3b}$$

et les conditions aux limites

$$\frac{\partial \psi}{\partial x}(0) = 0, \ \frac{\partial \psi}{\partial x}(1) = 1, \tag{1.4a}$$

$$\psi(1) = -1, \tag{1.4b}$$

$$\frac{\partial n}{\partial x}(0) = 0, \tag{1.4c}$$

$$\frac{\partial P}{\partial x}(1) + \frac{\partial \psi}{\partial x}(1)n(1) = \frac{\partial P}{\partial x}(1) + n(1) = 0. \tag{1.4d}$$

Tout d'abord, les points de discrétisation spatiale utilisés sont les points de Gauss – Lobatto – Legendre, c'est-à-dire les zéros du polynôme $(1 - X^2)L'_N$ où L_N est le N-ième polynôme de Legendre. Nous les notons ξ_k pour $0 \leq k \leq N$ avec $-1 = \xi_0 < \xi_1 < \cdots < \xi_N = 1$. Ces points ne sont pas uniformes sur l'intervalle $[-1, 1]$: la distance entre deux points consécutifs est en $O(N^{-2})$ au voisinage des extrémités et en $O(N^{-1})$ vers le milieu de l'intervalle. Ils se concentrent donc près des bords de $[-1, 1]$.

Comme les points de Gauss – Lobatto – Legendre sont compris entre -1 et 1, nous translatons la variable d'espace en $\xi = 2x - 1$. Par ailleurs, nous translatons également le potentiel gravitationnel ψ en $\widetilde{\psi} = \psi + 1$ afin d'avoir une condition de Dirichlet homogène pour $\widetilde{\psi}$ en $\xi = 1$. Pour plus de clarté, nous noterons $\widetilde{\psi}$ par ψ dans la suite.

Nous introduisons également les polynômes interpolateurs de Lagrange l_k, $0 \leq k \leq N$ associés aux points ξ_k. Nous écrivons alors la formulation variationnelle associée au système (1.2)– (1.4) en utilisant comme fonctions test les fonctions l_k définies précédemment ; nous nous servons également de l'équation suivante, qui vient de l'équation d'état (1.3) :

$$\frac{\partial P}{\partial \xi} = 2\theta c(\eta(\xi))\frac{\partial n}{\partial \xi}, \tag{1.5}$$

où la fonction c vaut

$$c(x) = \frac{J_{1/2}(x)}{J_{-1/2}(x)}; \tag{1.6}$$

nous obtenons pour tout k dans $\{1 \cdots N - 1\}$,

(1.7)

$$\int_{-1}^{1} \frac{\partial n}{\partial t}(\xi, t) l_k(\xi)(\xi + 1)^2 d\xi$$
$$= -4 \int_{-1}^{1} \left(2\theta c(\eta(\xi))\frac{\partial n}{\partial \xi} + n\frac{\partial \psi}{\partial \xi}\right)(\xi, t)\frac{\partial l_k}{\partial \xi}(\xi)(\xi + 1)^2 d\xi,$$

et

$$\int_{-1}^{1} \frac{\partial \psi}{\partial \xi}(\xi, t)\frac{\partial l_k}{\partial \xi}(\xi)(\xi + 1)^2 d\xi = -\pi \int_{-1}^{1} n(\xi, t) l_k(\xi)(\xi + 1)^2 d\xi, \tag{1.8}$$

et la fonction η est déterminée par l'équation suivante

$$J_{1/2}(\eta(\xi)) = \frac{4\pi}{\mu\theta^{3/2}} n(\xi). \tag{1.9}$$

Nous discrétisons maintenant la formulation variationnelle (1.7)-(1.8) en espace.

Soit $\mathbb{P}_N$ l'espace des polynômes de degré N définis sur $[-1, 1]$. La méthode de collocation consiste à approcher les fonctions n et ψ que l'on cherche à déterminer par leur décomposition sur la base $(l_k)_{0\leq k\leq N}$ de $\mathbb{P}_N$, c'est-à-dire on approche ψ, par exemple, par son interpolation polynômiale

$$I_N(\psi)(\xi) = \sum_{k=0}^{N} \psi(\xi_k) l_k(\xi). \tag{1.10}$$

Il nous suffit donc de déterminer la valeur de ψ aux points de Gauss – Lobatto – Legendre, c'est-à-dire de calculer le vecteur $\Psi = (\psi(\xi_k))_k$. De même pour la densité n, nous cherchons à calculer le vecteur $X = (n(\xi_k))_k$.

Nous remplaçons également les intégrales intervenant dans la formulation variationnelle par la formule de quadrature obtenue grâce à la formule de Gauss – Lobatto suivante :

$$\forall \Phi \in \mathbb{P}_{2N-1}, \int_{-1}^{1} \Phi(x)dx = \sum_{i=0}^{N} \Phi(\xi_i)\rho_i; \tag{1.11}$$

les ρ_k sont les poids de la formule de Gauss – Lobatto – Legendre et sont connus explicitement (cf. le livre de C. Bernardi et Y. Maday [**9**]).

On discrétise également la dérivée de n au point ξ en

$$\frac{\partial n}{\partial \xi}(\xi_i) = \sum_{j=0}^{N} n(\xi_j) l'_j(\xi_i) \tag{1.12}$$

et la dérivée de ψ au point ξ en

$$\frac{\partial \psi}{\partial \xi}(\xi_i) = \sum_{j=0}^{N} \psi(\xi_j) l'_j(\xi_i). \tag{1.13}$$

La première équation de la formulation variationnelle (1.7) discrétisée s'écrit alors

(1.14)

$$\begin{aligned}\frac{\partial n}{\partial t}(\xi_k)(\xi_k+1)^2\rho_k = -8\theta \sum_{j=0}^{N} \left(\sum_{i=1}^{N-1} l'_k(\xi_i) l'_j(\xi_i) c(\eta(\xi_i))(\xi_i+1)^2 \rho_i \right) n(\xi_j) \\ - 4 \sum_{i=1}^{N-1} l'_k(\xi_i) n(\xi_i) \frac{\partial \psi}{\partial \xi}(\xi_i)(\xi_i+1)^2 \rho_i.\end{aligned}$$

On utilise alors les deux conditions aux bords (1.4c) et (1.4d) que l'on discrétise grâce à l'équation (1.12). On peut ainsi exprimer les deux quantités $n(\xi_0)$ et $n(\xi_N)$ en fonction des $n(\xi_k)$ pour $k \in \{1 \cdots N-1\}$ et on injecte les expressions ainsi obtenues dans (1.14). Cette équation peut ainsi s'écrire sous la forme

$$\frac{\partial n}{\partial t}(\xi_k)(\xi_k+1)^2\rho_k = -8\theta \sum_{j=1}^{N-1} A_{k,j} n(\xi_j) - 4b_k(n,\psi), \tag{1.15}$$

où $b_k(n)$ est le terme non-linéaire

$$b_k(n) = \sum_{i=1}^{N-1} l'_k(\xi_i) n(\xi_i) \frac{\partial \psi}{\partial \xi}(\xi_i)(\xi_i+1)^2\rho_i.$$

On note D la matrice diagonale composée des éléments $(1+\xi_k)^2\rho_k$. Le système formé des équations (1.15) pour $k \in \{1, \cdots, N-1\}$ s'écrit alors

$$D\frac{\partial X}{\partial t} = -8\theta AX - 4B(X,\Psi). \tag{1.16}$$

Quant à la deuxième équation (1.8) de la formulation variationnelle, on la discrétise en

$$\sum_{j=0}^{N-1} \left(\sum_{i=1}^{N} l'_k(\xi_i) l'_j(\xi_i)(\xi_i+1)^2\rho_i \right) \psi(\xi_j) = -\pi n(\xi_k)(\xi_k+1)^2\rho_k. \tag{1.17}$$

En discrétisant la condition aux bords (1.4a) grâce à l'équation (1.13), on obtient une expression de $\psi(\xi_0)$ en fonction des $\psi(\xi_k)$ pour $1 \leq k \leq N-1$; on remplace alors cette expression dans (1.17).

Le vecteur Ψ est alors solution d'un système linéaire que l'on peut écrire :

$$C\Psi = -\pi DX. \tag{1.18}$$

Expliquons enfin la discrétisation en temps et le déroulement du programme à partir des équations (1.16) et (1.18).

Soit Δt le pas de temps. On suppose que l'on connaît le vecteur densité X aux temps $\Delta t, \cdots, m\Delta t$ et on souhaite le calculer au temps $(m+1)\Delta t$. On notera X^m et Ψ^m les vecteurs approximations de la densité et du vecteur potentiel au temps $m\Delta t$.

On commence par calculer le vecteur potentiel au temps $m\Delta t$ en résolvant le système (1.18) où la densité du second membre est évaluée de façon explicite, c'est-à-dire

$$C\Psi^m = -\pi DX^m.$$

On calcule alors la dérivée de ψ grâce à l'équation (1.13).

Ensuite, pour pouvoir calculer le terme $c(\eta(\xi_j))$ de (1.14), on résout en chaque point ξ_j l'équation

$$J_{1/2}(\eta(\xi_j)) = \frac{4\pi}{\mu\theta^{3/2}} n(\xi_j)$$

grâce à la méthode de Newton. La dérivée de $J_{1/2}(\eta)$ s'exprime très facilement en fonction de $J_{-1/2}$ d'après l'équation (1.1).

Finalement, on calcule la densité au temps $(m+1)\Delta t$ grâce au système (1.16) avec une évaluation implicite du terme linéaire du second membre et une évaluation explicite du terme non-linéaire $b(n)$, c'est-à-dire on résout le système

$$(D + 8\theta\Delta t A)X^{m+1} = DX^m - 4\Delta t B(X^m, \Psi^m).$$

2. Effondrements, explosions et hystérésis : étude numérique de systèmes auto-gravitants de fermions

Thermodynamics, collapses, explosions and hysteresis in self-gravitating systems

Pierre-Henri Chavanis, Magali Ribot, Carole Rosier and Clément Sire

Abstract : *The purpose of this paper is two-fold. Firstly, we review different kinds of kinetic equations that have been proposed to describe the dynamics of self-gravitating systems (stellar systems, self-gravitating Brownian particles, gaseous stars,...). We show that these equations have a thermodynamical structure and we emphasize the distinction between a microcanonical and a canonical description. Secondly, we use the simplest dynamical model to illustrate the general features of the thermodynamics of self-gravitating systems. Specifically, we numerically solve the Smoluchowski-Poisson system for self-gravitating Brownian fermions and illustrate the notions of collapse, explosion and hysteresis that are predicted by mean-field theory. We also mention the connexion of our work with the chemotactic aggregation of bacterial populations in biology.*

2.1. Introduction. Self-gravitating systems have a very special thermodynamics due to the unshielded long-range nature of the gravitational force (Padmanabhan [**88**]). These problems were first discussed by Antonov [**2**] and Lynden-Bell [**80**] in the case of classical point mass stars. Antonov discovered the absence of entropy maximum below a critical energy and Lynden-Bell interpreted the resulting instability as a "gravothermal catastrophe" (Lynden-Bell & Wood [**83**]). When a small-scale cut-off is introduced, e.g. hard spheres or quantum degeneracy, self-gravitating systems exhibit different kinds of phase transitions (of zeroth, first and second order) between gaseous and condensed states. These phase transitions occur both in canonical and microcanonical ensemble and the phase diagram versus cut-off parameter displays two critical points, one in each ensemble (Chavanis [**29**], Chavanis & Ispolatov [**33**], Chavanis & Rieutord [**35**]).

In the present paper, we propose to illustrate these phase transitions with a simple model of gravitational dynamics. Namely, we consider a model of self-gravitating Brownian fermions in a strong friction limit where we can neglect the inertia of the particles. Quantum degeneracy is introduced to provide an effective small-scale cut-off and properly define a condensed phase. We shall solve therefore the fermionic Smoluchowski-Poisson system. This is a *toy model* of self-gravitating systems, but it exhibits exactly the same phenomena (collapses, explosions and hysteresis) as more realistic models and it has the great

advantage of being amenable to a full analytical or numerical description. In addition, this model has direct applications in biology in the context of the chemotaxis of bacterial populations (Murray [**86**]). Indeed, bacteries such as *E. Coli* secrete a substance that has a long-range attracting effect on the other organisms. Under certain circumstances, this self-attraction can trigger a phenomenon of collapse similar in some respect to the collapse of self-gravitating systems. For bacteries, we can safely neglect inertia so that the evolution of the population is described by the Smoluchowski-Poisson system with a small-scale regularization. Of course, in biology, the small-scale regularization is not due to quantum mechanics. However, our results are relatively independent on the precise form of the cut-off.

In Secs. 2.2 and 2.3, we review different models of gravitational dynamics. This review is important to precisely state the domains of application of these models and avoid confusion and misunderstanding. This is particularly relevant since different communities (applied mathematicians, biologists, statistical physicists,...) are interested by the physics of systems with long-range interactions, for which gravity is of fundamental importance (Dauxois et al [**44**]). We show that all these models (corresponding to stellar systems, self-gravitating Brownian particles, gaseous stars,...) have a thermodynamical structure which corresponds either to a microcanonical or a canonical description. Corresponding H-theorems are derived. This distinction between microcanonical and canonical description is important because, as is well-known, statistical ensembles are inequivalent for self-gravitating systems so that the stability limits differ from one ensemble to the other (e.g., Chavanis [**32**]).

In Sec. 2.4, we specifically consider the fermionic Smoluchowski-Poisson system in a canonical setting (fixed temperature T). In agreement with thermodynamical prediction, we show that this system converges towards a statistical equilibrium state for $T > T_c$ and collapses for $T < T_c$. In the non-degenerate limit, the collapse is self-similar and leads to a finite time-singularity (Chavanis, Rosier & Sire [**40**]). In the case of fermions, the collapse stops when the core of the system becomes degenerate. The resulting "core-halo" structures forming the condensed phase have been calculated previously (Chavanis & Sommeria [**36**], Chavanis [**29**]). If, now, the condensate (fermion ball) is heated above a critical temperature $T_* > T_c$, it undergoes an "explosion", reverse to the collapse, and returns to the gaseous phase. We numerically exhibit an *hysteretic cycle* between the gaseous phase and the condensed phase by varying the temperature between T_* and T_c.

Collapses, explosions and hysteresis also occur in the Hamiltonian N-stars problem (the archetypal model of stellar dynamics). These phenomena are fundamental features of the thermodynamics of self-gravitating systems. They were predicted from mean-field theory (Stahl

et al [**114**], Chavanis & Ispolatov [**33**], Chavanis [**26**]) and have been confirmed numerically by Ispolatov & Karttunen [**67**] for the gravitational N-body problem by using Molecular Dynamics methods. The present approach, considering a different model of gravitational dynamics, exhibits the same phenomena and allows for a more detailed analysis.

2.2. Statistical mechanics of classical self-gravitating systems.

2.2.1. *The N-body problem.* Consider a system of N stars (or classical particles) in gravitational interaction. We assume that the system is isolated so that it conserves mass, energy, angular momentum and impulse. We ignore relativistic effects and consider Newton's theory of gravitation. The equations of motion can be cast in a Hamiltonian form

$$m\frac{d\mathbf{r}_i}{dt}=\frac{\partial H}{\partial \mathbf{v}_i},\qquad m\frac{d\mathbf{v}_i}{dt}=-\frac{\partial H}{\partial \mathbf{r}_i}, \tag{2.1}$$

where H is the Hamiltonian

$$H=\frac{1}{2}\sum_{i=1}^{N}mv_i^2-\sum_{i<j}\frac{Gm^2}{|\mathbf{r}_i-\mathbf{r}_j|}. \tag{2.2}$$

This N-body problem is the correct starting point in the description of globular clusters and elliptical galaxies (Binney & Tremaine [**15**]). When N is large, we can try to describe the evolution of the system in a statistical sense. Since the system is isolated, the relevant ensemble is the *microcanonical ensemble*. The statistical mechanics of self-gravitating systems was initiated by Antonov [**2**] and Lynden-Bell [**80**].

For long-range systems, the thermodynamic limit is unusual (de Vega & Sanchez [**48**], Chavanis & Rieutord [**35**]) and does *not* correspond to $N,V\to+\infty$ at fixed N/V. When the correct thermodynamical limit is taken, the mean-field approximation is exact and the statistical equilibrium state is obtained by maximizing the Boltzmann entropy

$$S=-\int f\ln f d^3\mathbf{r}d^3\mathbf{v}, \tag{2.3}$$

at fixed mass

$$M=\int \rho d^3\mathbf{r}, \tag{2.4}$$

and energy

$$E=\frac{1}{2}\int fv^2d^3\mathbf{r}d^3\mathbf{v}+\frac{1}{2}\int \rho\Phi d^3\mathbf{r}. \tag{2.5}$$

In these equations, $f(\mathbf{r},\mathbf{v})$ is the distribution function and $\rho(\mathbf{r})=\int fd^3\mathbf{v}$ the spatial density. The gravitational potential is related to

the density via the Poisson equation

$$\Delta\Phi = 4\pi G\rho. \tag{2.6}$$

Introducing Lagrange multipliers for each constraint and writing the variational problem in the form

$$\delta S - \beta\delta E - \alpha\delta M = 0, \tag{2.7}$$

we obtain the isothermal distribution

$$f = A' e^{-\beta\epsilon} \Rightarrow \rho = A e^{-\beta\Phi}, \tag{2.8}$$

where $\epsilon = \frac{v^2}{2} + \Phi$ is the energy of a star by unit of mass and $\beta = m/kT$ the inverse temperature. The equilibrium state is obtained by solving the Boltzmann-Poisson equation

$$\Delta\Phi = 4\pi G A e^{-\beta\Phi}, \tag{2.9}$$

and relating the Lagrange multipliers to the constraints (Padmanabhan [**88**], Chavanis [**38**]). The conservation of angular momentum

$$\mathbf{L} = \int f\ \mathbf{r}\times\mathbf{v}\ d^3\mathbf{r} d^3\mathbf{v}, \tag{2.10}$$

can be easily incorporated in the variational principle by introducing a Lagrange multiplier $-\beta\mathbf{\Omega}$ (Chavanis Rieutord [**35**]). Equation (2.9) remains valid provided that ϵ is replaced by the Jacobi energy $\epsilon_J \equiv \epsilon - \mathbf{\Omega}\cdot\mathbf{j}$, where $\mathbf{j} = \mathbf{r}\times\mathbf{v}$ is the angular momentum by unit of mass. We note that $\epsilon_J = \frac{w^2}{2} + \Phi_{eff}$, where $\mathbf{w} = \mathbf{v} - \mathbf{\Omega}\times\mathbf{r}$ is the relative velocity and $\Phi_{eff} = \Phi - \frac{1}{2}(\mathbf{\Omega}\times\mathbf{r})^2$ is the effective potential accounting for inertial forces.

To prevent complete evaporation and make the statistical mechanics well-defined, we need to enclose the system within a spherical box of radius R (otherwise the Boltzmann entropy has no maximum). Physically, the box is a crude model to take into account a spatial confinement of the system due to tidal effects (for globular clusters) or incomplete relaxation (for elliptical galaxies). If the system is confined within a box, local maxima of entropy at fixed mass and energy exist. They correspond to long-lived metastable states. To obtain true equilibrium states (global entropy maxima) a small-scale cut-off must be introduced. In Sec. 2.3, we shall consider quantum effects which provide an effective small-scale cut-off due to Pauli's exclusion principle.

2.2.2. *The Landau equation.* The self-gravitating N-body system achieves two successive equilibrium states. Mathematically, this is due to the non-commutation of the limits $t \to +\infty$ and $N \to +\infty$ (Chavanis [**30, 32**]). For fixed $N \gg 1$ and $t \to +\infty$, the system is expected to reach a statistical equilibrium state. The evolution is due to stellar encounters and this "collisional" regime is described by the Landau

equation

$$\frac{\partial f}{\partial t} + \mathbf{v}\frac{\partial f}{\partial \mathbf{r}} + \mathbf{F}\frac{\partial f}{\partial \mathbf{v}} = \frac{\partial}{\partial v^\mu}\int d^3\mathbf{v}_1\; K^{\mu\nu}\left\{ f_1 \frac{\partial f}{\partial v^\nu} - f\frac{\partial f_1}{\partial v_1^\nu}\right\}, \tag{2.11}$$

where $K^{\mu\nu}$ is the tensor

$$K^{\mu\nu} = \frac{A}{u}\left(\delta^{\mu\nu} - \frac{u^\mu u^\nu}{u^2}\right), \tag{2.12}$$

and $\mathbf{u} = \mathbf{v}_1 - \mathbf{v}$ is the relative velocity (we have denoted by A the quantity $2\pi G^2 m \ln(L_{max}/L_{min})$). This equation conserves mass and energy and satisfies a H-theorem [1] for the Boltzmann entropy (2.3), i.e. $\dot{S} \geq 0$ with $\dot{S} = 0$ at statistical equilibrium (Balescu [**6**]). Therefore, due the the development of encounters between stars, the system is expected to relax towards the statistical equilibrium state (2.8). In fact, this is the case only if the energy is sufficiently high (or the box radius sufficiently low). Above the Antonov limit $\Lambda = -ER/GM^2 = 0.335$, the system undergoes a gravitational collapse called "gravothermal catastrophe". The collapse is self-similar and leads to a finite time singularity (the central density becomes infinite in a finite time). Lancellotti & Kiessling [**76**] consider the full Landau-Poisson system and argue that the density profile behaves as $\rho \sim r^{-\alpha}$ with $\alpha = 3$ at large distances. This differs from the usually reported value $\alpha = 2.21$ obtained with the orbit averaged Fokker-Planck equation (Cohn [**43**]). Lancellotti & Kiessling argue that the orbit averaged Fokker-Planck equation does not describe the collapse accurately. We point out that Lancellotti & Kiessling do not prove that their equation for the scaling profile has solutions. It *may* happen that this equation has no solution. If this is the case, the dimensional arguments leading to $\alpha = 3$ would not be correct. We note also that $\alpha = 3$ leads to an infinite mass in the core of the distribution at the collapse time which is not physical.

2.2.3. *The Vlasov equation.* The timescale of collisional relaxation is given by $t_{relax} \sim \frac{N}{\ln N} t_D$ (Binney & Tremaine [**15**]), where $t_D = (3\pi/16G\rho)^{1/2}$ is the dynamical time. Therefore, this collisional description is well-suited to globular clusters for which $N \sim 10^5$, $t_D \sim 10^5$ years, age $\sim 10^{10}$ years. On the other hand, for elliptical galaxies

[1] We emphasize that the Landau equation is only marginally valid for self-gravitating systems because the diffusion coefficient diverges logarithmically with the distance and there is no evident shielding contrary to the plasma case. In addition, the true kinetic equation for self-gravitating systems is integrodifferential and non-Markovian so that the H-theorem cannot be proved (Kandrup [**69**]). These space and time delocalizations can induce a deviation to the Boltzmann distribution for *intermediate* collisional times. The decay of the diffusion coefficient with the velocity can also generate some anomalies in the tail of the velocity distribution. These complicated effects have not been studied in detail so far.

($N \sim 10^{11}$, $t_D \sim 10^8$ years, age $\sim 10^{10}$ years), the evolution is encounterless and governed by the Vlasov-Poisson system

$$\frac{\partial f}{\partial t} + \mathbf{v}\frac{\partial f}{\partial \mathbf{r}} + \mathbf{F}\frac{\partial f}{\partial \mathbf{v}} = 0, \tag{2.13}$$

$$\Delta\Phi = 4\pi G \int f d^3\mathbf{v}. \tag{2.14}$$

Formally, the Vlasov equation is valid for $t \ll t_{relax}$ and $N \to +\infty$. Since t_{relax} rapidly increases with the number of particles, the domain of validity of the Vlasov regime is extremely long.

The Vlasov-Poisson system is known to develop very complex filaments as a result of a mixing process in phase space. If we introduce a coarse-graining procedure, the coarse-grained distribution function $\overline{f}(\mathbf{r}, \mathbf{v}, t)$ will reach a *metaequilibrium* state $\overline{f}(\mathbf{r}, \mathbf{v})$ on a timescale of a few dynamical times. This process is known as "phase mixing" and "violent relaxation". Lynden-Bell [**81**] has tried to describe this metaequilibrium state in terms of statistical mechanics. In the dilute approximation, he predicts an isothermal distribution function

$$\overline{f} = Ae^{-\beta\epsilon}. \tag{2.15}$$

This distribution function is similar to Eq. (2.8), but it is reached on a much smaller timescale as a result of a collisionless relaxation. Of course, the time dependent coarse-grained distribution function $\overline{f}(\mathbf{r}, \mathbf{v}, t)$ is *not* a solution of the Vlasov equation (2.13) and we can wonder whether it is possible to write down a relaxation equation for it. A kinetic equation can be obtained by developing a quasilinear theory of the Vlasov-Poisson system (Kadomtsev & Pogutse [**68**], Severne & Luwel [**111**], Chavanis [**28**]). In the two-levels approximation of the theory and in the dilute limit, the equation satisfied by $\overline{f}(\mathbf{r}, \mathbf{v}, t)$ is similar to the Landau equation (2.11) but with a different value of the constant A and a completely different physical interpretation.

It can be shown that the functionals

$$S = -\int C(f) d^3\mathbf{r} d^3\mathbf{v}, \tag{2.16}$$

where C is a convex function, are conserved by the Vlasov equation on a fine-grained scale, i.e. $\dot{S}[f] = 0$ at any time. However, the functionals (2.16) increase on a coarse-grained scale in the sense that $S[\overline{f}(t)] \geq S[f(0)]$ (a monotonic increase is not implied) where it is assumed that, at $t = 0$, the distribution function is not mixed, i.e. $\overline{f}(0) = f(0)$. For that reason, the functionals (2.16) are called H-functions (Tremaine, Hénon & Lynden-Bell [**122**]). Boltzmann and Tsallis functionals are particular H-functions leading to isothermal or polytropic distribution functions. It can be shown that the variational

problem

$$\text{Max} \quad S[f] \quad \text{at fixed} \quad E[f],\ M[f], \tag{2.17}$$

where S is any H-function determines a stationary solution of the Vlasov equation of the form $f = f(\epsilon)$ with strong (nonlinear) dynamical stability properties. Such solutions can be achieved as a result of (incomplete) violent relaxation. Since this dynamical stability criterion is similar to the usual thermodynamical stability criterion (maximization of the Boltzmann entropy S_B at fixed mass and energy) we can use a *thermodynamical analogy* to analyze the dynamical stability of collisionless stellar systems (Chavanis [**32**]). In this analogy, the variational problem (2.17) is similar to a microcanonical stability condition.

2.2.4. *The Kramers-Chandrasekhar equation.* In his early work on stellar dynamics, Chandrasekhar [**25**] analyzed the "collisional" evolution of stars in a galaxy by using an analogy with Brownian motion. He considered an infinite and homogeneous medium and introduced the stochastic equations

$$\frac{d\mathbf{r}}{dt} = \mathbf{v}, \qquad \frac{d\mathbf{v}}{dt} = \mathbf{F} - \xi\mathbf{v} + \mathbf{R}(t), \tag{2.18}$$

where $\mathbf{F}$ is the mean gravitational force, $-\xi\mathbf{v}$ is a friction force and $\mathbf{R}(t)$ is a stochastic noise that is δ-correlated in time

$$\langle \mathbf{R}(t)\mathbf{R}(t')\rangle = 6D\delta(t-t'). \tag{2.19}$$

In Chandrasekhar's work, the dynamical friction and the stochastic force result from the encounters between stars (i.e., from the deviation to the smooth gravitational force $\mathbf{F}$ due to the system as a whole). The Fokker-Planck equation associated with the stochastic process (2.18) is the Kramers equation

$$\frac{\partial f}{\partial t} = \frac{\partial}{\partial \mathbf{v}}\left(D\frac{\partial f}{\partial \mathbf{v}} + \xi f\mathbf{v}\right). \tag{2.20}$$

The condition that the Boltzmann distribution (2.8) is a stationary solution of this equation leads to the Einstein relation

$$\xi = D\beta. \tag{2.21}$$

It should be first emphasized that this model cannot describe rigorously the evolution of the N-star system. Indeed, this equation does not conserve energy contrary to the Landau equation. Furthermore, it is not galilean invariant since the friction force is proportional to the velocity $\mathbf{v}$, not the relative velocity $\mathbf{v} - \mathbf{u}(\mathbf{r}, t)$. In fact, in Chandrasekhar's approach, Eq. (2.20) describes the relaxation of a *test* star in a bath of field stars at statistical equilibrium with inverse temperature β and mean velocity $\mathbf{u}(\mathbf{r}) = \mathbf{0}$. In that context, the Kramers equation can be derived from the Landau equation (2.11) in a thermal bath approximation (Balescu [**6**]). The diffusion coefficient can thus be explicitly

evaluated. Since the system is in contact with a heat reservoir, the relevant statistical ensemble is the canonical ensemble.

2.2.5. *Self-gravitating Brownian particles.* We can also consider a system of self-gravitating Brownian particles described by the stochastic equations (2.18)- (2.19). This is a *toy model* of gravitational dynamics which extends the classical Brownian model (Risken [**103**]) to the case of self-interacting particles. In this context, the friction is due to the presence of an inert gas and the stochastic force is due to classical Brownian motion, turbulence, or any other stochastic effect. We do *not* assume that the medium is homogeneous, so that we have to solve the Kramers equation

$$\frac{\partial f}{\partial t}+\mathbf{v}\frac{\partial f}{\partial \mathbf{r}}+\mathbf{F}\frac{\partial f}{\partial \mathbf{v}}=\frac{\partial}{\partial \mathbf{v}}\left(D\frac{\partial f}{\partial \mathbf{v}}+\xi f\mathbf{v}\right), \tag{2.22}$$

coupled to the Poisson equation (2.6). This makes the study much more complicated than usual. Up to date, we do not know any astrophysical application of this model although there could be connexions with the process of planetesimal formation in the solar nebula (see Sec. 2.2.7). Whatever, this model is interesting to develop because it possesses a thermodynamical structure and presents the same features as more realistic models (isothermal distributions, collapse, phase transitions,...). For this model, the relevant ensemble is the canonical one.

To simplify the problem further, we can consider the strong friction limit $\xi\to+\infty$, or equivalently the limit of large times $t\gg\xi^{-1}$. In that approximation, the velocity distribution becomes Maxwellian and the Kramers equation reduces to the Smoluchowski equation

$$\frac{\partial \rho}{\partial t}=\nabla\left[\frac{1}{\xi}(T\nabla\rho+\rho\nabla\Phi)\right]. \tag{2.23}$$

It can be shown that the Kramers equation decreases the free energy $F[f]=E[f]-TS[f]$, i.e. $\dot{F}\leq 0$ and $\dot{F}=0$ at statistical equilibrium (canonical H-theorem). Similarly, the Smoluchowski equation decreases the free energy $F[\rho]$ which is obtained from $F[f]$ by assuming that the velocity distribution is Maxwellian (this is rigorously valid in the strong friction limit). This leads to the expression

$$F=T\int\rho\ln\rho\ d^3\mathbf{r}+\frac{1}{2}\int\rho\Phi\ d^3\mathbf{r}. \tag{2.24}$$

The Kramers-Poisson and Smoluchowski-Poisson systems were first introduced by Chavanis, Sommeria & Robert [**37**] to describe the violent relaxation of collisionless stellar systems on a coarse-grained scale. They were presented as a small-scale parameterization of mixing in the Vlasov-Poisson system. They can also be obtained with the heuristic argument that the strong fluctuations of the gravitational potential during violent relaxation plays the same role as "collisions" (Lynden-Bell [**81**]). In order to account for the conservation of energy (which

is important in the context of violent relaxation), Chavanis et al [**37**] proposed to let the temperature depend on time, i.e. $T = T(t)$, so as to satisfy the energy constraint. They showed that the Kramers and Smoluchowski equations satisfy a H-theorem and that they can be derived from a variational principle called Maximum Entropy Production Principle (MEPP). This is only a very rough model of violent relaxation but it takes into account important features of the process (increase of entropy, virialization,...). It can also be used as a numerical algorithm to compute equilibrium states of the coarse-grained Vlasov-Poisson system.

Of course, if we introduce the stochastic equations (2.18) and (2.19) *ad hoc*, the self-gravitating Brownian gas model is perfectly well-defined mathematically and it possesses a lot of attractive properties. The Smoluchowski-Poisson system has been studied recently by Chavanis, Rosier & Sire [**40**] and Sire & Chavanis [**112, 113, 41**] in the physical literature. Rigorous results have been obtained in parallel in the mathematical literature (see Rosier [**105**] and Biler & Nadzieja [**14**] for a connexion). It is found that the Smoluchowski-Poisson system converges to a statistical equilibrium state for $T > T_c = GMm/2.52R$ and that it collapses and creates a finite time singularity for $T < T_c$. The density profile decreases as r^{-2} at large distances. Very nicely, the invariant density profile can be obtained analytically (Chavanis, Rosier & Sire [**40**]). The Smoluchowski-Poisson system also provides a relevant model for the chemotaxis of bacterial populations (Murray [**86**]). Therefore, studying self-gravitating systems has direct applications in biology!

2.2.6. *Self-gravitating gaseous systems.* We now discuss a third model of gravitational dynamics, namely the case of a self-gravitating gas. The equations of the problem are the Navier-Stokes equations

$$\frac{\partial \rho}{\partial t} + \nabla(\rho \mathbf{u}) = 0, \tag{2.25}$$

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\nabla)\mathbf{u} = -\frac{1}{\rho}\nabla p - \nabla\Phi + \frac{\eta}{\rho}\Delta\mathbf{u} + \frac{1}{\rho}(\zeta + \frac{\eta}{3})\nabla(\nabla\mathbf{u}), \tag{2.26}$$

coupled to the Poisson equation (2.6). These equations are used to analyze the stability of stars and determine their pulsation periods. We assume that the fluid is barotropic so that $p = p(\rho)$. It is easy to check that the free energy

$$F = \int \rho \frac{\mathbf{u}^2}{2} d^3\mathbf{r} + \frac{1}{2}\int \rho\Phi d^3\mathbf{r} + \int \rho \int_0^\rho \frac{p(\rho')}{\rho'^2} d\rho' d^3\mathbf{r}, \tag{2.27}$$

decreases during the evolution, i.e. $\dot{F} \leq 0$ and $\dot{F} = 0$ if, and only if, the system is in hydrostatic equilibrium with exact balance between pressure and gravity. The first term in Eq. (2.27) is the mean kinetic

energy, the second term is the gravitational energy and the third term is the entropy for a barotropic gas (a detailer). For an isothermal gas, $p = \rho T$ (with constant T), so that

$$F = \int \rho \frac{\mathbf{u}^2}{2} d^3\mathbf{r} + \frac{1}{2} \int \rho \Phi d^3\mathbf{r} + T \int \rho \ln \rho d^3\mathbf{r}. \tag{2.28}$$

For the Jeans-Euler equations obtained by setting $\eta = \zeta = 0$ (no viscosity), the free energy is rigorously conserved, i.e. $\dot{F} = 0$ even if the system is not in hydrostatic equilibrium. However, it can be shown that the variational problem

$$\text{Min} \quad F[\rho] \quad \text{at fixed} \quad M[f], \tag{2.29}$$

where F is any F-function of the form (2.27) determines a stationary solution of the Jeans-Euler equations, satisfying the hydrostatic balance $\nabla p = -\rho \nabla \Phi$, with strong (nonlinear) dynamical stability properties. Since this dynamical stability criterion is similar to the usual thermodynamical stability criterion (minimization of the Boltzmann free energy $F_B = E - TS_B$ at fixed mass and temperature) we can use a *thermodynamical analogy* to analyze the dynamical stability of gaseous stars (Chavanis [**32**]). In this analogy, the variational problem (2.29) is similar to a canonical stability condition.

The collapse of the isothermal gas has been described by Penston [**95**] in the case of an infinite domain. He finds that the collapse is self-similar and develops a finite time-singularity. For $T \neq 0$, the density profile decreases as r^{-2}, as in the Brownian model (Chavanis et al. [**40**]).

2.2.7. *The two-fluids model.* An extension of the previous model is provided by the two-fluids model. This model describe the dynamics of dust particles (first fluid) in a gas (second fluid). The two fluids are coupled through a friction force. The self-gravity of the dust particles is explicitly taken into account. Assuming an isothermal equation of state, the equations governing the motion of the dust fluid are

$$\frac{\partial \rho}{\partial t} + \nabla(\rho \mathbf{u}) = 0, \tag{2.30}$$

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\nabla)\mathbf{u} = -\frac{T}{\rho}\nabla\rho - \nabla\Phi - \xi(\mathbf{u} - \mathbf{u}_{gas}), \tag{2.31}$$

where $\mathbf{u}$ is the velocity of the dust component and $\mathbf{u}_{gas}$ the velocity of the gas component. Similar equations should be written for the gas. We shall assume in a first step that the gas is motionless and that the retroaction of the dust particles is negligible. Therefore, we set $\mathbf{u}_{gas} = \mathbf{0}$, thereby obtaining

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\nabla)\mathbf{u} = -\frac{T}{\rho}\nabla\rho - \nabla\Phi - \xi\mathbf{u}. \tag{2.32}$$

The system of equations (2.30) and (2.32) is mathematically well-posed and it has isothermal solutions of the form (2.8) as stationary states. Furthermore, we can easily check that the free energy (2.28) decreases i.e. $\dot{F} \leq 0$ and $\dot{F} = 0$ at hydrostatic equilibrium.

The hydrodynamic equations (2.30) and (2.32) can also be obtained with the Brownian model of Sec. 2.2.5 (Chavanis et al. [**37**]). Considering the hierarchy of moments of the Kramers equation (2.22) and closing this hierarchy by assuming that the velocity distribution is Maxwellian returns precisely Eqs. (2.30) and (2.32). Taking now the strong friction limit $\xi \to +\infty$ (i.e. neglecting the inertial terms in Eq. (2.32)) returns the Smoluchowski equation (2.23). In this limit, the free energy reduces to

$$F = \frac{1}{2}\int \rho\Phi d^3\mathbf{r} + T\int \rho \ln \rho d^3\mathbf{r}, \tag{2.33}$$

since $\mathbf{u} = O(\xi^{-1})$. This expression coincides with Eq. (2.24). Therefore, the SP system can also be justified as a particular limit of a two-fluids model.

This two-fluids model can have astrophysical applications in relation with, e.g., planetesimal formation. One fluid corresponds to the dust particles and the other to the solar nebula. The friction parameter is given by the Stokes or the Epstein law depending on the location in the solar nebula and the size of the particles (e.g., Chavanis [**27**]). Of course, to be realistic, we need to account for the Keplerian rotation of the protoplanetary disk and possibly the presence of vortices (Barge & Sommeria [**7**], Tanga et al. [**120**], Chavanis [**27**]). Numerical simulations of the dust-gas-vortex interactions have been made by Bracco et al. [**19**]. Self-gravity remains to be introduced in this model.

2.3. Statistical mechanics of self-gravitating fermions.

2.3.1. *Equilibrium states.* In order to describe phase transitions in self-gravitating systems, we need to introduce a small-scale regularization to properly define a condensed phase. We could consider the case of hard-spheres (Aronson & Hansen [**3**], Stahl et al. [**114**]) or use a softened potential (Follana & Laliena [**60**], Chavanis & Ispolatov [**33**]) in order to prevent the singularity of the naked r^{-1} potential. Instead, we shall consider the case of quantum particles (fermions) because the Pauli exclusion principle is a fundamental concept in physics and self-gravitating fermions have several applications in astrophysics (white dwarf and neutron stars, violent relaxation of collisionless stellar systems, massive neutrinos in dark matter etc...). Furthermore, the complete phase diagram of self-gravitating fermions has already been calculated (Chavanis [**29**]) so that we can directly use these results.

We briefly recall the basic equations that govern the equilibrium configurations of the self-gravitating Fermi gas at finite temperature.

The Fermi-Dirac distribution is obtained by maximizing the Fermi-Dirac entropy

$$S = -\int \left\{ f \ln f + (\eta_0 - f) \ln(\eta_0 - f) \right\} d^3\mathbf{r} d^3\mathbf{v},$$

at fixed mass, energy and angular momentum (microcanonical description). We assume here that $\mathbf{L} = \mathbf{0}$ so that $\mathbf{\Omega} = \mathbf{0}$. We recall that, for quantum particles, $\eta_0 = (2s+1)m^4/(2\pi\hbar)^3$ is the maximum allowable value of the distribution function fixed by Pauli's exclusion principle (s is the spin). Writing the variational principle as in Eq. (2.7), we get

$$f = \frac{\eta_0}{1 + \lambda e^{\beta\epsilon}}. \tag{2.34}$$

The Fermi-Dirac distribution can also be obtained by minimizing the free energy $F = E - TS$ at fixed T and M (canonical description). Microcanonical and canonical descriptions yield the same equilibrium distributions (first order variations) but the stability of the solution (second order variations) differs from one ensemble to the other (Chavanis [**29**]). The same variational problem arises in the statistical theory of violent relaxation for a completely different reason (Lynden-Bell [**81**]). In the two-levels approximation $f = \{0, \eta_0\}$ of the theory, the coarse-grained distribution function predicted by Lynden-Bell coincides with the Fermi-Dirac distribution with another interpretation of the degeneracy (Chavanis & Sommeria [**36**]).

We shall now extend the kinetic theories described in Sec. 2.2 to the case of quantum particles. This will provide new dynamical models that can be used to study the nature of phase transitions in self-gravitating systems. Only the main steps of the derivations will be given. The reader is refered to the papers of Kaniadakis [**70**] and Chavanis [**31, 39**] for more details and generalizations.

2.3.2. *The fermionic Landau equation.* Let us first consider an isolated system of self-gravitating fermions. We consider, as a starting point, the fermionic Boltzmann equation

$$\begin{aligned} \frac{df}{dt} &= \int d^3\mathbf{v}_1 d\Omega \; w(\mathbf{v} + \frac{\mathbf{\Delta}}{2}, \mathbf{v}_1 - \frac{\mathbf{\Delta}}{2}; \mathbf{\Delta}) \\ &\times \left\{ f'(\eta_0 - f) f_1'(\eta_0 - f_1) - f(\eta_0 - f') f_1(\eta_0 - f_1') \right\}, \end{aligned} \tag{2.35}$$

where $d\Omega$ is the element of solid angle and w the transition probability depending on the form of interaction (we have noted $f = f(\mathbf{r}, \mathbf{v}, t)$, $f_1 = f(\mathbf{r}, \mathbf{v}_1, t)$, $f' = f(\mathbf{r}, \mathbf{v}', t)$, $f_1' = f(\mathbf{r}, \mathbf{v}_1', t)$). The Boltzmann equation appropriate to quantum particles (fermions) was first introduced by Uehling & Uhlenbeck [**123**].

If the potential of interaction is the gravitational potential, we can implement a weak deflexion approximation. Expanding the Boltzmann

equation in Taylor series for $|\Delta\mathbf{v}| \ll |\mathbf{v}|, |\mathbf{v}_1|$, we obtain the fermionic Landau equation (Chavanis [**39**])

$$\frac{df}{dt} = \frac{\partial}{\partial v^\mu}\int d^3\mathbf{v}_1\ K^{\mu\nu} \times\left\{ f_1(\eta_0 - f_1)\frac{\partial f}{\partial v^\nu} - f(\eta_0 - f)\frac{\partial f_1}{\partial v_1^\nu}\right\}, \tag{2.36}$$

where $K^{\mu\nu}$ is the tensor defined in Eq. (2.12). The Boltzmann and the Landau equations have a microcanonical thermodynamic structure. They conserve mass and energy and increase the Fermi-Dirac entropy (2.34), i.e. $\dot{S} \geq 0$ and $\dot{S} = 0$ at statistical equilibrium.

We also note that in the two-levels approximation of the quasilinear theory of violent relaxation (Kadomtsev & Pogutse [**68**], Severne & Luwel [**111**], Chavanis [**28**]), the coarse-grained distribution function $\overline{f}(\mathbf{r}, \mathbf{v}, t)$ satisfies a fermionic Landau equation of the form (2.35) with a different constant A and a completely different physical interpretation.

2.3.3. *The fermionic Kramers equation.* We consider the Brownian model of Sec. 2.3.3 except that now we need to account for the Pauli exclusion principle due to the quantum nature of the particles (fermions). We start from the Master equation

$$\frac{\partial f}{\partial t} = \int [\pi(t, \mathbf{v}' \to \mathbf{v}) - \pi(t, \mathbf{v} \to \mathbf{v}')]d^3\mathbf{v}', \tag{2.37}$$

where $\pi(t, \mathbf{v} \to \mathbf{v}')$ is the transition probability from $\mathbf{v}$ to $\mathbf{v}'$. The Master equation is the general starting point for all Markovian stochastic processes. If the particles are fermions, the transition rate can be written $\pi(t, \mathbf{v} \to \mathbf{v}') = w(t, \mathbf{v}, \Delta\mathbf{v})f(\mathbf{v})(\eta_0 - f(\mathbf{v}'))$ where $\Delta\mathbf{v} = \mathbf{v}' - \mathbf{v}$ is the velocity deviation and the term $\eta_0 - f(\mathbf{v}')$ accounts for the exclusion principle. We shall now assume that $|\Delta\mathbf{v}| \ll |\mathbf{v}|, |\mathbf{v}_1|$ (diffusion approximation) and expand the Master equation in Taylor series. To second order in the expansion, we get the fermionic Fokker-Planck equation

$$\frac{\partial f}{\partial t} = \frac{\partial}{\partial v^\mu}\left[\zeta^{\mu\nu}\frac{\partial f}{\partial v^\nu} + f(\eta_0 - f)\left(\frac{\partial \zeta^{\mu\nu}}{\partial v^\nu} + \zeta^\mu\right)\right], \tag{2.38}$$

where

$$\zeta^\mu = -\frac{\langle \Delta v^\mu\rangle}{\Delta t} = -\int w(t, \mathbf{v}, \Delta\mathbf{v})\Delta v^\mu d(\Delta\mathbf{v}), \tag{2.39}$$

$$\zeta^{\mu\nu} = \frac{1}{2}\frac{\langle \Delta v^\mu \Delta v^\nu\rangle}{\Delta t} = \int w(t, \mathbf{v}, \Delta\mathbf{v})\Delta v^\mu \Delta v^\nu d(\Delta\mathbf{v}), \tag{2.40}$$

are the first (friction) and second (diffusion) moments of the velocity deviation. For the Ornstein-Uhlenbeck stochastic process (2.18) (2.19), the moments (2.39) and (2.40) can be easily evaluated and we obtain

the fermionic Kramers equation

$$\frac{\partial f}{\partial t} = \frac{\partial}{\partial \mathbf{v}}\left[D\frac{\partial f}{\partial \mathbf{v}} + \xi f(\eta_0 - f)\mathbf{v}\right]. \tag{2.41}$$

The condition that the Fermi-Dirac distribution is a stationary solution of Eq. (2.41) requires that the diffusion and the friction are related by an Einstein relation

$$\xi = D\beta. \tag{2.42}$$

This Brownian model corresponds to a canonical description in which the temperature is related to the strength of the stochastic force $\mathbf{R}(t) \sim T^{1/2}$. More precisely, Eq. (2.41) describes the relaxation of a *test* particle in a bath of field particles at statistical equilibrium with the Fermi-Dirac distribution function. In that context, the fermionic Kramers equation can be derived from the fermionic Landau equation (2.36) in a thermal bath approximation (Chavanis [**39**]). The diffusion coefficient can thus be explicitly evaluated.

2.3.4. *Self-gravitating Brownian fermions.* We now consider a gas of self-gravitating Brownian fermions as a *toy model* of gravitational dynamics. We thus consider the fermionic Kramers equation

$$\frac{\partial f}{\partial t} + \mathbf{v}\frac{\partial f}{\partial \mathbf{r}} + \mathbf{F}\frac{\partial f}{\partial \mathbf{v}} = \frac{\partial}{\partial \mathbf{v}}\left[D\left(\frac{1}{\eta_0 - f}\frac{\partial f}{\partial \mathbf{v}} + \beta f\mathbf{v}\right)\right], \tag{2.43}$$

coupled to the Poisson equation (2.6). For simplicity, we shall assume that the diffusion coefficient D is constant and set $\xi = D\beta$ (friction coefficient). We write the fermionic Kramers equation under the form (2.43), instead of Eq. (2.41), because the derivation of the Smoluchowski equation is more direct in that case (see Lemou et al. [**78**] for a generalization). It is straightforward to check that the fermionic Kramers equation (2.43) decreases the free energy

$$\begin{aligned} F = \frac{1}{2}\int f v^2 d^3\mathbf{r} d^3\mathbf{v} + \frac{1}{2}\int \rho\Phi d^3\mathbf{r} \\ +T\int\left\{f\ln f + (\eta_0 - f)\ln(\eta_0 - f)\right\}d^3\mathbf{r}d^3\mathbf{v}, \end{aligned} \tag{2.44}$$

constructed with the Fermi entropy (2.34), i.e. $\dot{F} \leq 0$ and $\dot{F} = 0$ at statistical equilibrium.

Starting from the fermionic Kramers equation (2.43), we can obtain hydrodynamical equations by writing the hierarchy of moments equations for f. We close this hierarchy of moment equations by making a local thermodynamical equilibrium (LTE) approximation, i.e. we assume that the distribution function is locally given by

$$f = \frac{\eta_0}{1 + \lambda(\mathbf{r}, t)e^{\beta\frac{w^2}{2}}}, \tag{2.45}$$

where $\mathbf{w} = \mathbf{v} - \mathbf{u}(\mathbf{r}, t)$. This distribution function minimizes the *local* Fermi-Dirac free energy at fixed density $\rho(\mathbf{r}, t)$, local velocity $\mathbf{u}(\mathbf{r}, t)$ and inverse temperature β. With this closure approximation, we get the damped Euler-Jeans equations (Chavanis et al [**37**])

$$\frac{\partial \rho}{\partial t} + \nabla(\rho \mathbf{u}) = 0, \tag{2.46}$$

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\nabla)\mathbf{u} = -\frac{1}{\rho}\nabla p - \nabla\Phi - \xi\mathbf{u}, \tag{2.47}$$

where $\rho = \int f d^3\mathbf{v}$ is the density and $p = \frac{1}{3}\int f w^2 d^3\mathbf{v}$ is the pressure. Using the LTE (2.45), we find that the equation of state $p = p(\rho)$ for a gas of fermions is given by the parametric equations

$$\rho = 4\sqrt{2}\pi\eta_0 T^{3/2} I_{1/2}(\lambda), \tag{2.48}$$

$$p = \frac{8}{3}\sqrt{2}\pi\eta_0 T^{5/2} I_{3/2}(\lambda), \tag{2.49}$$

where we have introduced the Fermi integrals

$$I_n(t) = \int_0^{+\infty} \frac{x^n}{1 + te^x} dx. \tag{2.50}$$

Using the LTE, we also note that the energy and the entropy can be expressed in terms of the hydrodynamical variables as

$$E = \frac{1}{2}\int \rho \mathbf{u}^2 d^3\mathbf{r} + \frac{3}{2}\int p d^3\mathbf{r} + \frac{1}{2}\int \rho\Phi d^3\mathbf{r}, \tag{2.51}$$

$$S = \int \rho \ln \lambda d^3\mathbf{r} + \frac{5}{2}\beta \int p d^3\mathbf{r}. \tag{2.52}$$

The damped Euler-Jeans equations decrease the free energy $F = E - TS$, or explicitly,

$$\begin{aligned} F = -T\int \rho \ln \lambda d^3\mathbf{r} - \int p d^3\mathbf{r} \\ + \frac{1}{2}\int \rho\Phi d^3\mathbf{r} + \frac{1}{2}\int \rho \mathbf{u}^2 d^3\mathbf{r}. \end{aligned} \tag{2.53}$$

We can explicitly check that $\dot{F} \leq 0$ with $\dot{F} = 0$ at statistical equilibrium (Chavanis et al [**37**]).

We now consider the large time limit (or strong friction limit $\xi \to +\infty$). Neglecting the inertial terms in Eq. (2.47) and substituting the resulting equation for $\mathbf{u}$ in Eq. (2.46), we obtain the fermionic Smoluchowski equation (Chavanis et al [**37**])

$$\frac{\partial \rho}{\partial t} = \nabla\Big[\frac{1}{\xi}(\nabla p + \rho\nabla\Phi)\Big]. \tag{2.54}$$

The fermionic Smoluchowski decreases the free energy (2.53) with $\mathbf{u} = \mathbf{0}$.

In fact, the fermionic Kramers equation (2.43) is a particular case of a generalized class of Fokker-Planck equations considered by Kaniadakis [**70**] and Chavanis [**31, 39**]. Generalized damped Euler-Jeans equations can be derived from these equations. They extend Eqs. (2.46) and (2.47) to an arbitrary barotropic equation of state $p = p(\rho)$. Quite generally, the free energy can be written

$$F = \int \rho \int_0^\rho \frac{p(\rho')}{\rho'^2} d\rho' d^3\mathbf{r} + \frac{1}{2}\int \rho\Phi d^3\mathbf{r} + \int \rho \frac{\mathbf{u}^2}{2} d^3\mathbf{r}. \tag{2.55}$$

For the Fermi-Dirac equation of state (2.48) (2.49), this expression is equivalent to Eq. (2.53). The damped Jeans-Euler equations (2.46) and (2.47) can also be obtained in the two-fluids model for an arbitrary equation of state (see sec. 2.2.7). Finally, taking the strong friction limit, we obtain the generalized Smoluchowski equation (2.54) for an arbitrary equation of state (Chavanis [**31**]). It is associated to the free energy (2.55) with $\mathbf{u} = \mathbf{0}$.

2.3.5. *Maximum Entropy Production Principle.* We finally show that the Kramers and Smoluchowski equations can be obtained from a general variational principle called Maximum Entropy Production Principle (Chavanis et al [**37**]). We present here the canonical version of this principle, using the free energy instead of the entropy.

Let us consider a conservative equation of the form

$$\frac{\partial f}{\partial t} + \mathbf{v}\frac{\partial f}{\partial \mathbf{r}} + \mathbf{F}\frac{\partial f}{\partial \mathbf{v}} = -\frac{\partial \mathbf{J}_f}{\partial \mathbf{v}}, \tag{2.56}$$

where $\mathbf{J}_f$ is an unknown current to be determined. Using Eqs. (2.56) and (2.34), we find that the time variation of the free energy (2.34) can be expressed as

$$\dot{F} = \int \mathbf{J}_f \left[T\frac{\partial}{\partial \mathbf{v}} \ln\left(\frac{f}{\eta_0 - f}\right) + \mathbf{v}\right]. \tag{2.57}$$

We consider $\dot{F}$ as a functional of $\mathbf{J}_f$ for fixed f (at any time). Minimizing $\dot{F}$ with the physical constraint

$$\int \frac{J_f^2}{2f} d^3\mathbf{r} d^3\mathbf{v} \leq C(\mathbf{r}, \mathbf{v}, t), \tag{2.58}$$

preventing the diffusion current from taking arbitrarily large values, we obtain the fermionic Kramers equation

$$\frac{df}{dt} = \frac{\partial}{\partial \mathbf{v}}\left\{ D\left[\frac{\partial f}{\partial \mathbf{v}} + \beta f(\eta_0 - f)\mathbf{v}\right]\right\}, \tag{2.59}$$

where $D(\mathbf{r}, \mathbf{v}, t)$ is a Lagrange multiplier accounting for (2.58). The Einstein relation $\xi = D\beta$ is automatically satisfied by the MEPP. The canonical H-theorem $\dot{F} \leq 0$ directly follows from Eqs. (2.57) and (2.59).

We note that Eq. (2.59) can be written

$$\frac{df}{dt} = \frac{\partial}{\partial \mathbf{v}}\left\{ D\beta f(\eta_0 - f)\frac{\partial}{\partial \mathbf{v}}\frac{\delta F}{\delta f}\right\}, \tag{2.60}$$

where $\delta F/\delta f$ is the functional derivative of $F[f]$. Generalized Kramers equations can be obtained by the same procedure, considering a larger class of free energy functionals (Chavanis [**31**]).

We can also derive the Smoluchowski equation from the MEPP. Let us consider a conservative equation of the form

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot \mathbf{J}_\rho, \tag{2.61}$$

where $\mathbf{J}_\rho$ is an unknown current to be determined. Using Eq. (2.61), we find that the time variation of the free energy (2.55), with $\mathbf{u} = \mathbf{0}$ can be written

$$\dot{F} = \int \frac{\mathbf{J}_\rho}{\rho}(\nabla p + \rho\nabla\Phi)d^3\mathbf{r}. \tag{2.62}$$

Minimizing $\dot{F}$ with the physical constraint

$$\int \frac{J_\rho^2}{2\rho}d^3\mathbf{r} \leq C(\mathbf{r}, t), \tag{2.63}$$

preventing the diffusion current from taking arbitrarily large values, we get

$$\frac{\partial \rho}{\partial t} = \nabla\Big[\frac{1}{\xi}(\nabla p + \rho\nabla\Phi)\Big], \tag{2.64}$$

where $\xi(\mathbf{r}, t)$ is a Lagrange multiplier accounting for constraint (2.63). The canonical H-theorem $\dot{F} \leq 0$ directly follows from Eqs. (2.62) and (2.64). This equation is valid for any equation of state including the case of fermions (Chavanis [**31**]). It can also be extended to the case of rotating bodies (Chavanis et al. [**37**], Lemou et al. [**78**]). Note that Eq. (2.64) can be rewritten

$$\frac{\partial \rho}{\partial t} = \nabla\left[\frac{1}{\xi}\rho\nabla\frac{\delta F}{\delta \rho}\right]. \tag{2.65}$$

Since Eq. (2.64) decreases the F-function (2.27) while conserving mass, it can be used as a *numerical algorithm* to construct arbitrary nonlinearly dynamically stable stationary solutions of the Euler-Jeans equations (see Sec. 2.2.6). This is of great practical interest since stellar models are not easy to construct, especially if we relax the condition of spherical symmetry (in the case of rotating systems). We emphasize that this numerical algorithm guarantees that all computed solutions are nonlinearly dynamically stable via the Euler-Jeans equations. This is at variance with numerical methods that directly solve the hydrostatic equation (with no guarantee of stability). This opens the way to many numerical studies and gives another interest to the equations

introduced in this paper. Similarly, the generalized Kramers equations proposed in Chavanis [**30**] can be used as numerical algorithms to construct a great variety of nonlinearly dynamically stable stationary solutions of the Vlasov equation.

2.4. Phase transitions in self-gravitating systems.

2.4.1. *General considerations.* In this section, we shall discuss the notions of collapse, explosion and hysteresis that occur in self-gravitating systems. We emphasize that these phenomena are common to all the dynamical models presented in the previous sections.

In Fig. 1, we reproduce the caloric curve obtained for classical particles without small-scale cut-off. By considering the maximization of the Boltzmann entropy at fixed mass and energy (microcanonical description), Antonov [**2**] noted that no statistical equilibrium state exists below a critical energy $E_c = -0.335GM^2/R$. In that case, the system will undergo gravitational collapse. Since the temperature increases during the contraction of the system, this has been called "gravothermal catastrophe" (Lynden-Bell & Wood [**83**]). Similarly, by considering a self-gravitating gas maintained at fixed temperature T (canonical description), Emden [**59**] and Bonnor [**16**] had previously noted that no hydrostatic equilibrium state exists below $T_c = GMm/2.52R$. Therefore, if the system is sufficiently cold, it will experience an "isothermal collapse". This collapse corresponds to the absence of minimum of free energy below T_c. It coincides with the Jeans instability criterion (Chavanis [**38**]).

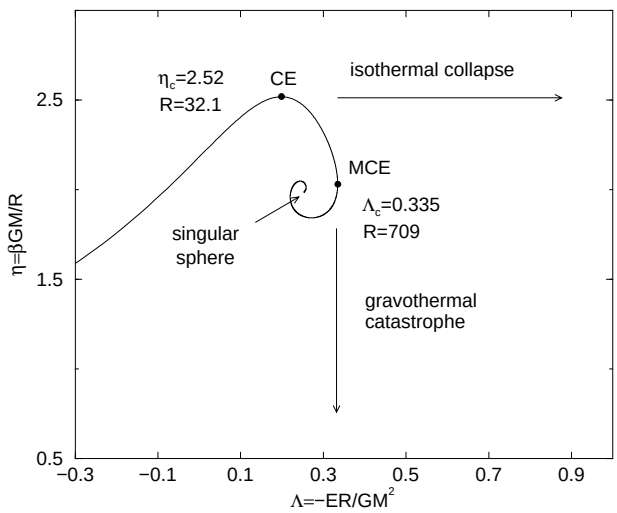


FIGURE 1. Caloric curve for an isothermal self-gravitating gas made of classical particles without small scale cut-off.

The notion of "explosion" is more subtle and can be evidenced only if a small-scale cut-off is introduced so that a condensed phase is well-defined. In Figs. 2 and 3, we reproduce the caloric curves obtained for self-gravitating fermions for different values of the degeneracy parameter (Chavanis [**29**]). The condensed phase exists only below a critical energy $E_*(\mu)$ or below a critical temperature $T_*(\mu)$ depending on the degeneracy parameter μ. Above these critical values, the system explodes. This is the phenomenon reversed to the collapse. Since the critical points (E_c, T_c) and (E_*, T_*) do not coincide an hysteresis occurs in both statistical ensembles (Chavanis [**26**]).

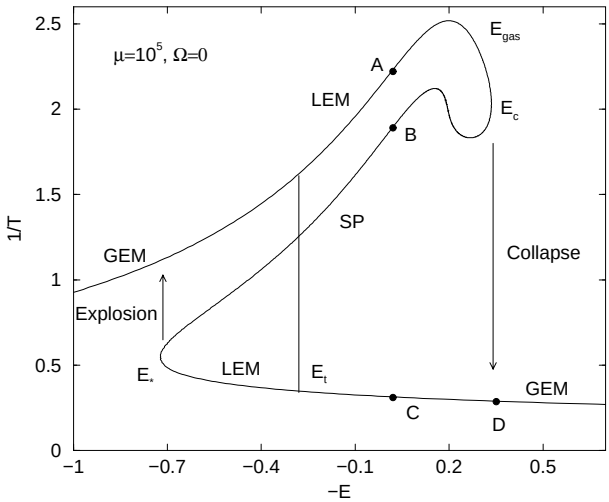


FIGURE 2. Caloric curve for the self-gravitating Fermi gas with a high value of degeneracy parameter $\mu = 10^5$ (small cut-off).

2.4.2. *Previous dynamical studies.* We shall now illustrate dynamically the notions of collapse, explosion and hysteresis that are predicted by statistical mechanics. One possibility would be to solve the N-body problem by introducing small-scale and large-scale cut-offs (see Sec. 2.2.1). The first numerical simulations of gravitational collapse for N-body systems were performed by Hénon [**64**]. He emphasized the notion of "self-similarity" (homology) and the formation of binaries at the end of the collapse. It is surprising that no astrophysicist conducted extensive numerical simulations of the N-stars problem to test the predictions of statistical mechanics [2] since the pioneering works of Hénon, Antonov and Lynden-Bell in the 1960's. Such simulations have

[2]Of course, there are many simulations of N-body systems in astrophysics and cosmology. However, we were not able to find any simulation respecting the strict, albeit artificial, conditions of statistical mechanics, e.g. the confinement of the system in a box.

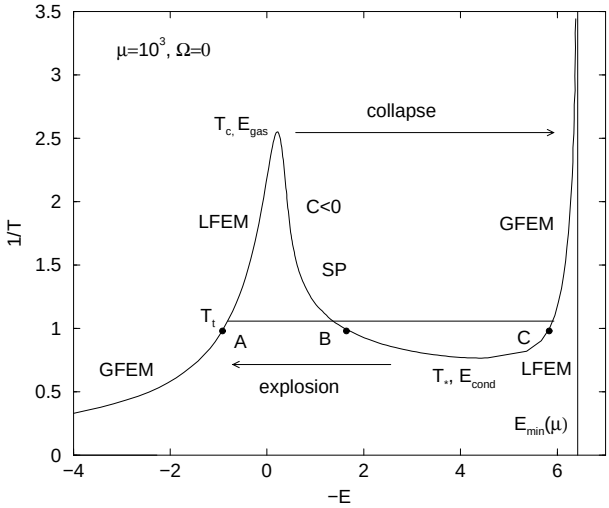


FIGURE 3. Caloric curve for the self-gravitating Fermi gas with a small value of degeneracy parameter $\mu = 10^3$ (high cut-off).

been done only recently by Ispolatov & Karttunen [**67**] using Molecular Dynamics methods. They showed that self-gravitating systems undergo collapses and explosions at the critical energies predicted by the mean-field theory. However, they could afford only a small number of particles ($N \sim 100 - 1000$) in their simulations which makes difficult to study the dynamics of the collapse or explosion in detail.

Another possibility to study the dynamics of stellar systems is to solve kinetic equations appropriate to the N-body system. The Landau-Poisson equation has never been solved numerically. The orbit averaged Fokker-Planck equation has been solved by Cohn [**43**]. He could beautifully follow the self-similar collapse of globular clusters. However, since he works in an infinite domain, there is no statistical equilibrium state in a strict sense and he could therefore not describe metastable states. Larson [**77**] had previously considered the collapse of a box-confined stellar system by solving the moments of the Fokker-Planck equation. He numerically confirmed the transition between equilibrium and collapsed states at the Antonov energy. A complete study of this transition has been made by Chavanis, Rosier & Sire [**40**] with the model of self-gravitating classical Brownian particles.

Concerning fluid systems, Penston [**95**] has solved the Euler-Jeans equations for an isothermal self-gravitating gas. He described the self-similar evolution of the "isothermal collapse". Lynden-Bell & Eggleton [**82**] used fluid equations with a transport of heat and exhibited the dynamics of the "gravothermal catastrophe". However, since they

work in an infinite domain, they could not exhibit bifurcations between metastable gaseous states and collapsed states.

2.4.3. *Model equations.* Since the numerical simulations of the N-body problem are relatively complicated, we shall illustrate the main features of the thermodynamics of self-gravitating systems (collapse, explosion and hysteresis) with simpler equations. We shall use the Brownian model presented in Sec. 2.2.5. In order to properly define a condensed phase, we shall consider the case of Brownian fermions. To simplify the problem further, we take the strong friction limit $\xi \to +\infty$. We emphasize that more realistic models will give the same qualitative results so that, for sake of illustration (which is our point here), we consider the simplest dynamical model. Hence, we solve the fermionic Smoluchowski-Poisson system (2.6)- (2.54) with the equation of state (2.48)- (2.49). The results depend on a dimensionless parameter μ (see below) which is called the "degeneracy parameter" (Chavanis & Sommeria [**36**]). For $\mu \to +\infty$, we recover the case of classical Brownian particles without cut-off (Chavanis, Rosier & Sire [**40**]).

We introduce dimensionless quantities

$$\mathbf{x} = \frac{\mathbf{r}}{R}, \quad n = \frac{\rho}{(M/R^3)}, \quad \psi = \frac{\Phi}{(GM/R)},$$
$$\tau = \frac{t}{(\frac{\xi R^3}{GM})}, \quad P = \frac{R^4}{GM^2}p, \qquad \Lambda = -\frac{ER}{GM^2},$$
$$\theta = \eta^{-1} = \frac{TR}{GM}, \quad \mu = \eta_0\sqrt{512\pi^4 G^3 M R^3}. \tag{2.66}$$

The equations of the problem become

$$\frac{\partial n}{\partial \tau} = \nabla(\nabla P + n\nabla\psi), \quad \Delta\psi = 4\pi n, \tag{2.67}$$

$$n = \frac{\mu}{4\pi}\theta^{3/2} I_{1/2}(\lambda), \quad P = \frac{\mu}{6\pi}\theta^{5/2} I_{3/2}(\lambda), \tag{2.68}$$

$$\int n d^3\mathbf{x} = 1, \quad -\Lambda = \frac{3}{2}\int P d^3\mathbf{x} + \frac{1}{2}\int n\psi d^3\mathbf{x}. \tag{2.69}$$

We shall solve these equations in a spherical box of radius $|\mathbf{x}| = 1$. The boundary condition implied by the conservation of mass is

$$(\nabla P + n\nabla\psi)\cdot\mathbf{n} = 0, \tag{2.70}$$

on the wall ($\mathbf{n}$ is a vector normal to the box). We note that the foregoing equations can be obtained from the dimensional ones by setting $R = G = M = \xi = 1$.

2.4.4. *Asymptotic limits.* The classical limit is recovered by taking $\lambda \to +\infty$ in Eq. (2.68). Using the asymptotic behavior of the Fermi integrals

$$I_n(\lambda) \sim \frac{\Gamma(n+1)}{\lambda}, \quad (\lambda \to +\infty) \tag{2.71}$$

we obtain the equation of state

$$p = \theta n. \tag{2.72}$$

The equations (2.67) with the isothermal equation of state (2.72) have been studied in previous works (Chavanis, Rosier & Sire [**40**] and Sire & Chavanis [**112, 113, 41**]).

The completely degenerate limit is obtained by taking $\lambda \to 0$. Using the asymptotic behavior of the Fermi integrals

$$I_n(\lambda) \sim \frac{1}{n+1}(-\ln\lambda)^{n+1}, \quad (\lambda \to 0) \tag{2.73}$$

we obtain the equation of state

$$p = Kn^{5/3}. \tag{2.74}$$

This corresponds to a polytrope of index 3/2 (white dwarf). The equations (2.67) with the equation of state (2.74) have been studied by Chavanis & Sire [**41**].

2.4.5. *Equation for the mass profile.* We shall restrict ourselves to spherically symmetric solutions in $D = 3$. In that case, the fermionic Smoluchowski equation can be rewritten

$$\frac{\partial n}{\partial \tau} = \frac{1}{x^2}\frac{\partial}{\partial x}\left\{x^2\left(\frac{\partial P}{\partial x} + n\frac{\partial \psi}{\partial x}\right)\right\}. \tag{2.75}$$

On the other hand, in this case of spherical symmetry, the Poisson equation is equivalent to the Gauss theorem

$$\frac{\partial \psi}{\partial x} = \frac{M(x,\tau)}{x^2}, \tag{2.76}$$

where

$$M(x,\tau) = \int_0^x n\ 4\pi x^2 dx, \tag{2.77}$$

is the mass contained within the sphere of radius x at time τ. By simple algebraic manipulations, we see that Eqs. (2.67)- (2.70) are equivalent to a *single* partial differential equation

$$\frac{\partial M}{\partial \tau} = 4\pi x^2 \frac{\partial P}{\partial x} + \frac{1}{x^2} M \frac{\partial M}{\partial x}, \tag{2.78}$$

with the equation of state

$$P = \frac{\mu}{6\pi}\theta^{5/2} I_{3/2}(\lambda), \quad \frac{1}{x^2}\frac{\partial M}{\partial x} = \mu\theta^{3/2} I_{1/2}(\lambda). \tag{2.79}$$

and the boundary conditions

$$M(0,\tau) = 0, \quad M(1,\tau) = 1. \tag{2.80}$$

2.4.6. *Numerical simulations.* We are now ready to follow numerically an hysteretic cycle in self-gravitating systems (see Fig. 4). The phase diagram of self-gravitating fermions versus degeneracy parameter has already been calculated in previous papers (Chavanis [**29**], Chavanis & Rieutord [**35**]). It presents two critical points, one in each ensemble. In the canonical ensemble, a first order phase transition appears at $\mu_{CTP} \simeq 83$. In the microcanonical ensemble, a first order phase transition appears at $\mu_{MTP} \simeq 2670$. We take a degeneracy parameter $\mu = 1000$ intermediate between these two values. The corresponding caloric curve is represented in Fig. 3. For the Brownian model that we consider, the relevant ensemble is the canonical ensemble. Therefore, our control parameter is the temperature T.

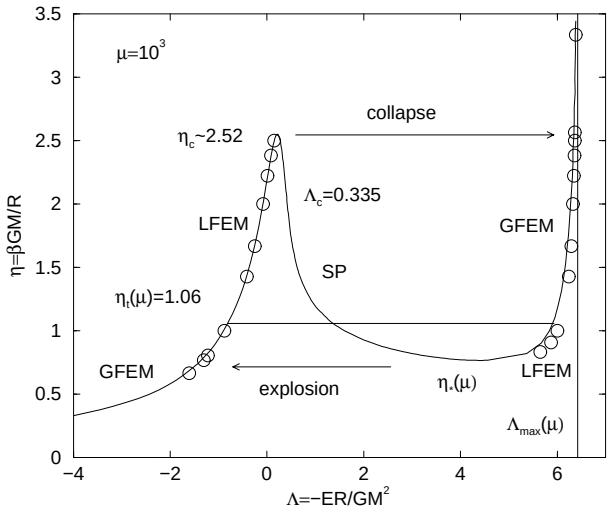


FIGURE 4. Numerically obtained points of the caloric curve for the self-gravitating Fermi gas with $\mu = 10^3$.

Our initial condition consists of a sphere of uniform density at temperature T_0. We start with a high temperature $T_0 = 1.5 > T_* = 1.24$ so that only the gaseous phase exists. Due to self-gravity, the homogeneous distribution is not an equilibrium state and the system quickly relaxes towards an inhomogeneous isothermal distribution with a small density contrast. When we slowly decrease the temperature, we follow a series of equilibria with higher and higher density contrast but no radical change of structure (they correspond to the gaseous states in Fig. 5). In particular, quantum mechanics effects (Pauli exclusion principle) are completely negligible at high temperatures. For $T > T_t$, where T_t is the transition temperature associated to a first order phase transition, the gaseous states are global minima of free energy $F = E - TS$. For

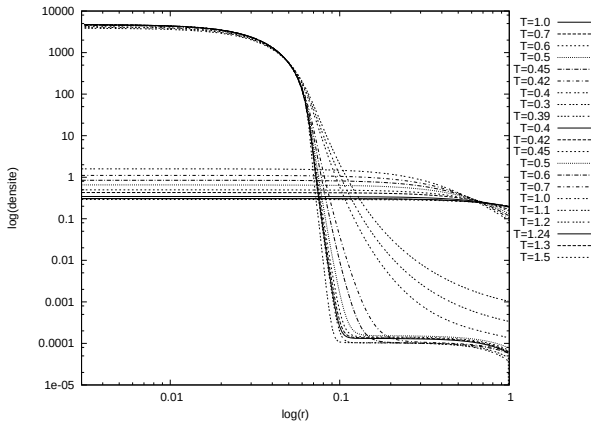


FIGURE 5. Equilibrium states profiles.

$T < T_t$, they become metastable (local minima of free energy). In the true problem where fluctuations are present, the system will finally converge towards the global minimum of free energy (condensed state) but this will take a very long time. Indeed, if $T \gg T_c$ the system has to cross a huge barrier of potential in order to trigger a phase transition to the core-halo phase. The probability of transition decreases exponentially rapidly with the number of particles. Since the Smoluchowski equation is derived in a mean-field limit in which formally $N \to +\infty$ (no fluctuations), the metastable branch has an infinite lifetime (provided that the system is initially prepared in this phase). Therefore, we can cross the "transition temperature" T_t without triggering the canonical first order phase transition.

However, at $T = T_c$, the metastable gaseous branch disappears [3] and the system undergoes an "isothermal collapse" (see Fig. 6). At intermediate times, the collapse is self-similar as in the no cut-off case (see Fig. 7). The density decreases as r^{-2} at large distances (Chavanis, Rosier & Sire [**40**]). When the central density becomes high enough, quantum effects come into play and the system feels the Pauli exclusion principle. Therefore, the gravitational contraction is balanced by quantum pressure (as in a white dwarf star) and the collapse stops. The resulting equilibrium state has a core-halo structure with a degenerate nucleus and an almost uniform envelop held by the box. If we keep decreasing temperature, the core becomes more and more degenerate and contains more and more mass (see Fig. 5). At $T = 0$ all the mass is in the completely degenerate nucleus and the envelop has been swallowed. The system is equivalent to a cold white dwarf star.

[3]In a more realistic problem taking into account fluctuations due to finite N effects, the collapse would occur slightly above T_c. Indeed, as we approach the critical point, the energy fluctuations increase tremendously (they are ideally infinite at T_c) while the barrier of potential decreases so the collapse occurs earlier. This point will be developed in a future work (in preparation).

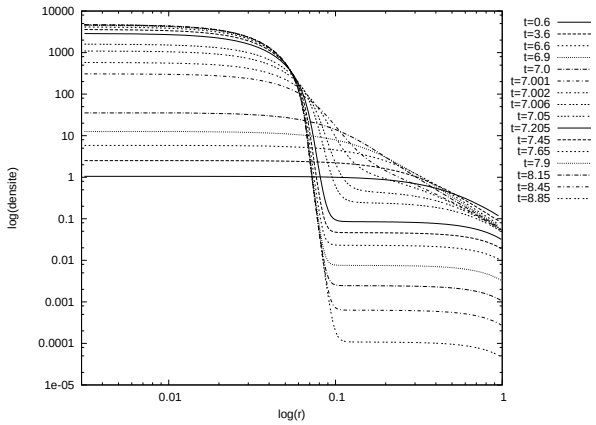


FIGURE 6. Collapse for $T = 0.39$ with cut-off ($\mu = 10^3$).

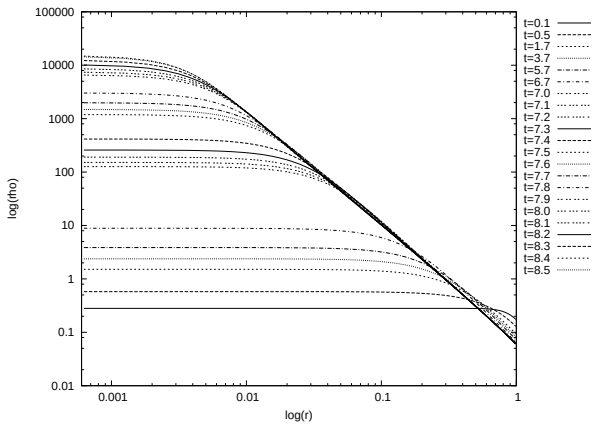


FIGURE 7. Collapse for $T = 0.39$ without cut-off.

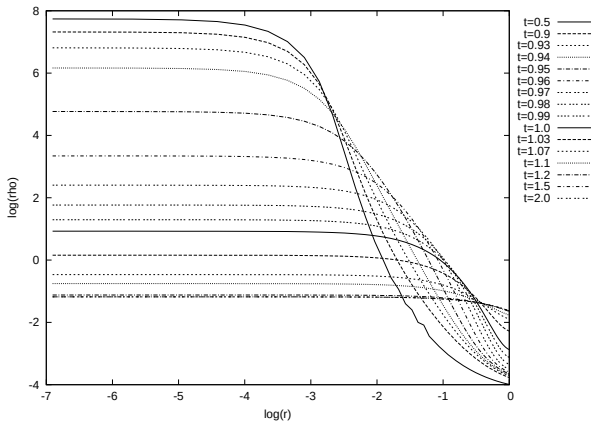


FIGURE 8. Explosion for $T = 1.24$.

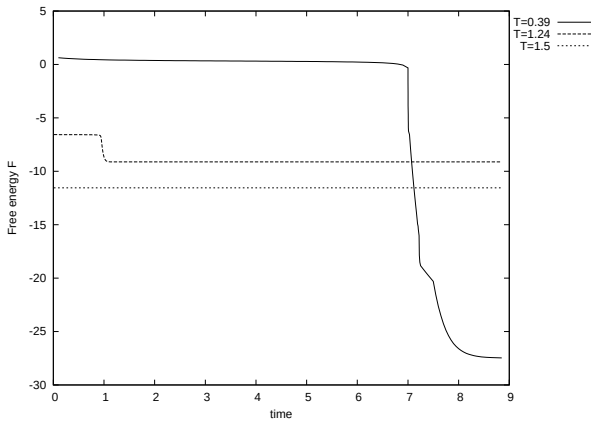


FIGURE 9. Time evolution of the free energy for different values of the temperature ($T = 0.39$, $T = 1.24$ and $T = 1.5$).

If we now increase the temperature, we follow the series of equilibria in the reversed direction. For $T < T_c$, only the condensed phase exists. For $T > T_c$ we remain in the condensed phase so that the collapse at $T = T_c$ is not reversible. For $T < T_t$ the core-halo states are global minima of free energy. For $T > T_t$ they become metastable (local minima) but, for the same reason as before, they are long-lived. In the mean-field limit that we are considering, the system remains in these equilibrium states for all times. However, at $T = T_*$, the condensed phase disappears and the system undergoes an isothermal "explosion" (Fig. 8). This explosion, reversed to the collapse, connects the core-halo phase to the gaseous phase. For higher temperatures, the system is gaseous with a smaller and smaller density contrast up to the uniform distribution as $T \to +\infty$.

Because of the long-time stability of the metastable states, the critical temperatures associated to the collapse and to the explosion do not coincide. This is the reason for the hysteresis phenomenon. Therefore, in the hysteretic region between T_* and T_c, the structure of the system depends on its history (Chavanis & Rieutord [**35**]). We are lead therefore to a notion of *basin of attraction*. Since there exist a local and a global minimum of free energy, the convergence of the system towards one of these states depends on the structure of the initial condition. If the initial condition is relatively uniform, the system will relax towards the gaseous state. Alternatively, if the initial condition is sufficiently peaked it is expected to relax towards the condensed phase. However, there does not seem to be any simple criterion to decide what the evolution will be and the complete structure of this basin of attraction is extremely complicated (Chavanis, Rosier & Sire [**40**]).

Bibliographie

[1] Georgios Akrivis, Michel Crouzeix, and Charalambos Makridakis. Implicit-explicit multistep finite element methods for nonlinear parabolic problems. *Math. Comp.*, 67(222) :457–477, 1998.

[2] V.A. Antonov. *Vest.Leningr.Gos. Univ.*, 7 :135, 1962.

[3] E.B. Aronson and C.J. Hansen. Thermal equilibrium states of a classical system with gravitation. *Astrophysical Journal*, 177 :145, 1972.

[4] A. Averbuch, A. Cohen, and M. Israeli. A stable and accurate explicit scheme for parabolic evolution equations. `http://www.ann.jussieu.fr/~cohen/para.ps.gz`, 1998.

[5] A. V. Balakrishnan. Fractional powers of closed operators and the semigroups generated by them. *Pacific J. Math.*, 10 :419–437, 1960.

[6] R. Balescu. *Statistical Mechanics of Charged Particles.* Interscience, New York, 1963.

[7] P. Barge and J. Sommeria. *Astronomy and Astrophysics*, 295 :1, 1995.

[8] Jöran Bergh and Jörgen Löfström. *Interpolation spaces. An introduction.* Springer-Verlag, Berlin, 1976. Grundlehren der Mathematischen Wissenschaften, No. 223.

[9] Christine Bernardi and Yvon Maday. *Approximations spectrales de problèmes aux limites elliptiques.* Springer-Verlag, Paris, 1992.

[10] Christophe Besse, Brigitte Bidégaray, and Stéphane Descombes. Order estimates in time of splitting methods for the nonlinear Schrödinger equation. *SIAM J. Numer. Anal.*, 40(1) :26–40 (electronic), 2002.

[11] P. Biler. Growth and accretion of mass in an astrophysical model. *Appl. Math. (Warsaw)*, 23(2) :179–189, 1995.

[12] P. Biler, P. Laurençot, and T. Nadzieja. On an evolution system descibing self-gravitating Fermi-Dirac particles. En préparation, 2003.

[13] P. Biler and T. Nadzieja. Growth and accretion of mass in an astrophysical model. II. *Appl. Math. (Warsaw)*, 23(3) :351–361, 1995.

[14] P. Biler and T. Nadzieja. Global and exploding solutions in a model of self-gravitating systems. En préparation, 2002.

[15] J. Binney and S. Tremaine. *Galactic Dynamics.* Princeton Series in Astrophysics, 1987.

[16] W. B. Bonnor. Boyle's law and gravitational instability. *Monthly Not. Roy. Astr. Soc.*, 116 :351–359 (1957), 1956.

[17] Folkmar A. Bornemann. An adaptive multilevel approach to parabolic equations. II. Variable-order time discretization based on a multiplicative error correction. *Impact Comput. Sci. Engrg.*, 3(2) :93–122, 1991.

[18] Folkmar A. Bornemann. An adaptive multilevel approach to parabolic equations. III. 2D error estimation and multilevel preconditioning. *Impact Comput. Sci. Engrg.*, 4(1) :1–45, 1992.

[19] A. Bracco, P.H. Chavanis, A. Provenzale, and E. Spiegel. *Phys. Fluids*, 11 :2280, 1999.

[20] Susanne C. Brenner. Preconditioning complicated finite elements by simple finite elements. *SIAM J. Sci. Comput.*, 17(5) :1269–1274, 1996.

[21] Peter N. Brown and Carol S. Woodward. Preconditioning strategies for fully implicit radiation diffusion with material-energy transfer. *SIAM J. Sci. Comput.*, 23(2) :499–516 (electronic), 2001. Copper Mountain Conference (2000).

[22] Roland Bulirsch and Josef Stoer. Numerical treatment of ordinary differential equations by extrapolation methods. *Numer. Math.*, 8 :1–13, 1966.

[23] Claudio Canuto, M. Yousuff Hussaini, Alfio Quarteroni, and Thomas A. Zang. *Spectral methods in fluid dynamics*. Springer Series in Computational Physics. Springer-Verlag, New York, 1988.

[24] Claudio Canuto and Alfio Quarteroni. Preconditioned minimal residual methods for Chebyshev spectral calculations. *J. Comput. Phys.*, 60(2) :315–337, 1985.

[25] S. Chandrasekhar. Brownian motion, dynamical friction and stellar dynamics. *Rev. Mod. Phys.*, 21 :383, 1949.

[26] P.H. Chavanis. In J. Zinn-Justin, D. Iagolnitzer, and V. Rivasseau, editors, *Proceedings of the TH-2002 conference on Theoretical Physics*, Annales Henri Poincaré.

[27] P.H. Chavanis. Trapping of dust by coherent vortices in the solar nebula. *Astronomy and Astrophysics*, 356 :1089, 2000.

[28] P.H. Chavanis. Statistical mechanics of violent relaxation in stellar systems. In N. Antonic, C.J. van Duijn, Jäger W., and A. Mikelic, editors, *Proceedings of the Conference on Multiscale Problems in Science and Technology*. Springer, 2001. [astro-ph/0212205].

[29] P.H. Chavanis. Phase transitions in self-gravitating systems. Self-gravitating fermions and hard spheres models. *Phys. Rev. E*, 65 :056123, 2002. [cond-mat/0109294].

[30] P.H. Chavanis. Statistical mechanics of two-dimensional vortices and stellar systems. In T. Dauxois, S. Ruffo, E. Arimondo, and Wilkens M., editors, *Dynamics and thermodynamics of systems with long range interactions*, Lecture Notes in Physics. Springer, 2002. [cond-mat/0212223].

[31] P.H. Chavanis. Generalized thermodynamics and Fokker-Planck equations. Applications to stellar dynamics and two-dimensional turbulence. *Phys. Rev. E*, 2003. Sous presse [cond-mat/0209096].

[32] P.H. Chavanis. Gravitational instability of isothermal and polytropic spheres. *Astronomy and Astrophysics*, 401 :15, 2003.

[33] P.H. Chavanis and I. Ispolatov. Phase diagram of self-attracting systems. *Phys. Rev. E*, 66 :036109, 2002. [cond-mat/0205204].

[34] P.H Chavanis, M. Ribot, C. Rosier, and C. Sire. Thermodynamics, collapses, explosions and hysteresis in self-gravitating systems. Soumis.

[35] P.H. Chavanis and M. Rieutord. Statistical mechanics and phase diagrams of rotating self-gravitating fermions. *Astronomy and Astrophysics*, 2003. Sous presse [astro-ph/0302594].

[36] P.H. Chavanis and J. Sommeria. Degenerate equilibrium states of collisionless stellar systems. *Mon. Not. R. Astron. Soc.*, 296 :569–578, 1998.

[37] P.H. Chavanis, J. Sommeria, and R. Robert. Statistical mechanics of two-dimensional vortices and collisionless stellar systems. *Astrophysical Journal*, 471 :385–399, 1996.

[38] Pierre-Henri Chavanis. Gravitational instability of finite isothermal spheres. *Astronomy and Astrophysics*, 381 :340, 2002.

[39] Pierre-Henri Chavanis. Generalized thermodynamics and kinetic equations : Boltzmann, Landau, Kramers and Smoluchowski. *Physica A*, 2003. Sous presse [cond-mat/0304073].

[40] Pierre-Henri Chavanis, Carole Rosier, and Clément Sire. Thermodynamics of self-gravitating systems. *Phys. Rev. E (3)*, 66(3) :036105, 19, 2002.

[41] Pierre-Henri Chavanis and Clément Sire. Anomalous diffusion and collapse of self–gravitating Langevin particles in D dimenions. *Phys. Rev. E*, 2003. Sous presse [cond-mat/0303088].

[42] Chichia Chiu and Noel Walkington. An ADI method for hysteretic reaction-diffusion systems. *SIAM J. Numer. Anal.*, 34(3) :1185–1206, 1997.

[43] H. Cohn. Late core collapse in star clusters and the gravothermal instability. *Astrophys. J.*, 242 :765, 1980.

[44] T. Dauxois, S. Ruffo, E. Arimondo, and M. Wilkens. An introduction. In T. Dauxois, S. Ruffo, E. Arimondo, and Wilkens M., editors, *Dynamics and thermodynamics of systems with long range interactions*, Lecture Notes in Physics. Springer, 2002. [cond-mat/0212223].

[45] P. de Mottoni and M. Schatzman. Development of interfaces in $\mathbf{R}^N$. *Proc. Roy. Soc. Edinburgh Sect. A*, 116(3-4) :207–220, 1990.

[46] Piero de Mottoni and Michelle Schatzman. Geometrical evolution of developed interfaces. *Trans. Amer. Math. Soc.*, 347(5) :1533–1589, 1995.

[47] Piero de Mottoni and Michelle Schatzman. The Thual-Fauve pulse : Skew stabilization. Technical Report 304, Équipe d'Analyse Numérique de Lyon, 1999. `http://numerix.univ-lyon1.fr/publis/publiv/1999/schatz1609/publi.ps.gz`.

[48] H.J. de Vega and N. Sanchez. Statistical mechanics of the self-gravitating gas I : Thermodynamic limit and phase diagrams. *Nucl.Phys.B*, 625 :409, 2002.

[49] Stéphane Descombes. Convergence of a splitting method of high order for reaction-diffusion systems. *Math. Comp.*, 70(236) :1481–1501, 2001.

[50] Stéphane Descombes and Marc Massot. Operator splitting for nonlinear reaction - diffusion systems with an entropic structure : singular perturbation and order reduction. *Numerische Mathematik*, 2004. À paraître.

[51] Stéphane Descombes and Magali Ribot. Convergence of the Peaceman-Rachford approximation for reaction-diffusion systems. *Numer. Math.*, 95(3) :503–525, 2003.

[52] Stéphane Descombes and Michelle Schatzman. Strang's formula for holomorphic semi-groups. *J. Math. Pures Appl. (9)*, 81(1) :93–114, 2002.

[53] M. Deville and E. Mund. Chebyshev pseudospectral solution of second-order elliptic equations with finite element preconditioning. *J. Comput. Phys.*, 60(3) :517–533, 1985.

[54] M. O. Deville and E. H. Mund. Finite-element preconditioning for pseudospectral solutions of elliptic problems. *SIAM J. Sci. Statist. Comput.*, 11(2) :311–342, 1990.

[55] B.O. Dia and M. Schatzman. On the order of extrapolation of integration formulae. Technical Report 275, Équipe d'Analyse Numérique de Lyon, 1998. `http://maply.univ-lyon1.fr/publis/publiv/1998/dia190298/texte.ps`.

[56] Boun Oumar Dia. *Méthodes de directions alternées d'ordre élevé en temps.* PhD thesis, Université Claude-Bernard Lyon 1, France, Juin 1996.

[57] Boun Oumar Dia and Michelle Schatzman. An estimate of the Kac transfer operator. *J. Funct. Anal.*, 145(1) :108–135, 1997.

[58] Jim Douglas, Jr. On the numerical integration of $\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 = \partial u/\partial t$ by implicit methods. *J. Soc. Indust. Appl. Math.*, 3 :42–65, 1955.

[59] R. Emden. *Gazkugeln.* Teubner Verlag, Berlin, 1907.

[60] E. Follana and V. Laliena. Thermodynamics of self-gravitating systems with softened potentials. *Phys. Rev. E*, 61 :6270, 2000.

[61] Luigi Gatteschi. Uniform approximation of Christoffel numbers for Jacobi weight. In *Numerical integration, III (Oberwolfach, 1987)*, volume 85 of *Internat. Schriftenreihe Numer. Math.*, pages 49–59. Birkhäuser, Basel, 1988.

[62] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations. I*, volume 8 of *Springer Series in Computational Mathematics.* Springer-Verlag, Berlin, second edition, 1993. Nonstiff problems.

[63] P. Haldenwang, G. Labrosse, S. Abboudi, and M. Deville. Chebyshev 3-D spectral and 2-D pseudospectral solvers for the Helmholtz equation. *J. Comput. Phys.*, 55(1) :115–128, 1984.

[64] M. Hénon. *Ann ; Astrophysic.*, 1961.

[65] Daniel Henry. *Geometric theory of semilinear parabolic equations.* Springer-Verlag, Berlin, 1981.

[66] Lars Hörmander. *The analysis of linear partial differential operators. I.* Springer Study Edition. Springer-Verlag, Berlin, second edition, 1990. Distribution theory and Fourier analysis.

[67] I. Ispolatov and M. Karttunen. Collapses and explosions in self-gravitating systems. [cond-mat/0302590], 2003.

[68] B.B. Kadomtsev and O.P. Poguste. Collisionless relaxation in systems with Coulomb interactions. *Phys. Rev. E*, 25 :1155, 1970.

[69] H.E. Kandrup. A generalized Landau equation for a system with a self-consistent mean field : derivation from an N-particle Liouville equation. *Astrophys. J.*, 244 :316, 1981.

[70] G. Kaniadakis. *Physica A*, 296 :405, 2001.

[71] Tosio Kato. *Perturbation theory for linear operators.* Springer-Verlag New York, Inc., New York, 1966. Die Grundlehren der mathematischen Wissenschaften, Band 132.

[72] Tosio Kato. *Perturbation theory for linear operators.* Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition.

[73] David Kinderlehrer and Chun Liu. Evolution of grain boundaries. *Math. Models Methods Appl. Sci.*, 11(4) :713–729, 2001.

[74] David Kinderlehrer, Irene Livshits, Florin Manolache, Anthony D. Rollett, and Shlomo Ta'asan. An approach to the mesoscale simulation of grain growth. In M. et al. Aindow, editor, *Influences of interface and dislocation behavior on microstructure evolution*, volume 652, Y1.5. of *Mat. Res. Soc. Symp. Proc.*, 2001.

[75] David Kinderlehrer, Irene Livshits, and Shlomo Ta'asan. A variational approach to modeling and simulation of grain growth. En préparation, 2003.

[76] C. Lancellotti and M. Kiessling. Self-gravitational collapse in stellar dynamics. *Astrophys. J.*, 549 :L93, 2001.

[77] R.B Larson. A method for computing the evolution of star clusters. *Monthly Notices of the Royal Astronomical Society*, 147 :323, 1970.

[78] M. Lemou, Laurençot P., and P.H. Chavanis. Technical report, 2003.

[79] J.-L. Lions and E. Magenes. *Non-homogeneous boundary value problems and applications. Vol. I.* Springer-Verlag, New York, 1972. Translated from the French by P. Kenneth, Die Grundlehren der mathematischen Wissenschaften, Band 181.

[80] D. Lynden-Bell. Runaway centers. In *Extrait du Bulletin Astronomique.* Série 3, Tome III, Fascicule 2 (Éditions du CNRS).

[81] D. Lynden-Bell. Statistical mechanics of violent relaxation in stellar systems. *Monthly Not. Roy. Astr. Soc.*, 136 :101, 1967.

[82] D. Lynden-Bell and P.P. Eggleton. On the consequences of the gravothermal catastrophe. *Mon.Not. R. Astr. Soc.*, 191 :483, 1980.

[83] D. Lynden-Bell and R. Wood. The gravothermal catastrophe in isothermal spheres and the inset of red-giants structure for stellar systems. *Monthly Not. Roy. Astr. Soc.*, 138 :495, 1968.

[84] G. Meurant. *Computer solution of large linear systems*, volume 28 of *Studies in Mathematics and its Applications.* North-Holland Publishing Co., Amsterdam, 1999.

[85] L. S. Mulholland and D. M. Sloan. The role of preconditioning in the solution of evolutionary partial differential equations by implicit Fourier pseudospectral methods. *J. Comput. Appl. Math.*, 42(2) :157–174, 1992.

[86] J. D. Murray. *Mathematical biology.* Springer-Verlag, Berlin, second edition, 1993.

[87] Steven A. Orszag. Spectral methods for problems in complex geometries. *J. Comput. Phys.*, 37(1) :70–92, 1980.

[88] Padmanabhan. Statistical mechanics of gravitating systems. *Phys. Rep.*, 188 :285, 1990.

[89] Seymour V. Parter. On the Legendre-Gauss-Lobatto points and weights. *J. Sci. Comput.*, 14(4) :347–355, 1999.

[90] Seymour V. Parter. Preconditioning Legendre special collocation methods for elliptic problems. I. Finite difference operators. *SIAM J. Numer. Anal.*, 39(1) :330–347 (electronic), 2001.

[91] Seymour V. Parter. Preconditioning Legendre spectral collocation methods for elliptic problems. II. Finite element operators. *SIAM J. Numer. Anal.*, 39(1) :348–362 (electronic), 2001.

[92] Seymour V. Parter and Ernest E. Rothman. Preconditioning Legendre spectral collocation approximations to elliptic problems. *SIAM J. Numer. Anal.*, 32(2) :333–385, 1995.

[93] A. Pazy. *Semigroups of linear operators and applications to partial differential equations*, volume 44 of *Applied Mathematical Sciences.* Springer-Verlag, New York, 1983.

[94] D. W. Peaceman and H. H. Rachford, Jr. The numerical solution of parabolic and elliptic differential equations. *J. Soc. Indust. Appl. Math.*, 3 :28–41, 1955.

[95] M.V. Penston. Dynamics of self–gravitating gaseous spheres iii. analytical results in the free-fall of isothermal cases. *Mon. Not. R. Astr. Soc.*, 144 :425, 1969.

[96] Michael Reed and Barry Simon. *Methods of modern mathematical physics. II. Fourier analysis, self-adjointness.* Academic Press [Harcourt Brace Jovanovich Publishers], New York, 1975.

[97] M. Ribot. Application of the Residual Smoothing Scheme to the preconditioning of spectral methods by finite elements methods. Soumis.

[98] M. Ribot. Asymptotic expansion of Legendre polynomials and of their critical points. Technical Report 359, Équipe d'Analyse Numérique de Lyon, 2002. `http://maply.univ-lyon1.fr/publis/publiv/2002/359/publi.ps.gz`.

[99] M. Ribot and M. Schatzman. Stability, convergence and order of the extrapolations of the Residual Smoothing Scheme in energy norm. Soumis.

[100] M. Ribot and M. Schatzman. Stability of the extrapolations of the Residual Smoothing Scheme. En préparation.

[101] Magali Ribot. A reaction-diffusion approach for grain growth. En préparation.

[102] L.F. Richardson. The approximate solution by finite-differences of physical problems involving differential equations, with an application to the stresses in a mansory dram. *Philos. Roy. Soc. London Ser.*, A 210 :307–357, 1910.

[103] H. Risken. *The Fokker-Planck equation.* Springer, 1989.

[104] Werner Romberg. Vereinfachte numerische Integration. *Norske Vid. Selsk. Forh., Trondheim*, 28 :30–36, 1955.

[105] Carole Rosier. Problème de Cauchy pour une équation parabolique modélisant la relaxation des systèmes stellaires auto-gravitants. *C. R. Acad. Sci. Paris Sér. I Math.*, 332(10) :903–908, 2001.

[106] Yousef Saad. *Iterative Methods for Sparse Linear Systems.* WPS, 1996.

[107] Michelle Schatzman. Stability of the Peaceman-Rachford approximation. *J. Funct. Anal.*, 162(1) :219–255, 1999.

[108] Michelle Schatzman. *Analyse numérique.* Dunod, Paris, 2001. Second edition.

[109] Michelle Schatzman. Numerical integration of reaction-diffusion systems. *Numer. Algorithms*, 31(1-4) :247–269, 2002. Numerical methods for ordinary differential equations (Auckland, 2001).

[110] Michelle Schatzman. Toward non commutative numerical analysis : high order integration in time. In *Proceedings of the Fifth International Conference on Spectral and High Order Methods (ICOSAHOM-01) (Uppsala)*, volume 17, pages 99–116, 2002.

[111] G. Severne and M. Luwel. Dynamic theory of collisionless relaxation. *Astrophys. & Space Sci.*, 72 :293, 1980.

[112] Clément Sire and Pierre-Henri Chavanis. Thermodynamics and collapse of self-gravitating Brownian particles in D dimensions. *Phys. Rev. E (3)*, 66(4) :046133, 23, 2002.

[113] Clément Sire and Pierre-Henri Chavanis. Post–collapse dynamics of self–gravitating brownian particles in d dimensions. [cond-mat/0306406], 2003.

[114] B. Stahl, M.K.H. Kiessling, and K. Schindler. Phase transitions in gravitating systems and the formation of condensed objects. *Plane Space Sci.*, 43 :271, 1994.

[115] R. Stańczy. Steady states for a system describing Fermi–Dirac particles. Technical Report 135, University of Wroclaw, 2003. `http://www.math.uni.wroc.pl/~mathbank/preprinty`.

[116] Gilbert Strang. Accurate partial difference methods. I. Linear Cauchy problems. *Arch. Rational Mech. Anal.*, 12 :392–402, 1963.

[117] Gilbert Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5 :506–517, 1968.

[118] Gábor Szegő. *Orthogonal polynomials*. American Mathematical Society, Providence, R.I., fourth edition, 1975. American Mathematical Society, Colloquium Publications, Vol. XXIII.

[119] S. Ta'asan, P. Yu, I. Livshits, D. Kinderlehrer, and J. Lee. Multiscale modeling and simulation of grain boundary evolution. In *Proc. 44th AIAA/ASME/ASCE/AHS Structures, Structural Dynamics, and Materials Conference*. À paraître.

[120] P. Tanga, A. Babiano, B. Dubrulle, and A. Provenzale. *Icarus*, 121 :158, A996.

[121] O. Thual and S. Fauve. Localized structures generated by subcritical instabilities. *J. Phys, France.*, 49 :1829–1833, 1988.

[122] S. Tremaine, M. Hénon, and D. Lynden-Bell. H-functions and mixing in violent relaxation. *Mon. Not. R. Astr. Soc.*, 219 :285, 1986.

[123] E.A. Uehling and G.E. Uhlenbeck. *Phys. Rev. E*, 43 :552, 1933.

[124] P. J. van der Houwen and B. P. Sommeijer. Improving the stability of predictor-corrector methods by residue smoothing. *IMA J. Numer. Anal.*, 10(3) :361–378, 1990.

[125] Kōsaku Yosida. *Functional analysis*. Springer-Verlag, Berlin, 1995. Reprint of the sixth (1980) edition.

Étude théorique de méthodes numériques pour les systèmes de réaction-diffusion ; application aux équations paraboliques non linéaires et non locales.

Résumé : On s'intéresse dans cette thèse à l'étude de méthodes numériques pour les systèmes de réaction-diffusion. Tout d'abord, on étudie le schéma par régularisation du résidu et ses extrapolations ; ce schéma introduit un préconditionneur en espace lors de la discrétisation en temps. On prouve la stabilité en norme usuelle et la convergence en norme d'énergie de cette méthode et on l'applique au préconditionnement de méthodes spectrales par des méthodes d'éléments finis. Cette application nécessite le calcul d'asymptotiques précises des polynômes de Legendre et de leurs extrema. On prouve aussi la convergence et l'ordre deux d'une méthode de splitting semi-discrétisée en temps pour les systèmes de réaction-diffusion, l'approximation de Peaceman-Rachford. Enfin, on applique ces méthodes à la simulation d'une équation parabolique non linéaire pour modéliser la croissance de grains et à une équation parabolique non locale venant de la mécanique statistique et modélisant les systèmes autogravitants de fermions.

Mots-clés : analyse numérique, méthodes de splitting, préconditionnement, systèmes de réaction-diffusion, asymptotiques de polynômes, croissance de grains, systèmes auto-gravitants de fermions.

Theoretical study of numerical methods for reaction-diffusion systems; application to non linear and non local parabolic equations.

Abstract: We are interested in the study of numerical methods for reaction-diffusion systems. We first consider the Residual Smoothing Scheme and its extrapolations; this scheme uses a spatial preconditioner for the time discretization. We prove the stability of this method for the usual norm and its convergence in energy norm and we apply this scheme to the preconditioning of spectral methods by finite elements methods. For this application, we need to compute precise asymptotic formulas of Legendre polynomials and of their extrema. Then, we study a semi-discretization in time of a splitting scheme, called the Peaceman-Rachford approximation and we show that this scheme is convergent and of order two. Eventually, we apply these methods to the simulation of a parabolic non linear equation modelizing grain growth and to the computation of solutions of a non local parabolic equation coming from statistical mechanics and modelizing the fermionic self-gravitating systems.

Keywords: numerical analysis, splitting methods, preconditioning, reaction -diffusion systems, asymptotics of polynomials, grain growth, fermionic self-gravitating systems.

Discipline : Mathématiques

Laboratoire : MAPLY (Mathématiques APpliquées de LYon), UMR CNRS 5585, Université Claude Bernard Lyon I, 69622 Villeurbanne Cedex.