Analyse des méthodes numériques de simulation et contrôle en chimie quantique

Gabriel Turinici

Applications Scientifiques du Calcul Intensif
ASCI
Laboratoire CNRS - UPR 9029
Bâtiment 506
Université Paris Sud
91405 Orsay Cedex, France

THÈSE

présentée pour l'obtention du titre de Docteur de l'Université Pierre et Marie Curie Spécialité: mathématiques appliquées

par
Gabriel TURINICI

sujet: *Analyse de méthodes numériques de simulation et contrôle en chimie quantique*

Soutenue le 20 décembre 2000 devant un jury composé de:

Rapporteurs :	Claude	LeBris
	Anthony T.	Patera
Examinateurs :	Pierre	Lallemand
	Olivier	Pironneau
	Marie-Madeleine	Rohmer
	Marius	Tucsnak
Directeur de thèse :	Yvon	Maday

A mes parents
A mes grands-parents
A tous ceux qui ont su m'être proches

Remerciements

Même si, par sa nature, une thèse se base sur une volonté et un effort individuel, j'ai eu l'occasion de comprendre personnellement comment elle est rendue possible et se trouve influencée par tous ceux avec qui j'ai interagi sur le plan académique ou personnel pendant ces années de travail. C'est avec plaisir que j'en remercierai ici quelques-uns tout en exprimant mes sincères excuses pour tous ceux que la maladresse (et pas l'oubli!) me fera omettre.

Je voudrais témoigner toute ma gratitude à Pierre Lallemand, Marie-Madeleine Rohmer et Marius Tucsnak pour l'honneur qu'ils m'ont fait de participer au jury de soutenance, à Claude Le Bris et Antony Patera pour le temps sacrifié à l'évaluation de ce travail et à Olivier Pironneau pour avoir présidé le jury.

Je dois mon entrée dans le monde des mathématiques appliquées à la chimie quantique à Yvon Maday. L'avoir comme directeur de thèse c'est une expérience difficile à témoigner sur quelques lignes. Son savoir d'être présent aux moments difficiles et sa façon de comprendre l'homme et non seulement l'apprentis m'ont toujours surpris autant que sa capacité singulière de partager son savoir.

Ma sympathie et mes remerciements s'adressent naturellement à Eric Cancès et Claude Le Bris qui ont su être non seulement des interlocuteurs d'exception mais aussi partenaires de qualité lors des nombreuses discutions ou actions communes. C'est un plaisir de remercier aussi leurs collaborateurs du CERMICS, Mathieu Pilot et Adel Ben Haj Yedder, pour les discutions et les collaborations fructueuses que nos rencontres ont occasionné.

Un merci particulier aux membres de l'Action Concertée Incitative "Contrôle par laser" dirigée par Claude Le Bris: Arne Keller, Benoit Soep, Claude Dion, Frederic Le Quéré, Osman Atabek, Pierre Rouchon... dont les interventions, remarques et questions ont été très instructives et stimulantes.

Travailler dans un environnement aussi particulier que l'ASCI a été une joie et un privilège. Son caractère interdisciplinaire m'a permis d'avoir des discutions très enrichissantes avec les chimistes du laboratoire, Gwang-Hi Jeung, Marc Benard et Marie-Madeleine Rohmer et même d'abuser de leur temps et de leur patience... Je me dois de remercier encore plus vivement Marc Benard et Marie-Madeleine Rohmer pour leur accueil chaleureux lors de mes missions a Strasbourg et pour la collaboration très concrète que ces missions ont développé. Merci également aux thésards de l'ASCI, notamment

Leonardo Baffico, Christophe Prud'homme, Laurent et Gaëlle Jeanfaivre, Francesca Rapetti, Thomas Guignon, Marc Montagnac, Martha Gonzalez... pour l'amitié qu'ils ont su ajouter aux discussions académiques. Un merci particulier à Monique Butin pour son aide et son amabilité pendant ces années.

Cette thèse doit beaucoup au directeur de l'ASCI, Pierre Lallemand, qui a toujours été prêt à donner des solutions constructives à tous les problèmes qu'un thésard peut rencontrer, soit ils académiques, logistiques ou administratifs.

Pendant mes années de thèse j'ai eu souvent l'occasion d'interagir avec les membres du Laboratoire d'Analyse Numérique (LAN) de l'Université de Paris 6. J'ai pu donc apprécier les qualitées aussi bien académiques et humaines de Olivier Pironneau, Christine Bernardi, Benoit Perthame, Frédéric Hecht, Yann Brenier, Doina Cioranescu, Albert Cohen, Liliane Ruprecht, Daniéle Boulic, Christian David, Stéphane Del Pino, Basarab Matei, Paul Métier ...

Je pense également à Marius Tucsnak et Cheng-Zhong Xu d'INRIA-Lorraine avec qui j'ai eu des interactions très enrichissantes et dont les conseils se sont avérés très fructueux.

Un merci particulier à Antony Patera pour les discussions et les idées qu'il a partagé avec moi pendant quelques rencontres qu'on a eu lors de ses séjours a Paris.

La collaboration continue que j'ai eu avec Herschel Rabitz pendant les dernières années de thèse a été très encourageante et fructueuse. Son regard sur la science, ses qualitées extra académiques et sa disponibilité font de lui un interlocuteur précieux et distingué. Merci aussi à Wusheng Zhu, de l'équipe de Herschel Rabitz, qui a su répondre avec beaucoup de patience et méthode à mes questions impétueuses sur le contrôle quantique.

Enfin, mes remerciements s'adressent à ma famille, à mes amis et à Monica qui ont su aider là où eux seulement peuvent le faire ...

Résumé. La chimie quantique est un domaine de recherche de plus en plus présent dans les préoccupations des mathématiciens appliqués. S'inscrivant dans cet mouvement, cette thèse propose quelques études d'analyse de méthodes numériques de simulation et contrôle en chimie quantique.

La Partie I est dédiée à la présentation du cadre général de la chimie quantique computationnelle.

Les techniques d'estimation a posteriori et leurs applications à la chimie quantique sont présentées dans la Partie II. Après une introduction aux méthodes a posteriori (chapitre II-1) on présente dans le chapitre II-2 une étude mathématique de la méthode des variables adiabatiques et on propose en particulier un estimateur a posteriori. Des simulations numériques qui illustrent les qualités de cet estimateur sont aussi présentées. Une étude d'analyse a posteriori de l'équation de Hartree-Fock est présentée en chapitre II-3. Outre la construction d'un intervalle de confiance pour localiser l'énergie de Hartree-Fock, la même méthode fournit un procédé d'amélioration des solutions numériques. Ces résultats théoriques ont été testés et implémentés sur un code de chimie quantique.

Une étude concernant le contrôle au niveau quantique des phénomènes chimiques est présentée dans la Partie III. Des résultats permettant de décider de la contrôlabilité d'un système de dimension finie sont présentés en III-1.2. Les critères sont simples à vérifier et donnent lieu à des interprétations intuitives. Des résultats théoriques complémentaires et des exemples numériques sont proposés dans III-1.3. Finalement, un code qui implémente des algorithmes génétiques pour l'étude des mécanismes de contrôle est décrit en chapitre III-2.

Abstract. Mathematics applied to quantum chemistry is an emerging research field. Placed within this context, this thesis presents some contributions to the analysis of numerical methods of simulation and control in quantum chemistry.

The Part I is dedicated to the presentation of the general framework of the computational quantum chemistry.

A posteriori estimation techniques and their applications to quantum chemistry are presented in Part II. After an introduction to the a posteriori methods (chapter II-1) we present in chapter II-2 a mathematical analysis of the adiabatic variable method and we propose in particular an a posteriori estimator. Numerical simulations that support the theory are also presented. An a posteriori analysis of the Hartree-Fock equation is presented in chapter II-3. In addition to the construction of a trust interval for the Hartree-Fock energy, the same method provides an improvement of the numerical solutions. The theoretic results have been tested and implemented on a quantum chemistry code.

A study concerning the control of the chemical phenomena at the quantum level is presented in Part III. Results allowing to asses the controllability of finite dimensional systems are presented in III-1.2. The criteria are easy to check and allow for intuitive understanding. Complementary theoretical results and numerical examples are proposed in III-1.3. Finally, a code that implements genetic algorithms to study control mechanisms is described in chapter III-2.

Table des matières

Table des matières **9**

I Introduction 11

1 Introduction **13**

2 Cadre général quantique **17**
2.1 Introduction à la théorie quantique 17
2.2 Espace de configurations 19
2.3 Équation de Schrödinger 20
2.4 Approximation de Born-Oppenheimer 21
2.5 Équation de Hartree-Fock 23
2.6 Aspects numériques, LCAO 26

II Methodes a posteriori 29

1 Techniques d'estimation a posteriori **31**
1.1 Introduction aux techniques d'estimation a posteriori 31
1.2 Principes mathématiques 32
1.2.1 Indicateur d'erreur 33
1.2.2 Bornes sur les fonctionnelles de la solution 34

2 Procédé de réduction adiabatique **37**
2.1 Construction de l'hamiltonien nucléaire 38
2.2 Présentation du procédé 40
2.2.1 Construction des bases adaptées 40
2.2.2 Définition et diagonalisation de l'hamiltonien réduit . . 42
2.3 Résultats théoriques et expérimentations numériques 44

3 Étude des équations de Hartree-Fock **65**

III Contrôle en chimie quantique 101

1 Contrôlabilité des systèmes quantiques 105
1.1 Considérations générales 105
1.2 Résultats théoriques 113
1.3 Applications 127

2 Algorithmes génétiques 143
2.1 Why GA? 144
2.2 Introduction to Genetic Algorithms 145
2.3 Implementation 147
2.4 Numerical results 152

A Annexe (Appendix) 165
A.1 Improvements of controllability results 165
A.1.1 Elimination of the periodicity hypothesis 165
A.1.2 Comments on the extension of controllability results for general connectivity graphs 167
A.2 Controllability of 3-level systems 169

Bibliographie générale 175

Index 183

Première partie

Introduction

Chapitre 1

Introduction

La chimie quantique est un domaine de recherche de plus en plus présent dans les préoccupations des mathématiciens appliqués. En France, cet intérêt se traduit par un nombre grandissant de thèses [86, 87, 88, 89, 91, 93] et des contributions originales, et aussi par des collaborations pluridisciplinaires [73]; on peut en effet constater que la recherche mathématique en chimie quantique a atteint le point où elle donne lieu a des publications communes avec les chimistes dans les revues spécialisés et dont les résultats sont implémentés dans des codes largement diffusés [89].

Partant de l'expérience acquise par les mathématiciens appliqués dans le domaine des simulations des équations d'évolution ou aux dérivées partielles, cette thèse se propose d'exploiter ces compétences dans le cadre de la chimie quantique selon deux sujets directeurs: les analyses a posteriori et le contrôle des équation d'évolution.

Partie I. Une introduction succincte à la mécanique quantique, aux méthodes et aux équations de la chimie quantique computationnelle ab initio est proposée en Partie I, chapitre 2. On y retrouve les équations de Schrödinger, l'approximation de Born-Oppenheimer et les équations de Hartree-Fock.

Partie II. Ayant comme seule base les principes de la mécanique quantique, les calculs *ab initio* sont utilisés pour connaître et prédire les propriétés quantiques des systèmes chimiques. La description de l'état des systèmes par l'intermédiaire de la fonction d'onde donne souvent naissance à des problèmes de taille trop importante pour des calculs directs (voir section 2.6 et chapitre 3). Des techniques d'approximation sont alors employées, mais peu de travaux théoriques existent qui puissent garantir la qualité des résultats ainsi obtenus. Il apparaît donc nécessaire de pouvoir quantifier la confiance

à mettre dans de tels calculs numériques. A cette fin on a fait appel à des méthodes utilisées dans le calcul scientifique sous le nom d'analyse a posteriori; ces méthodes ont pour but la construction de quantités calculables seulement à partir de la solution approchée obtenue qui donnent des indications qualitatives et/ou quantitatives sur la précision du calcul fait.

Les techniques d'estimation a posteriori et leurs applications à la chimie quantique sont présentées dans la Partie II. Après une introduction rapide aux méthodes d'estimation a posteriori (chap. 1) on présente dans le chapitre 2 une étude mathématique rigoureuse de la méthode des variables adiabatiques, largement employée par les chimistes afin de réduire la dimension des systèmes à résoudre lors de la simulation du mouvement nucléaire. Faute d'explication théorique, beaucoup d'empirisme existe encore sur ce type d'approche; il nous a paru important de présenter une étude mathématique rigoureuse de cette approximation et en particulier de proposer un estimateur a posteriori qui pourrait permettre de vérifier l'hypothèse d'adiabaticité faite sur certaines variables et qui constitue la base du processus d'approximation. Des simulations numériques qui illustrent les qualités de cet estimateur dans des cas couramment utilisés par les chimistes sont aussi présentées.

Le calcul de l'énergie électronique est une préoccupation majeure de la chimie quantique computationnelle et la méthode de Hartree-Fock est un des moyens ab initio les plus utilisés. Il nous paraît légitime là encore de se poser le problème de la **fiabilité** des ces calculs, d'autant plus que très souvent on y rencontre toute une diversité d'approximations dont les justifications mathématiques sont assez peu développées ou pas du tout. Une étude d'analyse a posteriori de l'équation de Hartree-Fock est présentée en chapitre 3. L'objectif est de construire, à partir d'une solution discrète obtenue par un calcul préalable, un intervalle de confiance pour localiser l'énergie de Hartree-Fock. Outre la construction de cet intervalle, il est montré que, due à la présence du problème variationnel sous-jacent, on peut proposer des solutions numériques plus précises que le calcul initial non seulement pour l'énergie mais aussi pour la fonction d'onde. Ces résultats théoriques ont été testés et implémentés sur un code de chimie quantique computationnelle.

Partie III. Rendre possible, influencer le cours ou changer le résultat d'une réaction chimique est au coeur même de la chimie expérimentale. Il n'est donc pas surprenant que beaucoup d'études existent actuellement dans le contrôle au niveau quantique des réactions chimiques.

Quand les réactants sont dans un certain état (supposé connu) le cours d'une réaction est régi par l'équation de Schrödinger dépendant du temps ayant comme donnée initiale l'état en question. Cependant, il est possible

de changer le résultat final par l'application de champs externes et éviter (ou minimiser) ainsi la formation de produits non désirés et maximiser les produits recherchés. Le champ externe destiné à influencer l'évolution d'une réaction peut être par exemple une ou plusieurs impulsions laser, dépendant du temps; la pratique montre que trouver une telle impulsion est un travail difficile, car les chimistes eux mêmes manquent à présent d'intuition physique sur les mécanismes de contrôle; des techniques de mathématique appliquée ont dûes être importées et ont commencé à donner des résultats positifs dans quelques cas particuliers. Une étude théorique et numérique concernant le contrôle des systèmes quantiques est présentée dans la partie III.

Une question théorique centrale concernant le contrôle en chimie quantique est le problème de la **contrôlabilité**, c'est à dire l'étude de ce que le modèle permet de réaliser à partir d'un état initial donné. Des résultats permettant de décider de la contrôlabilité d'un système discret (de dimension finie) sont présentés en section 1.2. Les critères sont simples à vérifier et donnent lieu à des interprétations intuitives.

Une technique utile pour étendre le champ d'application des résultats théoriques de la section 1.2, des applications, ainsi que des considérations sur les connexions qui existent entre les lois de conservation et les "défauts" de contrôlabilité sont présentées en section 1.3. D'autres généralisations du théorème de contrôlabilité ont été regroupées dans l'annexe A.

Les algorithmes génétiques (AG) sont parmi les méthodes de simulation les plus utilisées pour trouver les impulsions qui réalisent le contrôle aussi bien pour des expériences réelles en laboratoire que lors des simulation numériques. A cette spécificité contribue la possibilité technologique de réaliser un nombre important d'impulsions laser par unité de temps et le peu de moyens nécessaires en terme de quantité de substance à contrôler (quelques dizaines de molécules suffisent). Sur le plan numérique des simulations classiques de contrôle optimal sont très difficiles, voire impossibles, à mettre en oeuvre pour de systèmes d'intérêt pratique ce qui rend compétitives les AG. Même si la qualité des solutions trouvées par les AG en laboratoire et encourageante, l'analyse du processus de contrôle se fait au cas par cas et ne donne pas encore de compréhension des phénomènes ni des réactions et donc n'a pas de pouvoir prédictif. Des simulations numériques et des études théoriques sont nécessaires pour trouver les bonnes approximations et décrire ces solutions. Dans cette perspective un code parallèle a été développé et est employé sur les machines de l'ASCI. Ce code est présenté en chapitre 2 ainsi qu'une introduction générale aux AG. Des techniques de *filtrage* ont été mises au point pour éliminer dans les solutions obtenues (par les AG) les parties qui ne contribuent pas activement au contrôle, ouvrant ainsi la voie pour l'interprétation et la compréhension des mécanismes de contrôle.

Chapitre 2

Cadre général des calculs en chimie quantique

2.1 Introduction à la théorie quantique

Le but de ce chapitre est de présenter le cadre de la chimie quantique computationnelle (non relativiste) dans une perspective pratique à l'usage du mathématicien appliqué ; on ne cherchera pas à "démontrer" les équations présentées mais plutôt à donner des indications intuitives là où de telles indications correspondent à des réalitées plus profondes de ce domaine; le lecteur intéressé à approfondir ce thème est invite à consulter les ouvrages dans les références de la partie [1] ou encore [85], et à compléter sa vision sur les variables de spin qui ne sont introduites ici que de façon sommaire.

La *théorie quantique* en physique est une description des particules élémentaires qui forment la matière et de leur interaction tout d'abord entre elles et aussi avec diverses formes d'énergie. Le nom "théorie quantique" vient du fait que la matière et l'énergie sont décrites en termes d'unités indivisibles appelées "quanta" (singulier "quantum"). Mentionnons que la physique classique diffère de la théorie quantique dont elle est une approximation et donne de très bon résultats sur les systèmes macroscopiques comme en mécanique des solides ou des fluides. Elle est beaucoup moins précise dès qu'il s'agit d'objets microscopiques tels que les atomes ou les manifestations de l'énergie à une très petite échelle.

La théorie quantique est plus générale que la physique classique et elle pourrait en principe prédire le comportement de tout système physique, chimique ou biologique; néanmoins expliquer en quantique le comportement des systèmes du monde quotidien est généralement[1] trop compliqué pour

[1] une exception est par exemple la conduction en phase solide qui est intrinsèquement

être pratiquement réalisable.

La théorie quantique ne spécifie pas seulement des règles propres pour la description de l'univers mais introduit aussi de nouvelles façons de penser sur la matière et l'énergie. Les particules élémentaires qui sont décrites par la théorie quantique n'ont pas de localisation, vitesse, chemin bien défini comme les objets classiques mais leur propriétés sont décrites en terme de probabilités que la propriété respective ait une certaine valeur. Par exemple le résultat d'un calcul quantique peut être la probabilité qu'un certaine particule soit dans une position donnée à un temps donné.

La description quantique des particules permet de comprendre comment celles-ci se combinent pour former des atomes ou molécules ou encore des substances plus complexes. C'est ici que commence le domaine de la chimie quantique computationnelle qui a pour but de déterminer certaines propriétées des substances chimiques en partant de leur description au niveau quantique. Pour chaque système le choix de la théorie qui sera utilisée pour le décrire est essentiel; quand de choix est fait en faveur d'un traitement purement quantique on parle des calculs *ab initio* ; quand des connaissances plus *classiques* sont injectées dans le modèle (les plus souvent afin de le simplifier) on aura recours à des modèles dits *(semi-)empiriques*.

On ne détaillera pas ici toute la zoologie des modèles en chimie quantique, il est pourtant utile de mentionner que notre travail se place dans le cadre des modèles **ab initio**. Ce choix se fait en fonction de la taille, de l'énergie ou encore des interactions à étudier dans le système respectif et est à mettre en relation avec le principe de la dualité onde-particule qui affirme que toute particule (système) peut être vue aussi comme une onde : plus la longueur d'onde (de Broglie) associée est petite plus le cadre naturel devient classique car la nature ondulatoire de la particule (du système) devient de plus en plus indétectable. La relation qui donne la longueur d'onde d'une particule en fonction de son impulsion permet de constater que celle-ci est extrêmement petite pour tout objet macroscopique (visible) et plus généralement pour tout objet beaucoup plus grand qu'un atome. Par exemple, la longueur d'onde d'une balle de tennis envoyée à $200km/h$ est $0.825 \times 10^{-34}m$ ce qui est 100 trillions de milliards fois (10^{24}) plus petit que la taille du plus petit atome dont celle-ci est constituée (l'atome d'hydrogène).

quantique et doit être traitée ainsi

2.2 Espace de configurations, fonction d'onde, observables

On décrira ensuite les notions de base de la chimie quantique sans trop détailler le traitement des variables de spin, on renvoie le lecteur à [1] ou par exemple [89] pag.20-42 pour plus de détails.

Contrairement à la physique classique, l'espace le plus adapté à la description des systèmes en physique quantique n'est pas l'espace tridimensionnel réel, mais un espace nommé *espace des configurations* qui représente l'ensemble de toutes les configurations possibles pour un système quantique. Supposons par exemple qu'on s'intéresse à un système à N particules P_1,...,P_N isolées et sans contraintes. Les degrés de liberté du système seront alors l'ensemble des coordonnées des particules $r_i = (x_i,y_i,z_i) \in \mathbb{R}^3$, $i = 1,...,N$ et l'ensemble des variables de spin $s_i \in \mathbb{S}_i$, $i = 1,...,N$ qu'on notera d'une manière globale $R_i = (r_i,s_i)$. Chaque variable de spin prend ses valeurs dans un ensemble fini $\mathbb{S}_i$ qui dépend de la particule; par exemple pour l'électron cet ensemble $\mathbb{S}_e$ a seulement 2 valeurs: *spin up* et *spin down*, ce qui est noté symboliquement $\mathbb{S}_e = \{|+>,|->\}$ ou encore $\mathbb{S}_e = \{-\frac{1}{2},\frac{1}{2}\}$. Comme dans les applications traitées dans cette thèse les variables de spin ne jouent pas de rôle, sinon d'alourdir les notations, on a choisi de ne pas entrer dans les détails sur le traitement de ces variables de spin et on a adapté notre présentation en conséquence. Soit donc $R = (R_1,...,R_N) \in R^{3N} \times \sqcap_{i=1}^{N} \mathbb{S}_i$ l'ensemble des degrés des liberté des particules. A chaque instant t le système quantique est alors complètement décrit par une fonction complexe $\Psi(t) \in L^2(R^{3N} \times \sqcap_{i=1}^{N} \mathbb{S}_i;\mathbb{C})^2$ de norme L^2 égale à 1 qui est appelée *fonction d'onde* du système, avec l'interprétation suivante: pour tout point $R \in R^{3N} \times \sqcap_{i=1}^{N} \mathbb{S}_i$, $|\Psi(R,t)|^2$ est la "probabilité"[3] que le système soit dans la configuration R à l'instant t. L'approximation de la physique classique qui consiste comme on l'a dit à remplacer cette fonction par une masse de Dirac est justifiée pour des objets classiques ou la dualité onde-particule ne joue par un rôle essentiel, mais donne des résultats faux pour des systèmes atomiques dont les caractéristiques ondulatoires plus prononcées induisent sur $\Psi(R,t)$ des structures complexes.

La transcription mathématique des notions de la physique classique fait intervenir la notion d'*observable*. A chaque quantité mesurable classique on associe un opérateur linéaire Hermitien (borné ou non-borné) et qui agit sur

[2] la mesure d'intégration dR sera la mesure canonique du produit $R^{3N} \times \sqcap_{i=1}^{N} \mathbb{S}_i$: mesures de Lebesgue pour chaque $\mathbb{R}^3$ et mesures discrètes uniformes sur chaque $\mathbb{S}_i$

[3] il faut en fait parler de la densité de probabilité car on est sur un produit faisant intervenir des espaces continus

la fonction d'onde. Parmi les opérateurs les plus importants on mentionne

- la position de la particule i par rapport a un axe (par exemple l'axe Ox), notée par x_i et qui agit comme une multiplication par x_i;
- l'impulsion de la particule i par rapport à un axe (par exemple l'axe Ox), notée par p_{x_i} et qui agit comme une dérivation par rapport à la variable x_i $p_{x_i} = -i\hbar\frac{\partial}{\partial x_i}$ où $\hbar = 6.63 \times 10^{-34} Js$ est la constante de Plank ;
- l'énergie cinétique T de la particule i, notée par T_i et donnée par $T_i = -\frac{\hbar^2}{m_i}\Delta_i$, ou Δ_i est l'opérateur de Laplace qui agit seulement sur les 3 coordonnées de r_i, $\Delta_i = \frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2}$ et m_i est la masse de la particule;
- l'énergie potentielle $V(r)$ qui agit comme une multiplication par la fonction $V(r)$;
- l'opérateur Hamiltonian qui correspond a l'observable "énergie" du système: $H = T + V$.

Les postulats de la mécanique quantique affirment que pour chaque mesure d'une observable (associée à un opérateur) A, les seuls résultats possibles sont des valeurs propres de A (quantisation des valeurs des variables dynamiques); l'état lui même n'étant pas nécessairement une fonction propre de A, la valeur mesurée pour une fonction d'onde donnée sera une variable aléatoire qui admet une valeur moyenne $< \Psi(R,t), A\Psi(R,t) >$. On à noté ici par $< \cdot,\cdot >$ le produit scalaire L^2 complexe canonique $< f,g >= \int_{\mathbb{R}^{3N}\times\Pi_{i=1}^{N}\mathbb{S}_i} f(R)\overline{g(R)}dR$. L'opérateur A étant Hermitien

$$< \Psi(R,t), A\Psi(R,t) >=< A\Psi(R,t), \Psi(R,t) > .$$

Cette quantité sera ensuite notée $< \Psi(R,t)|A|\Psi(R,t) >$ (notation dite *braket*).

2.3 Équation de Schrödinger

La dynamique d'un système quantique est régie par *l'équation de Schrödinger dépendant du temps*:

$$i\hbar\frac{\partial}{\partial t}\Psi(R,t) = H\Psi(R,t). \tag{2.3.1}$$

Quand l'Hamiltonian du système ne dépend pas du temps une solution remarquable de (2.3.1) est l'évolution stationnaire $\Psi(R,t) = e^{-\frac{i}{\hbar}Et}\Psi(R,0)$ où

$\Psi(R) = \Psi(R,0)$ est une fonction propre de H de valeur propre correspondante E:

$$H\Psi(R) \ = E\Psi(R), \ \|\Psi(R)\|_{L^2} = 1. \tag{2.3.2}$$

L'équation (2.3.2) est appelée *l'équation de Schrödinger indépendant du temps* ; quand E est la plus petite valeur propre de H, (2.3.2) admet une autre interprétation. On considère le problème de minimisation de l'énergie

$$\inf\{E =< \Psi(R)|H|\Psi(R) >; \Psi(R) \in \mathcal{A}, \ \|\Psi(R)\|_{L^2} = 1\} \tag{2.3.3}$$

où $\mathcal{A}$ est l'espace des fonctions physiquement admissibles. L'espace $\mathcal{A}$ dépend du nombre de bosons et fermions du système et est généralement un sous-espace vectoriel strict de $L^2(R^{3N} \times \sqcap_{i=1}^{N} \mathbb{S}_i; \mathbb{C})$, (voir plus loin dans section 2.4 l'introduction de l'hypothèse de Born-Oppenheimer et dans 2.5 le modèle Hartree-Fock) ; pour ne pas entrer trop en détails sur les variables de spin et les symétries de la fonction d'onde nous remettrons à plus tard la description de $\mathcal{A}$.

Un minimiseur de (2.3.3) (lorsqu'il existe) est appelée *état fondamental* du système ; c'est donc en particulier un état de plus basse énergie et aussi un état physiquement *stable* ; toute autre solution de (2.3.3) est appelée *état excité*.

2.4 Approximation de Born-Oppenheimer

Pour les applications en chimie quantique auxquelles on s'intéresse, le système est un ensemble d'électrons et de noyaux de masses m_i, de charges Z_i et de coordonnées $r_i \in \mathcal{R}^3$; l'opérateur H s'écrit comme la somme d'un opérateur *énergie cinétique* T et d'un autre appelé V, $H = T + V$; dans V les seules interactions qui nous intéressent sont l'attraction et répulsion coulombiennes entre les particules :

$$T = -\frac{\hbar^2}{2} \sum_{i=1}^{N} \frac{1}{m_i} \Delta_i \tag{2.4.4}$$

$$V = \sum_{i<j} \frac{Z_i Z_j}{4\pi\epsilon_0} \frac{1}{|r_i - r_j|} \tag{2.4.5}$$

On à noté ici par ϵ_0 la permitivité du vide.

Étant donnée la taille des problèmes (2.3.1) ou (2.3.2) (on rappelle que r est dans un espace à $3N$ dimensions !!!) les approches directes de discrétisation totale de l'espace des solutions ne conduisent à des résolutions effectives que pour les systèmes extrêmement simples (atome d'hydrogène). Il faut alors introduire des approximations pour les cas auxquels la chimie s'intéresse. L'approximation la plus utilisée est celle dite de *Born-Openheimer* qui consiste, tenant compte du fait que la masse d'un noyau est beaucoup plus importante que la masse d'un électron (ce qui donnerait à impulsion constante un mouvement électronique très rapide autour des noyaux), à supposer que les électrons s'adaptent instantanément à la "position" des noyaux. Du point de vue de la physique ceci revient à séparer l'évolution du système en mouvement électronique et mouvement nucléaire et à remplacer le premier par un opérateur potentiel lors du traitement du dernier. Plus précisément, si on indice les noyaux par n,n' et les électrons par e,e' et on pose $R = (R_n,R_e)$ (R_n = ensemble des degrés de liberté nucléaires, R_e = ensemble des degrés de liberté électroniques), on peut écrire l'opérateur hamiltonien H sous la forme

$$H = -\frac{\hbar^2}{2}\sum_n \frac{1}{m_n}\Delta_n + \frac{1}{2}\sum_{n,n'} \frac{Z_n Z_{n'}}{4\pi\epsilon_0}\frac{1}{|r_n - r_{n'}|}$$
$$-\frac{\hbar^2}{2}\sum_e \frac{1}{m_e}\Delta_e + \frac{1}{2}\sum_{n,e} \frac{Z_n Z_e}{4\pi\epsilon_0}\frac{1}{|r_n - r_e|} + \frac{1}{2}\sum_{e,e'} \frac{Z_e Z_{e'}}{4\pi\epsilon_0}\frac{1}{|r_e - r_{e'}|}. \quad (2.4.6)$$

On note alors, pour toute "position" R_n des noyaux, par $H_e(R_n)$ l'hamiltonien (dit *électronique*) représente par les trois derniers termes de (2.4.6):

$$H_e(R_n) = -\frac{\hbar^2}{2}\sum_e \frac{1}{m_e}\Delta_e + \frac{1}{2}\sum_{n,e} \frac{Z_n Z_e}{4\pi\epsilon_0}\frac{1}{|r_n - r_e|} + \frac{1}{2}\sum_{e,e'} \frac{Z_e Z_{e'}}{4\pi\epsilon_0}\frac{1}{|r_e - r_{e'}|} \quad (2.4.7)$$

On diagonalise ensuite $H_e(R_n)$ obtenant une fonction propre $\Phi_{R_n}(R_e)$ appelée *fonction d'onde électronique* (qui dépend paramètriquement des degrés de liberté nucléaires R_n et a comme variables seulement les degrés de liberté électroniques R_e) et une valeur propre $E(R_n)$:

$$H_e(R_n)\Phi_{R_n}(R_e) = E(R_n)\Phi_{R_n}(R_e). \quad (2.4.8)$$

Pour calculer l'énergie totale du système on diagonalise ensuite l'*hamiltonien nucléaire* défini par:

$$H_n = -\frac{\hbar^2}{2}\sum_n \frac{1}{m_n}\Delta_n + \frac{1}{2}\sum_{n,n'} \frac{Z_n Z_{n'}}{4\pi\epsilon_0}\frac{1}{|r_n - r_{n'}|} + E(R_n). \quad (2.4.9)$$

agissant sur la *fonction d'onde nucléaire* (qui dépend seulement de R_n). La solution du problème aux valeurs propres

$$H_n\Psi(R_n) = E\Psi(R_n) \tag{2.4.10}$$

permet alors de proposer une valeur pour l'énergie du système, à savoir la valeur propre E, et un candidat pour la fonction d'onde du système, à savoir $\Psi(R_n)\Phi_{R_n}(R_e)$.

Remarque 2.4.1 *En l'absence d'interaction avec l'extérieur, l'hamiltonien nucléaire défini dans (2.4.9) n'agit pas sur les variables de spin; si le système est aux noyaux discernables (i.e. chaque noyau est présent en un seul exemplaire) on peut alors résoudre l'équation de Schrödinger pour chaque configuration de spin possible sans imposer de contraintes sur les variables spatiales, ce qui revient à supposer que les noyaux n'ont pas de variable de spin.*

2.5 Équation de Hartree-Fock

On se place dans l'étape de diagonalisation de l'hamiltonien électronique dans le paradigme Born-Oppenheimer (voir ci-dessus) et on considère un système dont les coordonnées des noyaux sont **fixées** en $\bar{r}_j \in \mathbb{R}^3$, $j = 1,...m$ et les variables de spin en $\bar{s}_j \in \mathbb{S}_j$, $j = 1,...m$ et on note par $r_i \in \mathbb{R}^3$ $i = 1,...,N_e$ les coordonnées des électrons. Afin de ne pas toujours écrire les facteurs supplémentaires $\hbar$ et ϵ_0 on fera les changements d'échelle nécessaires pour travailler en unités atomiques. L'hamiltonien électronique devient alors :

$$H_e = -\sum_{i=1}^{N_e}\Delta_i + \sum_{i=1}^{N_e}V(r_i) + \sum_{i<j}\frac{1}{|r_i - r_j|} \tag{2.5.11}$$

$$V(r) = -\sum_{j=1}^{m}\frac{Z_j}{|r - \bar{r}_j|} \tag{2.5.12}$$

Le but est de calculer la première valeur propre de H_e qui est l'énergie électronique, c'est à dire trouver

$$E = \inf\Big\{(H_e\Phi,\Phi)|\Phi \in \mathcal{A}_e, \|\Phi\|_{L^2} = 1\Big\} \tag{2.5.13}$$

où $\mathcal{A}_e$ est le sous-espace vectoriel de $H^1(\mathbb{R}^{3N_e} \times \sqcap_{i=1}^{N_e}\mathbb{S}_i)$ des fonctions d'onde physiquement admissibles.

Comme on à déjà pu le voir, la dimension du système est trop importante pour un calcul direct ; par exemple dans le cas (simple !) de l'eau (H_2O) le nombre d'électrons est 10 donc l'opérateur H_e agit sur des fonctions à 30 variables (sans compter les variables de spin). Une discrétisation minimale (chaque variable vit sur $\mathbb{R}$ tout entier !) de 10 points par direction donnerait une matrice $10^{30} \times 10^{30}$ **impossible** à calculer en pratique. Il y à donc besoin d'approximations simplificatrices dont la plus connue est celle dite de *Hartree-Fock* . Pour ne pas compliquer la présentation on introduira cette approximation seulement dans sa forme RHF (*Restricted Hartree-Fock*). On supposera donc que le nombre d'électrons dans le système est pair (ce qui sera toujours le cas dans tout ce qui suit).

Le concept de départ est celui de *paire d'électrons de Lewis* qui dit que lorsque le nombre d'électrons dans le système est pair on peut les grouper deux par deux et associer à chaque paire une variable spatiale dans $\mathbb{R}^3$. à cette variable il correspond un électron de type *spin up* et un autre de type *spin down*. Symboliquement, à la place de N_e variables de position et N_e variables de spin on aura seulement $N_{pe} = N_e/2$ variables de position pour caractériser le système; on designera par $r_1,...,r_{N_{pe}}$ ces variables. Une fois cette *approximation* faite la fonction d'onde est cherchée dans $H^1(\mathbb{R}^{3N_{pe}}; \mathbb{C})$

Une propriété de la fonction d'onde exploitée dans l'approximation de Hartree-Fock est l'antisymétrie:

$$\Phi(...,r_i,...,r_j,...) = -\Phi(...,r_j,...,r_i,...), \;\; i,j = 1,...,N_{pe}$$

qui découle des principes fondamentaux de la mécanique quantique (appliqués à ce cas) et dont une justification (intuitive!) réside dans le principes d'exclusion de Pauli[4] et celui d'indiscernabilité des électrons[5].

Le minimum (2.5.13) va donc porter sur l'espace des fonctions antisymétriques de $H^1(\mathbb{R}^{3N_{pe}}; \mathbb{C})$ dont un sous-espace strict mais remarquable est celui des fonctions qui s'écrivent comme un déterminant (dit de *Slater*):

$$\Phi(r_1,...,r_{N_{pe}}) = \frac{1}{\sqrt{N_{pe}!}} \det(\Phi_i(r_j)),$$

les Φ_i, $i = 1,...,N_{pe}$ étant des fonctions de $H^1(\mathbb{R}^3; \mathbb{C})$ qu'on va choisir orthogonales (L^2) deux à deux.

[4]Le principe de Pauli affirme qu'on ne peut pas trouver deux électrons du même spin au même endroit, ou, d'une manière équivalente, que la probabilité associée à une configuration avec deux électrons au même endroit est zéro, ce qui pour notre cas s'écrit : $\Phi(...,r_i = r,...,r_j = r,...) = 0$.

[5]ce principe qui affirme que tous les électrons sont identiques est une propriété particulière de la fonction d'onde et s'écrit dans notre cas : $|\Phi(...,r_i,...,r_j,...)| = |\Phi(...,r_j,...,r_i,...)|$, $i,j = 1,...,N_{pe}$.

Introduisons la *matrice de densité* $\rho(x,y) = \sum_{i=1}^{N_{pe}} \Phi_i(x)\overline{\Phi_i(y)}$ et la fonction *densité électronique* $\rho(x) = \rho(x,x)$.

L'énergie associée à la fonction d'onde devient :

$$\begin{aligned}\mathcal{E}^{HF}(\Phi_1,...,\Phi_{N_{pe}}) &= \sum_{i=1}^{N_{pe}} \int_{\mathbb{R}^3} \left(|\nabla\Phi_i|^2 + V\ |\Phi_i|^2\right) + \frac{1}{2}\int\int_{\mathbb{R}^3\times\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|}dxdy \\ &\quad -\frac{1}{2}\int\int_{\mathbb{R}^3\times\mathbb{R}^3} \frac{|\rho(x,y)|^2}{|x-y|}dxdy \qquad (2.5.14)\end{aligned}$$

et le nouveau problème s'écrit:

$$\inf\left\{\ \mathcal{E}^{HF}(\Phi_1,...,\Phi_{N_{pe}})|\Phi_i \in H^1(\mathbb{R}^3),\ <\Phi_i,\Phi_j>=\delta_{ij},\ i,j=1,...,N_{pe}\right\} \qquad (2.5.15)$$

En écrivant l'équation d'Euler Lagrange associée à ce problème de minimisation sous contraintes on aboutit à

$$F_\Phi\Phi_i = \sum_{i=1}^{N_{pe}} \tilde{\epsilon}_{ij}\Phi_j,\ \ i=1,...,N_{pe} \qquad (2.5.16)$$

où l'opérateur F_Φ, appellé *opérateur de Fock*, est défini par

$$F_\Phi\Phi_i = -\Delta\Phi_i + V\Phi_i + (\rho\star\frac{1}{|x|})\Phi_i - \int\int_{\mathbb{R}^3\times\mathbb{R}^3} \frac{\rho(x,y)\Phi_i(y)}{|x-y|}dxdy \qquad (2.5.17)$$

et $\star$ est le produit de convolution

$$(f\star g)(x) = \int_{\mathbb{R}^3} f(x-y)g(y)dy. \qquad (2.5.18)$$

Notons que dans (2.5.16) la matrice $\tilde{\epsilon} = (\tilde{\epsilon}_{ij})_{i,j=1,N_{pe}}$ est symétrique. Comme Φ et F_Φ sont invariants par multiplication du vecteur $(\Phi_1,...,\Phi_{N_{pe}})$ par une matrice unitaire, on peut diagonaliser $\tilde{\epsilon}$ et écrire la célèbre équation de Hartree-Fock:

$$F_\Phi\Phi_i = \epsilon_i\Phi_i,\ \ i=1,...,N_{pe} \qquad (2.5.19)$$

Remarque 2.5.2 *L'équation de Hartree-Fock (2.5.19) est un problème aux valeurs propres non linéaire car la matrice de densité $\rho(x,y)$ et la densité $\rho(x)$ dépendent de la solution $(\phi_1,...,\phi_{N_{pe}}) \in H^1(\mathbb{R}^3;\mathbb{C})^{N_{pe}}$; de plus ce problème est non local à cause du produit de convolution qui entre dans la définition de l'opérateur de Fock F_Φ.*

2.6 Aspects numériques des calculs éléctroniques. Approximation LCAO.

La résolution numérique de l'équation (2.5.19) repose sur une approximation de Galerkin de l'espace $L^2(\mathbb{R}^3)$. Soit δ un paramètre de discrétisation et notons par G^δ le sous-espace discret de $L^2(\mathbb{R}^3)$ engendré par les fonctions h_1^δ, ..., $h_{n_\delta}^\delta$. Le problème discret associé à (2.5.19) consiste à chercher $\Phi^\delta = (\Phi_i^\delta)_{i=1}^{N_{pe}} = (\sum_{i=1}^{n_\delta} C_{ij} h_j^\delta)_{i=1}^{N_{pe}} \in (G^\delta)^{N_{pe}}$ et $(\epsilon_i^\delta)_{i=1}^{N_{pe}}$ tel que l'égalité $F_{\Phi^\delta}\Phi_i^\delta = \epsilon_i^\delta \Phi_i^\delta$ soit satisfaite pour tout $i = 1,...,N_{pe}$ dans le dual de G^δ.

Le calcul de la matrice de F_{Φ^δ} fait intervenir des contributions de la forme $< F_{\Phi^\delta} h_\alpha^\delta, h_\beta^\delta >_{L^2,L^2}$. Tenant compte de la définition de (2.5.17) de F_{Φ^δ} il sera nécessaire de calculer pour tous $i,j,k,l = 1,...,n_\delta$ l'intégrale

$$\int\int_{\mathbb{R}^3\times\mathbb{R}^3} \frac{h_i^\delta(x)\overline{h_j^\delta(x)h_k^\delta(y)}h_l^\delta(y)}{|x-y|}dxdy \tag{2.6.20}$$

appelée intégrale biélectronique.

Ce constat à des conséquences importantes pour la résolution numérique des équations de Hartree-Fock. On en déduit que la complexité algorithmique du calcul est de l'ordre au moins n_δ^4 et aussi que généralement les bases discrètes à utiliser doivent permettre un calcul très rapide des contributions (2.6.20). Pour répondre à cette nécessité, les h_i^δ sont choisies comme (somme de) fonctions Gaussiennes, ce qui permet de calculer les termes (2.6.20) par des formules analytiques [5] p.410-416. Ce choix est aussi envisageable sur le plan de l'analyse numérique car les fonctions Gaussiennes engendrent tour l'espace $L^2(\mathbb{R}^3)$ et donc des études de convergence peuvent être développées.

Malheureusement, la complexité de l'ordre 4 de l'algorithme requiert en pratique l'utilisation des bases de petite taille ce qui rend impossible toute analyse asymptotique. Pour garantir la convergence vers des solutions acceptables, les fonctions h_i^δ utilisées sont choisies comme somme de Gaussiennes optimisées pour reproduire au mieux les solutions de Hartree-Fock pour un seul noyau. La base $\{h_i^\delta; i = 1,...,N_{pe}\}$ sera ainsi une approximation de l'ensemble des orbitales atomiques[6] de chaque atome présent dans la molécule (parfois élargie avec d'autre fonctions spéciales). On parle alors de l'approximation *Linear Combination of Atomic Orbitals* (*LCAO*).

En pratique ce choix permet d'obtenir des très bons résultats. L'analyse des propriétés d'approximation des bases LCAO est pourtant difficile

[6] une orbitale atomique est une fonction d'onde mono-électronique qui correspond à un noyau donné

car l'ensemble des orbitales atomiques n'engendre pas tout l'espace $L^2(\mathbb{R}^3)$[7]. De plus, des optimisations additionnelles (voir [4] p.81, 86-87) sont mises en oeuvre pour tenir compte des effets d'écran, de la corrélation ... Ce changement continu et au cas par cas des bases de discrétisation rend les études d'analyse asymptotique classique (dite à priori) très difficiles et en fait peu intéressantes pour les chimistes. Plus utiles pour les praticiens s'avèrent être les techniques pour mesurer la qualité des solutions obtenues. Un exemple [27] de ce type d'analyse (appelée a posteriori) est présenté au chapitre 3 de la partie II .

[7] ces orbitales sont les fonctions propres correspondant aux valeurs propres discrètes d'un opérateur qui a aussi du spectre continu

Deuxième partie

Methodes a posteriori

Chapitre 1

Techniques d'estimation a posteriori

1.1 Introduction aux techniques d'estimation a posteriori

Pour un nombre de plus en plus important de problèmes posés dans le cadre industriel, la simulation numérique de problèmes régis par des équations aux dérivées partielles (*EDP*) a acquis suffisamment de fiabilité pour être utilisée en alternative à des expériences réelles. Néanmoins, la majeure partie des problèmes reste hors d'atteinte avec une fiabilité suffisante à cause de complexités trop importantes (taille mémoire, temps calcul....). Pour les problèmes qui sont à la limite des possibilités actuelles, il est intéressant de disposer d'outils annexes permettant de valider les calculs effectués. Dans cette optique, on ne peut se contenter des résultats de l'analyse numérique "classique" qui traitent de la convergence asymptotique des méthodes numériques utilisées, puisque on ne sait justement pas si on est dans le régime asymptotique. Depuis plusieurs années on a vu se développer au contraire des techniques d'estimation d'erreur (dites *a posteriori* par opposition aux méthodes asymptotiques dites *a priori*) où la qualité de l'approximation est exprimée en termes constructifs de la solution calculée.

Ces techniques ont connu ces dernières années des développements importants en mécanique des fluides et en mécanique des structures (on réfère en particulier à [28] pour une synthèse de ces techniques). On remarque néanmoins que pour les problèmes de calculs de modes propres d'une part [22, 28] et pour des applications plus "*exotiques*" dans les domaines de la chimie quantique computationnelle, peu de travaux existent.

Ce chapitre sera dédié à la présentation d'un certain nombre de résultats

dans ces directions, tout d'abord pour des estimations *a posteriori* de calculs de solution d'un problème nucléaire et ensuite sur la détermination explicite de bornes précises de valeurs propres pour des problèmes de type Hartree-Fock.

1.2 Principes mathématiques des analyses a posteriori

Dans ce qui suit on présentera d'une manière succinte les bases mathématiques des méthodes a posteriori. Avant d'entrer dans les détails techniques on ajoutera quelques mots sur le lieu et l'utilité de l'analyse a posteriori dans le domaine des simulations numériques.

Supposons qu'on veuille résoudre des EDP ou trouver des valeurs et fonctions propres (diagonaliser) d'opérateurs intervenant dans des EDP. On emploie alors des algorithmes de calcul dont la justification repose sur une analyse classique (a priori) qui étudie le comportement asymptotique de la solution discrète en fonction de la discrétisation spatio-temporelle choisie ; le résultat typique d'une telle analyse est un théorème de convergence qui montre qu'au fur et à mesure que la discrétisation s'affine l'algorithme choisi converge **vers la bonne solution**. Cette analyse donne aussi éventuellement divers estimations asymptotiques sur la vitesse de convergence.

En pratique pourtant il n'est pas rare de travailler sur des discrétisations dont la finesse n'est pas suffisante et qui sont difficiles à affiner davantage à cause d'un coût élevé ; par exemple dans le cas des problèmes aux valeurs propres la complexité cubique dans la dimension des vecteurs limite rapidement les possibilités de raffinement. Il apparaît alors le besoin d'un outil pour apprécier la fiabilité du calcul fait. En particulier, en chimie quantique computationnelle on fait souvent des comparaisons entre les énergies de deux configurations d'une molécule afin d'en déterminer la plus stable [89]. Il est alors important de s'assurer que la somme des erreurs faites sur les calculs est moins importante que la différence entre ces deux énergies, sinon on n'a aucun contrôle sur la fiabilité de la conclusion qu'on tire de la comparaison des deux valeurs.

Le cas de la chimie quantique computationnelle a ceci de particulier qu'on travaille souvent avec de gros codes qu'il est difficile de remplacer et sur lesquelles il existe toute une tradition concernant leur applicabilité et leur mise en oeuvre et peu de techniques pour une analyse quantitative des résultats.

C'est dans ce contexte qu'on s'intéresse à des procédés qui puissent quantifier la confiance que l'on peut avoir dans le résultat d'un calcul numérique.

Une réponse possible nous est donnée par les méthodes a posteriori dont le principe est présenté ci-dessous.

Supposons qu'on ait résolu numériquement le problème qui nous intéresse et qu'on soit en possession d'une solution approchée. L'analyse a posteriori nous permet alors de construire certaines quantités qui dépendent seulement des données de sortie (notre solution approchée) et qui nous aident à valider quantitativement le résultat ; lorsque ces quantités sont des estimations d'erreur globales on aura obtenu un critère pratique d'arrêt (on vérifie si on a atteint la précision voulue), et lorsque ce sont des estimations d'erreur locales on aura des indications sur les parties de l'espace continu qu'il faut discrétiser davantage (dans une perspective adaptative).

Pour simplifier la présentation des méthodes a posteriori on les a divisées en deux types: *bornes sur fonctionnelles de la solution* et *indicateur d'erreur* (*estimateur d'erreur*). Le premier type est à utiliser lorsque le but final est de calculer une certaine fonctionnelle de la solution (par exemple une observable) et se propose de donner des bornes explicites calculables a posteriori pour cette fonctionnelle appliquée à la solution discrète dont on dispose ; le deuxième type est employé plutôt dans le cadre des approches de résolution adaptatives ou encore lorsque son coût CPU plus petit le rend plus convenable que le premier type d'approche.

Remarque 1.2.3 *Il est d'usage, pour des raisons d'efficacité et surtout lors des approches adaptatives, d'exiger que l'estimateur ou les bornes mentionnées ci-dessus soit calculables plus facilement que la solution approchée elle-même; cette demande peut être néanmoins relaxée en fonction de la nécessité de pouvoir quantifier la fiabilité du résultat.*

1.2.1 Indicateur d'erreur

Les méthodes dont on présentera brièvement les assises mathématiques sont liées aux travaux de Babuška [16], Bernardi [18], Ladevèze [19], Oden [20], Pousin et Rappaz [26], Verfürth [28, 29], ... et partent du principe que l'erreur est équivalente au résidu. Plus précisément, soit $(H,\|.\|)$ un espace Hilbertien et F une fonction F de classe C^1 entre H et son dual H'. Supposons que le problème à résoudre est écrit sous la forme:

Trouver u dans H qui annule F:

$$F(u) = 0 \tag{1.2.1}$$

Si on note par u_0 une solution exacte de ce problème qu'on voudrait calculer (mais qui n'est pas connue explicitement), on peut écrire

$$\|F(u)\|_{H'} = \|F(u) - F(u_0)\|_{H'} = \|DF(u_0)(u-u_0) + o(\|u-u_0\|)\|_{H'}$$

où $DF(u_0)$ est la différentielle de F en u_0. Si l'on suppose que $DF(u_0)$ est un isomorphisme de H dans H' alors on voit que pour u assez proche de u_0 on aura une équivalence entre $\|u-u_0\|$ et $\|DF(u_0)(u-u_0)\|_{H'}$ et par conséquent on en déduit l'existence de deux constantes c et C (qui dépendent de $DF(u_0)$) telles que *pour* $\|u-u_0\|$ *suffisamment petit*:

$$c\|u-u_0\| \leq \|F(u)\|_{H'} \leq C\|u-u_0\|$$

où, d'une autre façon:

$$C^{-1}\|F(u)\|_{H'} \leq \|u-u_0\| \leq c^{-1}\|F(u)\|_{H'} \tag{1.2.2}$$

On a ainsi obtenu une quantité ($F(u)$) calculable a posteriori dépendant seulement du problème et de la solution approchée proposée u qui nous donne une indication (à l'aide des estimations **bilatérales** !) sur l'erreur $\|u-u_0\|$ faite sur la solution; de plus le calcul de cette quantité n'est pas trop élaboré, il faut appliquer F et calculer la norme du résultat dans l'espace dual[1]. Le désavantage de la méthode réside dans le fait qu'on ne connaît pas explicitement les constantes qui interviennent dans (1.2.2) ni les normes les plus optimales dans lesquelles il faut évaluer $F(u)$. Néanmoins ceci peut nous être utile si on veut par exemple comparer différentes méthodes ou pour connaître les endroits où il faut mettre davantage de points de discrétisation où lorsque on est en présence de propriétés d'uniformité de F qui nous permettent de prédire les constantes qui interviennent.

1.2.2 Bornes sur les fonctionnelles de la solution

En complément des approches dont le principe a été exposé ci-dessus on présentera par la suite une autre catégorie de méthodes d'analyse a posteriori ([21, 22, 23, 24, 25]) dont le but est de quantifier la fiabilité des **résultats en sortie** d'une simulation numérique; cette approche fournit des outils d'une précision élevée par rapport aux méthodes antérieurement présentées et a l'avantage d'une connaissance explicite de toutes les constantes qui interviennent, au prix, bien sur, d'un coût de calcul plus élevé.

[1] en général le calcul de la norme dans l'espace dual fait lui aussi l'objet de quelques simplifications comme on le verra plus loin dans l'analyse de la méthode de réduction adiabatique

La méthode a été initialement développée dans le cadre des simulations en ingénierie où la résolution de problèmes stationnaires ou d'évolution est un pas intermédiaire vers le calcul de certaines caractéristiques de ces systèmes (déformation, pression, portance, traînée) qui s'expriment explicitement à l'aide de la solution du problème considéré. Notons par α la solution du problème associé, par s la caractéristique à calculer et par l la fonctionnelle qui permet de calculer s à partir de α: $s = l(\alpha)$. Dans ce contexte le but d'une analyse a posteriori n'est pas de donner des indications sur l'erreur faite sur la solution α mais plutôt sur l'erreur faite sur la caractéristique s à calculer. À partir d'une solution approchée α_h dont on dispose une telle méthode permet de trouver des **bornes** inférieures $s_-(\alpha_h)$ et supérieures $s_+(\alpha_h)$ du résultat en sortie recherché : $s_-(\alpha_h) \leq s \leq s_+(\alpha_h)$, ce qui introduit le concept d'*intervalle de confiance*, c'est a dire l'intervalle $[s_-(\alpha_h), s_+(\alpha_h)]$ (dépendant de la solution approchée α_h) où on est sûr de trouver la valeur (exacte) s de la propriété à calculer ; moins cet intervalle est large plus notre calcul est précis.

La méthode des bornes pour des résultats en sortie sera utilisée pour l'équation de Hartree-Fock ; c'est dans cette optique qu'on choisi de la présenter très rapidement et d'une façon simplifiée pour le cas du problème aux valeurs propres suivant: trouver le premier mode propre du Laplacien sur un domaine borné $\Omega \subset \mathbb{R}^3$. Il s'agit donc de résoudre:

$$\begin{cases} -\Delta u = \lambda u \quad dans \quad \Omega \\ u \in H_0^1(\Omega), \quad \|u\|_{L^2(\Omega)} = 1 \\ u = 0 \quad sur \quad \partial\Omega \end{cases} \tag{1.2.3}$$

Soit (u_0, λ_0) une solution exacte de (1.2.3) et soit $l(\cdot,\cdot)$ une certaine fonctionnelle définie sur l'espace des solutions $H_0^1(\Omega) \times \mathbb{R}$ aux valeurs réelles et qui sera (pour des raisons de clarté) $l(u,\lambda) = \lambda$

Soit $(u, \lambda = (\nabla u, \nabla u))$ une approximation de (u_0, λ_0) normalisée par $\|u\|_{L^2(\Omega)} = 1$ ($(\cdot,\cdot)$ est le produit scalaire L^2) ; notons par $e = u_0 - u$ la différence entre la fonction propre exacte et son approximation. Alors on peut écrire :

$$\begin{aligned} \lambda_0 - \lambda = (\nabla u_0, \nabla u_0) - (\nabla u, \nabla u) = (\nabla e, \nabla(u_0 + u)) = \\ (\nabla e, \nabla e) + 2(\nabla u, \nabla e) = (\nabla e, \nabla e) + 2(-\Delta u, e) \end{aligned} \tag{1.2.4}$$

Quand e est petit en norme H^1, parmi les deux termes du membre de droite $(\nabla e, \nabla e)$ est "généralement" plus petit que $2(-\Delta u, e) = 2(\nabla u, \nabla e)$

car il fait intervenir la puissance deuxième du gradient de e ; on essaie alors d'écrire le terme $2(-\Delta u,e)$ à l'aide d'une quantité contenant ∇e. Plus précisément, soit $\hat{e} \in H_0^1(\Omega)$ solution de :

$$-\Delta \hat{e} = \Delta u + \lambda u \quad dans\ \Omega \tag{1.2.5}$$

Alors on peut écrire:

$$\begin{aligned} \lambda_0 - \lambda &= (\nabla e,\nabla e) + 2(\lambda u + \Delta \hat{e},e) = \\ &= (\nabla e,\nabla e) - 2(\nabla \hat{e},\nabla e) + 2\lambda(u,u_0 - u). \end{aligned} \tag{1.2.6}$$

D'autre part comme $\|u_0\|_{L^2(\Omega)} = \|u\|_{L^2(\Omega)} = 1$:

$$2(u,u_0 - u) = 2(u,u_0) - 2 = 2(u,u_0) - (u,u) - (u_0,u_0) = -(e,e) \tag{1.2.7}$$

donc

$$\lambda_0 - \lambda \ = -(\nabla \hat{e},\nabla \hat{e}) + (\nabla(e - \hat{e}),\nabla(e - \hat{e})) - \lambda(e,e) \tag{1.2.8}$$

Par une analyse asymptotique faisant intervenir les propriétés des discrétisations utilisées pour résoudre le problème (1.2.3) on montre ensuite que pour une large classe de discrétisations le dernier terme est d'ordre plus petit que le premier et tenant compte de la **positivité** du deuxième terme on obtient l'inégalité asymptotique:

$$\lambda_0 \geq \lambda - (\nabla \hat{e},\nabla \hat{e}) + ...$$

Comme d'autre part la première valeur propre vérifie l'inégalité variationnelle $\lambda_0 \leq \lambda$ on a ainsi obtenu des bornes sur la première valeur propre:

$$\lambda - (\nabla \hat{e},\nabla \hat{e}) + ... \leq \lambda_0 \leq \lambda$$

Notons que les bornes font intervenir seulement une fonction calculable par (1.2.5) à partir de la solution approchée (u, λ) et ne dépendent pas de la solution exacte (u_0, λ_0)[2]. De plus on peut montrer que cettes bornes sont asymptotiquement optimales [22] ; la pratique confirme elle aussi les résultats théoriques [22].

Remarque 1.2.4 *Notons que l'équation à résoudre (1.2.5) est plus simple que le problème de départ, car on remplace un problème aux valeur propres par une inversion de Laplacien. Ainsi l'effort demandé pour obtenir les bornes a posteriori est plus petit que celui fait pour résoudre le problème initial*[3]*. De plus ici toutes les quantités sont connues explicitement.*

[2]La solution de (1.2.5) peut par exemple être approchée par des techniques standard de résolution d'EDP ce qui introduit une étape de discrétisation supplementaire.

[3]Des réductions supplémentaires du temps de calcul requis par la résolution de (1.2.5) peuvent être réalisées par des techniques de relaxation sur sous-domaines, tout en préservant les bornes globales.

Chapitre 2

Étude du procédé de réduction adiabatique

Un des problèmes rencontrés dans le calcul scientifique en chimie quantique [32, 33, 34, 35] est la recherche des valeurs/fonctions propres de l'hamiltonien nucléaire (voir partie I section 2.4 formule (2.4.9)) ayant une énergie plus petite qu'une valeur E_{MAX} fixée à l'avance, ce qui en langage des Chimistes revient a trouver le fondamental et les premiers états excités (de vibration / rotation).

Le nombre de variables intervenant dans ce problème est important et comme de plus on s'intéresse à un spectre assez large, la taille de la base de discrétisation est grande à un point tel qu'elle interdit souvent tout calcul; pour donner une idée de l'état de l'art dans le domaine, notons qu'aujourd'hui on peut travailler avec des molécules bi- et tri-atomiques, on peut encore mener des calculs complets (avec des approximations fortement simplificatrices) pour les quadri-atomiques mais pour plus de 4 noyaux il n'y a pas de méthode directe efficace qui prenne en compte tous les degrés de liberté du système. On est alors amené à chercher des méthodes pour réduire le nombre de fonctions de base. Comme il s'agit d'un algorithme de diagonalisation (donc de complexité générique cubique dans la dimension de l'espace de discrétisation[1]) cette réduction potentielle est importante et en effet une diminution d'un facteur de 3 ou 4 de la base peut permettre de traiter numériquement un problème qui ne l'était pas avant cette réduction.

La *réduction adiabatique (pseudo-)spectrale* est l'un de ces procédés, largement utilisé dans la pratique, (voir [32, 34] et les références). Faute d'explication théorique, beaucoup d'empirisme existe encore sur ce type d'approche;

[1] Même si, notamment lorsque on s'intéresse seulement à un petit nombre de valeurs propres, des techniques d'approximation permettant de réduire la complexité algorithmique existent, elle reste toujours sur-linéaire dans le cas qui nous intéressent.

il nous a paru important de présenter une étude mathématique rigoureuse de cette approximation et en particulier de proposer un estimateur a posteriori qui pourrait permettre de vérifier l'hypothèse d'adiabaticité faite sur certaines variables et qui constitue la base du processus d'approximation.

Le travail théorique [38, 39] a été confirmé [39] par une simulation numérique partant d'un code de Claude Leforestier du Laboratoire Structure et Dynamique des Systèmes Moléculaires et Solides de l'Université de Montpellier 2.

2.1 Construction de l'hamiltonien nucléaire

Le problème qu'on se propose de résoudre est donc la recherche des fonctions et valeurs propres de l'hamiltonien nucléaire (voir (2.4.9)). En accord avec la remarque 2.4.1 on considérera que les seuls degrés de liberté des noyaux sont leur coordonnées.

L'hypothèse de Born-Oppenheimer (voir partie I, section 2.4) amène à un hamiltonien nucléaire de la forme $H = T + V$ où V désigne un potentiel (i.e. une fonction des coordonnées nucléaires) supposé connu dans les cas qui nous intéressent (soit empiriquement soit par calcul électronique préalable, par exemple Hartree-Fock) et T est l'opérateur d'énergie cinétique.

Une étape importante dans l'écriture de l'hamiltonien nucléaire est le choix du système de coordonnées. Notre système sera supposé isolé et sans rotation ; l'équation de Schrödinger (indépendante du temps) est à considérer alors dans l'ensemble des degrés de liberté internes, c'est à dire l'ensemble des $3N$ coordonnées de tous les N noyaux privé des trois coordonnées de translation et trois angles de rotation (mouvement solide), ce qui résulte en $3N-6$ coordonnées indépendantes. Une façon générale (qu'on ne détaillera pas ici) d'écrire ceci est par exemple le *formalisme d'Eckart* (voir [6, 7]).

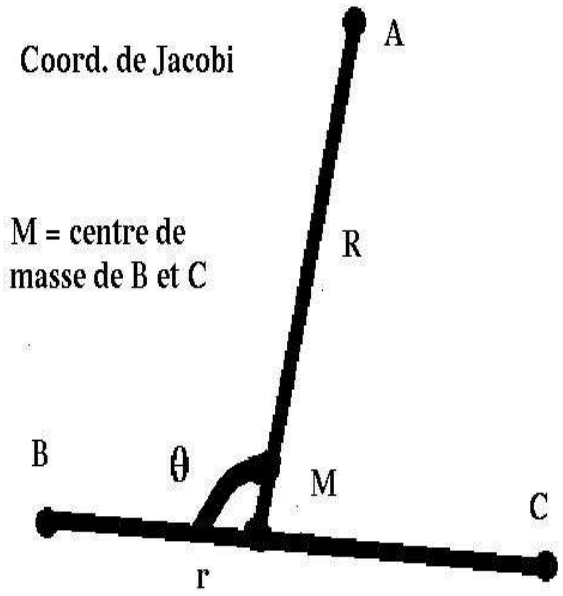
Le système qu'on étudiera par la suite est une molécule triatomique, ayant donc 3 degrés de liberté internes. On notera par A, B et C les trois particules (noyaux), de masses m_A, m_B et m_C. Un choix très classique pour le système de coordonnées est celui des *coordonnées de Jacobi* [32, 33] ($r = BC$, $R = AM$, $\theta = \widehat{BMA}$), voir figure 2.1.1.

L'opérateur Hamiltonien s'écrit alors

$$H_0 = -\frac{\hbar^2}{2\mu}\frac{\partial^2}{\partial R^2} - \frac{\hbar^2}{2\mu_{BC}}\frac{\partial^2}{\partial r^2} - \frac{\hbar^2}{2I}\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\sin\theta\frac{\partial}{\partial\theta} + V(R,r,\theta) \qquad (2.1.1)$$

avec les notations

$$\frac{1}{\mu} = \frac{1}{m_A} + \frac{1}{\mu_{BC}},\ \frac{1}{\mu_{BC}} = \frac{1}{m_B} + \frac{1}{m_C},\ \frac{1}{I} = \frac{1}{\mu R^2} + \frac{1}{\mu_{BC} r^2}$$



TAB. 2.1.1:

Afin de simplifier l'écriture des équations on fera le changement d'échelle nécessaire sur les variables R et r pour transformer les deux premiers termes dans (2.1.1) en $-\partial_{rr}-\partial_{RR}$. Notons par $f(R,r)$ le coefficient devant le troisième terme en (2.1.1) et faisons le changement de variable $z = \cos\theta$. On peut alors écrire

$$H = -\partial_{RR} - \partial_{rr} - f(R,r)\partial_z(1-z^2)\partial_z + V(R,r,z) = T_{R,r,z} + V. \qquad (2.1.2)$$

Pour des raisons physiques liées à la localisation des noyaux, les fonctions propres de H qui nous intéressent ont une décroissance très rapide au-delà d'un rectangle tridimensionnel centré dans la "position d'équilibre" du système. Par souci d'efficacité, les approches numériques pour la diagonalisation de H utiliseront cette information et consisteront à chercher les fonctions propres dans $L^2(\Omega)$ où Ω est une partie bornée de $\mathbb{R}^3$. Afin de rendre les notations plus lisibles, on choisira Ω égal à $]-1,1[^3$; les coordonnées R et r sont donc à considérer comme des déviations par rapport à une position d'équilibre ($\theta = \arccos(z)$ garde sa signification physique dans la définition des coordonnées de Jacobi). Enfin, pour des raisons évoquées ci-dessus, les conditions aux limites pour les fonctions d'onde seront zéro pour toute valeur de R ou r égale a ± 1.

Remarque 2.1.5 *La justification mathématique des conditions aux limites*

repose sur la définition de la forme bilinéaire associée à H ($T_{R,r,z}$)

$$a(u,v) = \int_{]-1,1[^3} \partial_x u\, \partial_x v + \partial_y u\, \partial_y v + (1-z^2)\partial_z u\, \partial_z v \qquad (2.1.3)$$

qui nous conduit à considérer l'espace $\{u \in L^2(]-1,1[^3); a(u,u) < \infty\}$ *ou, de manière équivalente*

$$\{u \in L^2(]-1,1[^3); u_x, u_y\ , \sqrt{1-z^2}u_z \in L^2(]-1,1[^3)\}.$$

Pour prendre en compte les conditions aux limites, on ajoute dans la définition de l'espace ci-dessus $u(\pm 1,r,z) = u(R,\pm 1,z) = 0$ *ce qui nous amène a définir*[2]

$$X_0^1 = \{u \in L^2(]-1,1[^3); u_x,\ u_y,\ \sqrt{1-z^2}u_z \in L^2(]-1,1[^3),$$
$$u(\pm 1,r,z) = u(R,\pm 1,z) = 0,\ R,r,z \in]-1,1[\}. \qquad (2.1.4)$$

Il n'y a pas de conditions à la limite sur la frontière $z = \pm 1$ *car, selon [12] page 69, les conditions* $u \in L^2$ *et* $\sqrt{1-z^2}\ u_z \in L^2$ *ne permettent pas de définir des valeurs de* u *en* $z = \pm 1$.

On est donc à même d'énoncer le problème à résoudre: trouver les fonctions $\Psi \in H^1(\Omega)$ et les valeurs $E \leq E_{MAX}$ solutions de

$$H\Psi = E\Psi,\ \Psi \in X_0^1,\ \|\Psi\|_{L^2(\Omega)} = 1. \qquad (2.1.5)$$

Il s'agit donc de trouver les premiers modes propres de l'hamiltonien; pour ce faire, on utilise une méthode itérative du type *Lanczos* qui repose sur le calcul d'une suite de vecteurs $\{\psi_n\}_n$ définis de manière récurente par $\psi_{n+1} = (H - \alpha_{n+1})\psi_n - \beta_n\ \psi_{n-1}$, où α_{n+1}, β_n sont des coefficients réels, $n = 1,2,...$ Du point de vue de l'efficacité la partie la plus coûteuse est le calcul de $H\psi_n$, car il n'existe pas en général de base de discrétisation dans laquelle aussi bien $T_{R,r,z}$ que $V(R,r,z)$ aient une forme (matricielle) simple; en effet si on choisit par exemple comme base les modes propres de la partie cinétique, $T_{R,r,z}$ sera alors diagonale mais on constate que la matrice du potentiel $V(R,r,z)$ est pleine. La solution est fournie par la construction de bases bien adaptées pour T et pour V et par leur utilisation alternative.

2.2 Présentation du procédé de réduction adiabatique

2.2.1 Construction des bases adaptées

On a vu dans la section précédente que pour appliquer facilement l'hamiltonien à un vecteur donné il est besoin d'une base convenable aussi bien

[2] voir dans la section 2.3 la justification de la notation

pour le potentiel que pour l'opérateur énergie cinétique. En général une base idéale pour ce calcul n'existe pas mais certaines équivalences de bases sont à prendre en compte lors de l'application de l'hamiltonien.

L'opérateur d'énergie cinétique étant $T_{R,r,z} = -\partial_{RR} - \partial_{rr} - f(R,r)\partial_z(1-z^2)\partial_z$, la base la plus naturelle qu'on peut lui associer est l'ensemble des produits tensoriels $\varphi_{k,\ell,n}(R,r,z)$ des fonctions propres sur $]-1,1[$ des opérateurs ∂_{rr} ,∂_{RR} et $\mathbb{A} = \partial_z(1-z^2)\partial_z$,

$$\varphi_{k,l,n}(R,r,z) = \sin\Big(\frac{k\pi}{2}(R+1)\Big)\sin\Big(\frac{l\pi}{2}(r+1)\Big)L_n(z), \tag{2.2.6}$$

pour (k,l,n) en $\mathbb{N}^3$ et L_n et le "n"-ème polynôme de Legendre (voir [10] p.233 pour définition et propriétés). En effet, dans une telle base les opérateurs ∂_{rr} et ∂_{RR} sont diagonaux et la matrice de $f(R,r)\partial_z(1-z^2)\partial_z$ est bloc-diagonale car les L_n sont des fonctions propres de $\partial_z(1-z^2)\partial_z$.

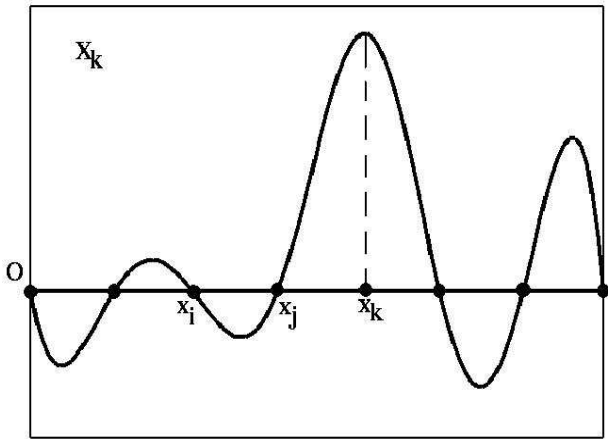
Pour mieux comprendre l'expression matricielle du potentiel dans une base on se placera pour commencer dans le cas uni-dimensionnel. Soit donc $\mathbb{P} = \{\chi_i(z)\ ;\ i = 1,..,N\}$ une base de fonctions deux à deux orthogonales; calculer la matrice d'un opérateur potentiel $\mathcal{V}(z)$ dans la base $\mathbb{P}$ revient à intégrer $\mathcal{V}\chi_i\ \chi_j$ sur le domaine de définition pour tous les χ_i,χ_j dans $\mathbb{P}$. Comme il est classique, ceci peut être fait par une formule d'intégration numérique. La recherche de la formule d'intégration la plus adaptée pour une base donnée part du constat que, dans les cas qui nous intéressent, on peut trouver à partir des relations $(\phi_i,\phi_j) = \delta_{ij}$ une autre base $\mathbb{X} = \{X_i\ ;\ i = 1,..,N\}$ qui génère le même espace discret, des points ζ_i et des poids ω_i tel que $X_p(\zeta_n) = 0$ si $p \neq n$, $p,n = 1,...,N$, et la formule de quadrature:

$$\int f(z)\ dz \simeq \sum_{i=1}^{N} \omega_i f(\zeta_i) \tag{2.2.7}$$

soit exacte pour tout produit $\chi_i\chi_j$, $i,j = 1,...,N$. On appellera ceci une **base localisée** [3]). Le changement de base $X_j = \sum_{i=1}^{N} U_{jn}\chi_n$ sera déterminé par une matrice unitaire U, où $U_{ij} = \sqrt{\omega_i}\chi_j(\zeta_i)$.

Si on calcule la matrice de $\mathcal{V}$ dans cette base par la formule d'intégration

[3]Cette appellation est motivée par le constat que plus N est grand, plus les X_i ressemblent à des masses de Dirac. Lorsque la base P est polynômiale, les points ζ_i sont les valeurs propres de l'opérateur "multiplication par la coordonnée z" et les X_i sont les fonctions propres associées (l'opérateur étant diagonalisé dans la base donnée $\mathbb{P}$). Il est facile à voir que si l'espace engendré était tout le L^2 on aurait comme fonctions propres des masses de Dirac et des valeurs propres tous les points de l'intervalle: $zf(z) = \lambda f(z) \Rightarrow z = \lambda$ et $f = \delta(z-\lambda)$; δ est le symbole de Kronecker.



TAB. 2.2.2: La "k"-ème fonction de la base localisée

numérique (2.2.7) (dont on néglige pour l'instant l'erreur) on obtient :

$$\int X_i \mathcal{V} X_j \, dz \simeq \sum_{n=1}^{N} \omega_n X_i(\zeta_n) \mathcal{V}(\zeta_n) X_j(\zeta_n) = \sum_{n=1}^{N} \omega_n \delta_{in} \delta_{jn} \mathcal{V}(\zeta_n)$$

ce qui nous montre que $\mathcal{V}$ sera représenté par la matrice diag$(d_n \mathcal{V}(\zeta_n))$ (avec des constantes d_n qui dépendent de ω_n). On a ainsi trouvé une base dans laquelle $\mathcal{V}$ est diagonale (si on néglige l'erreur d'intégration).

Dans notre cas les bases unidimensionnelles employées dans (2.2.6) sont soit des sinusoïdes soit des bases de polynômes de Legendre ; dans le premier cas la formule d'intégration numérique correspondante par (2.2.7) est aux points équidistants et dans le deuxième cas celle-ci est la formule d'intégration de Gauss aux points de Gauss-Legendre; toutes les deux ont des propriétés d'approximation et d'interpolation optimales [10, 14] (par rapport au nombre de points employés).

2.2.2 Définition et diagonalisation de l'hamiltonien réduit

Le procédé de réduction adiabatique a pour but de construire à partir des fonctions $\varphi_{k,l,n}$ une base réduite dans laquelle le problème aux valeurs propres sur l'hamiltonien (2.1.2) sera résolu. La démarche est la suivante:

1. On identifie par une analyse en modes normaux autour de l'équilibre une variable spéciale pour le système qu'on va appeler **la coordonnée**

adiabatique [4]. Ici il s'agira de la variable z.

2. On considère l'hamiltonien obtenu en enlevant les termes cinétiques dans la coordonnée adiabatique. Celui-ci sera appelé *hamiltonien réduit*, il s'agit ici de:

$$H^r := -\partial_{RR} \ - \partial_{rr} + \ V \tag{2.2.8}$$

   On le diagonalise par une procédure **très rapide**. En effet on ramène le problème 3D à un petit nombre de problèmes 2D en fixant la valeur de la coordonnée adiabatique. C'est ici qu'intervient l'intuition physique, la variable adiabatique étant d'une certaine façon celle qui permet de décrire au mieux l'hamiltonien total par son action en des valeurs fixées.
3. Puisqu'on cherche les vecteurs propres ayant une énergie plus petite que E_{MAX}, on ne garde parmi les vecteurs propres calculés à l'étape 2 que ceux dont l'énergie est plus petite que $(1+\epsilon)E_{MAX}$ (où $\epsilon > 0$).
4. En tensorisant les vecteurs obtenus au point 3 avec des fonctions caractéristiques de la variable adiabatique on définit une ensemble réduit de vecteurs où l'on cherche à diagonaliser H.

Quelques précisions s'imposent sur la description du procédé. Supposons avoir choisi comme espace discret l'espace $X_{M,N}$ engendre par $\{\varphi_{k,\ell,n}; 1 \leq k,l \leq M, 0 \leq n \leq N\}$. Soit $\{\zeta_i\}_{1\leq i\leq N+1}$ les points d'intégration de la formule de Gauss-Legendre correspondante aux L_n, $n = 0,...,N$, c'est a dire les racines $\{\zeta_i\}_{1\leq i\leq N+1}$ du polynôme de Legendre L_{N+1} de degré $N+1$. A ces points on associe comme on a vu ci-dessus la base des polynômes caractéristiques de degré $\leq N$, $\{h_j\}_{1\leq j\leq N+1}$ tels que $h_j(\zeta_i) = \delta_{i,j}$. L'observation essentielle est que (si on néglige les erreurs d'intégration numérique) dans la base $\Big\{v_{k,l,n} = \sin\Big(\frac{k\pi}{2}(R+1)\Big)\sin\Big(\frac{l\pi}{2}(r+1)\Big)h_j(z); 1 \leq k,l \leq M, 0 \leq n \leq N\Big\}$ l'hamiltonien réduit est diagonal par blocs; en effet, le opérateurs ∂_{rr} et ∂_{RR} sont déjà diagonaux; de plus les éléments de matrice de l'opérateur potentiel V seront zéro pour des indices n, n' différents dans la variable z :

$$\begin{aligned}\int (v_{k,l,n} V v_{k',l',n'})(R,r,z)dRdrdz &= \sum_{i=1}^{N+1} \int (v_{k,l,n} V v_{k',l',n'})(R,r,\zeta_i)\rho_i dRdr \\ &= \sum_{i=1}^{N+1} \int \sin\Big(\frac{k\pi}{2}(R+1)\Big)\sin\Big(\frac{l\pi}{2}(r+1)\Big)\delta_{ni}\cdot V(R,r,\zeta_i)\cdot \\ &\quad \sin\Big(\frac{k'\pi}{2}(R+1)\Big)\sin\Big(\frac{l'\pi}{2}(r+1)\Big)\delta_{n'i}\rho_i dRdr = 0 \ (n \neq n')\end{aligned}$$

[4] Il convient de remarquer que, même si certaines similitudes existent, cette appellation n'est pas directement liée a l'interprétation classique de séparation de mouvements.

ou $\{\zeta_i\}_{1\leq i\leq N+1}$ sont les points de Gauss et $\{\rho_i\}_{1\leq i\leq N+1}$ sont les poids de la formule de quadrature de Gauss-Legendre a $N+1$ points.

D'une manière plus rigoureuse la diagonalisation de H^r est approchée par la diagonalisation de l'*hamiltonien localisé* H^r_δ défini par

$$(H^r_\delta\varphi,\psi) = \int_{]-1,1[^2} \sum_{i=1}^{N+1} \Big((\partial_R\psi\partial_R\varphi + \partial_r\psi\partial_r\varphi)(R,r,\zeta_i) + (V\psi\varphi)(R,r,\zeta_i) \Big) \rho_i dRdr.$$

Remarque 2.2.6 *Diagonaliser H^r_δ revient à diagonaliser sur $X_{M,0}$ plusieurs opérateurs 2D : $-\partial_{RR} - \partial_{rr} + V(.,.,\zeta_i)$ pour tout i, $1 \leq i \leq N+1$. En plus de la réduction de dimension par rapport au cas 3D, cette diagonalisation est trivialement parallélisable, et sera donc d'un coût bien moindre que celui de la diagonalisation de H.*

Notons par $\big(\Phi_{p,q,i}(R,r)\big)_{1\leq p,q\leq M}$ les vecteurs propres normalisés L^2 de $-\partial_{RR} - \partial_{rr} + V(.,.,\zeta_i)$ et par $\big(\Lambda_{p,q,i}\big)_{1\leq p,q\leq M}$ les valeurs propres correspondantes, $i = 1,...,N+1$. L'approximation finale du problème sera la recherche dans l'espace linéaire engendré par les fonctions à 3 variables $\Phi_{p,q,i}(R,r)h_i(z)$ qui correspondent aux valeurs propres $\Lambda_{p,q,i} \leq (1+\epsilon)E_{MAX}$ des valeurs et fonctions propres de H_δ défini par

$$\begin{aligned}(H_\delta\varphi,\psi) \quad &= \textstyle\int_{]-1,1[^3} \partial_R\psi\partial_R\varphi + \partial_r\psi\partial_r\varphi + (1-z^2)\partial_z\psi\partial_z\varphi dRdrdz \\ &\quad + \textstyle\int_{]-1,1[^2} \sum_{i=1}^{N+1} V(R,r,\zeta_i)(\psi\varphi)(R,r,\zeta_i)\rho_i dRdr \qquad (2.2.9)\end{aligned}$$

Remarque 2.2.7 *La méthode peut être étendue pour le traitement des cas avec plus de 3 variables par une application récursive de la procédure ci-dessus. En fait on considérera quelques variables comme adiabatiques jusqu'à ce qu'on arrive à des matrices facil à diagonaliser. On réfère à [35] pour un exemple avec quatre noyaux (6 variables).*

2.3 Résultats théoriques et expérimentations numériques

Cette partie reprend une étude [39] effectuée en collaboration avec Yvon Maday et soumise pour publication dans Mathematical Modelling and Numerical Analysis.

Mathematical Modelling and Numerical Analysis Will be set by the publisher
Modélisation Mathématique et Analyse Numérique

e-mail: maday@ann.jussieu.fr e-mail: turinici@asci.fr

NUMERICAL ANALYSIS OF THE ADIABATIC VARIABLE METHOD FOR THE APPROXIMATION OF THE NUCLEAR HAMILTONIAN

Yvon Maday[1,2] and Gabriel Turinici[1]

Abstract. Many problems in quantum chemistry deal with the computation of fundamental or excited states of molecules and lead to the resolution of eigenvalue problems. One of the major difficulties in these computations lies in the very large dimension of the systems to be solved. Indeed these eigenfunctions depend on $3n$ variables where n stands for the number of particles (electrons and/or nucleari) in the molecule. In order to diminish the size of the systems to be solved, the chemists have proposed many interesting ideas. Among those stands the adiabatic variable method; we present in this paper a mathematical analysis of this approximation and propose, in particular, an a posteriori estimate that might allow for verifying the adiabaticity hypothesis that is done on some variables; numerical simulations that support the a posteriori estimators obtained theoretically are also presented.

Résumé. De nombreux problèmes en chimie quantique portent sur le calcul d'états fondamentaux ou excités de molécules et conduisent à la résolution de problèmes aux valeurs propres. Une des difficultés majeures dans ces calculs est la très grande dimension des systèmes qui sont en présence lors des simulations numériques. En effet les modes propres recherchés sont fonctions de $3n$ variables ou n est le nombre de particules (électrons ou noyaux) de la molécule. Afin de réduire la dimension des systèmes à résoudre les chimistes fourmillent d'idées intéressantes qui permettent d'approcher le système complet. La méthode des variables adiabatiques entre dans ce cadre et nous présentons ici une étude mathématique rigoureuse de cette approximation. En particulier nous proposons un estimateur a posteriori qui pourrait permettre de vérifier l'hypothèse d'adiabaticité faite sur certaines variables ; des simulations numeriques qui implémentent cet estimateur sont aussi présentées.

AMS Subject Classification. 65N25, 35P15, 81V55.

The dates will be set by the publisher.

1. Introduction

One problem frequently encountered in computational quantum chemistry (cf. [8]- [13]) consists in the evaluation of the eigenmodes of some Hamiltonian operator corresponding to eigenvalues smaller than some prescribed value E_{MAX}.

Under the Born-Oppenheimer approximation the nuclear Hamiltonian operator can be written as $H = T + V$ where V stands for the potential multiplicative part (assumed to be known by a previous electronic ab-initio computation or by empirical means) and T is the kinetic (Laplace) operator.

The number of independent variables being important any argument leading to the simplification of the behavior of the solution allows to enlarge the class of molecules that can be treated.

Firstly it seems natural to introduce the first eigenmodes of the Laplace operator written in the coordinate system and search for the eigenmodes of the Hamiltonian operator in this modal basis. In order to do so we use some Lanczos-type iterative method which relies on the computation of a vector sequence $\{\psi_n\}_n$ defined recursively by:

$$\psi_{n+1} = c_0 H(\psi_n) - c_1\ \psi_{n-1}. \tag{1}$$

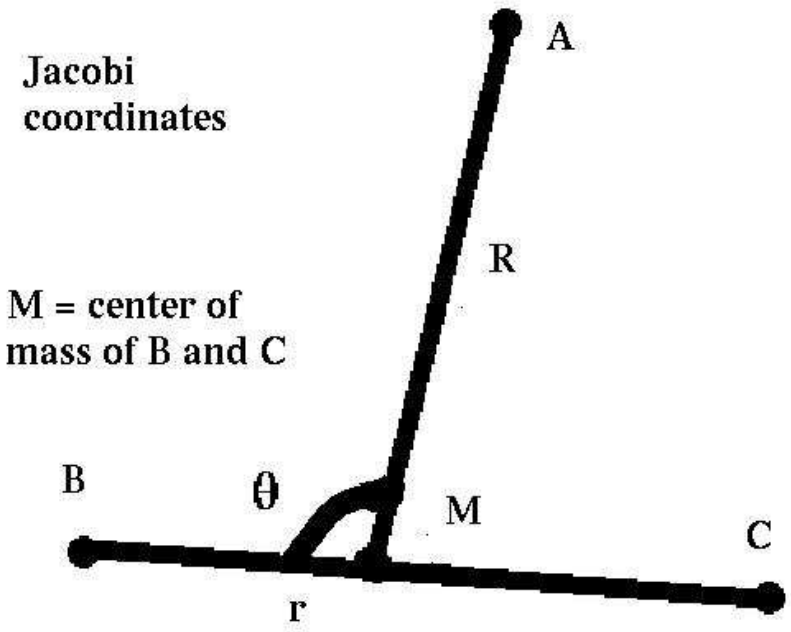
Keywords and phrases: a posteriori estimator, adiabatic variable method, computational quantum chemistry, nuclear hamiltonian

[1] ASCI, UPS 9029, bâtiment 506, Université Paris-Sud, 91405 Orsay

[2] Analyse numerique, Université Paris VI, place Jussieu, 75252 Paris cedex 05, France

In terms of CPU time the most expensive part is to apply the Hamiltonian operator H to ψ_n. In fact, even if the chosen basis is well adapted for the Laplace operator (such that it is diagonal), the potential operator matrix is full. In general we are interested in determining a large part of the spectrum, the size of the discretization basis (and hence the size of matrices involved) is usually so large that it forbids any computation. We are then lead to search for methods allowing us to further reduce the number of basis functions. The *pseudo-spectral adiabatic variable method* proposed in [8], [9] is one such pertinent discretization tool that seems to give quite good results in practice.

Its principle is presented below for a triatomic molecule.



Let the Laplace operator be written in Jacobi coordinates (R, r, θ) (cf. [8]), and let us assume that we want to find a function ψ on the open brick[1] $\Omega =]-1,1[^2 \times]0,\pi[$ of $\mathbb{R}^3$ such that :

$$\tilde{H}\psi = E\psi, \quad \text{with } \tilde{H} = \tilde{T}_{R,r,\theta} + V = -\partial_{RR} - \partial_{rr} - \frac{f(R,r)}{\sin\theta}\partial_\theta \sin\theta \partial_\theta + V, \tag{2}$$

where the function ψ has to satisfy

$$\psi(\pm 1, r, \theta) = \psi(R, \pm 1, \theta) = 0 \ , \|\psi\|_{L^2(\Omega)} = 1. \tag{3}$$

Then

1. We identify by a normal-mode analysis around the equilibrium position some special variable for our system named **the adiabatic variable**. Here it will be θ and we write the Hamiltonian using the coordinate transformation $z = \cos\theta$.

$$H = -\partial_{RR} - \partial_{rr} - f(R,r)\partial_z(1-z^2)\partial_z + V = T_{R,r,z} + V. \tag{4}$$

[1] The initial range for R, r is mapped by affine transformations into $]-1,1[$; the coordinates R, r are to be considered henceforth as relative deviations from some equilibrium position; note that the physical meaning of θ is preserved.

2. We consider the Hamiltonian operator obtained by removing the terms containing derivatives in the adiabatic variable; we call it **reduced Hamiltonian**, here it is

$$H^r := T_{R,r,z} \ - (-f(R,r)\partial_z(1-z^2)\partial_z) + \ V = -\partial_{RR} \ - \partial_{rr} + \ V \tag{5}$$

   and diagonalize it by a **fast** procedure. In fact the 3D problem is reduced to a small number of 2D problems by freezing the values of the adiabatic coordinate. It is here that the physical intuition comes into play, the adiabatic variable being in a certain way the one that allows us to accurately describe the total hamiltonian by its action in a small number of fixed values.
3. Since we are looking for eigenmodes with a corresponding energy smaller than E_{MAX}, we keep among the vectors obtained in step 2 only those with energy smaller than $(1+\epsilon)E_{MAX}$ (where $\epsilon > 0$).
4. We construct by tensor product of the vectors obtained in step 3 with characteristic functions of the adiabatic variable a reduced basis used to finally diagonalize the full hamiltonian operator H.

In practice this procedure gives good results. However the choice of the adiabatic variable(s) and/or coordinate system affects substantially its efficiency. Therefore it seems interesting to give some a priori estimates to help intuition in the choice of the adiabatic variable for a given system and to complement this analysis by a posteriori estimators so as to decide about its usefulness once the computation is over and also in order to confirm the choice of ϵ used in the truncation[2].

Before proceeding with the different error analysis, it is important to introduce the choice of the values of the adiabatic variable that are being frozen during step 2. These are the Gauss quadrature points for that variable. This choice can be justified by at least two reasons. The first one is that these points are optimal for the evaluation (through quadrature formulas) of integrals involved in the computation of the action of the potential over the vectors required in the Lanczos recurrence. The second argument is that this set of points is optimal for interpolating in the linear space of polynomials spanned by the first eigenmodes of the differential operator $\partial_z(1-z^2)\partial_z$ in the adiabatic variable, i.e. the Legendre polynomials $\{L_n\}_n$. The values we freeze are therefore the Gauss-Legendre points, namely the zeroes $\{\zeta_i\}_{1\le i\le N+1}$ of the Legendre polynomial L_{N+1} of degree $N+1$. It is classical to associate to these points a (localized) basis containing characteristic polynomials of degree $\le N$, $\{h_j\}_{1\le j\le N+1}$ such that $h_j(\zeta_i) = \delta_{i,j}$, $i,j = 1,...N+1$ (Kronecker symbol).

We introduce the interpolation operator $\mathcal{J}_N$ from $\mathcal{C}^0(]-1,1[)$ to $\mathcal{P}_N(]-1,1[)$ on these nodes. This operator has optimal approximation properties (cf. [1] Thm.13.2, p.299), that is for any real $\sigma > \frac{1}{2}$, there exists some constant $c > 0$ such that

$$\forall v \in H^\sigma(]-1,1[), \quad \|v - \mathcal{J}_N v\|_{L^2(]-1,1[)} \le cN^{-\sigma}\|v\|_{H^\sigma(]-1,1[)}. \tag{6}$$

2. A PRIORI ANALYSIS

We propose this analysis for the case of the triatomic system (2) - (3) where for simplicity we set $f(R,r) \equiv 1$. This a priori analysis is not the main purpose of the paper and serves only as preliminary verification of the pertinency of the algorithm. More detailed analysis is presented in next section. As we have already seen, the discretization has 2 steps. Firstly we introduce the eigenfunctions of the operator $T_{R,r,z}$ on $L^2(]-1,1[^3)$, here $\varphi_{k,\ell,n}(R,r,z) = \sin(\frac{k\pi}{2}(R+1))\sin(\frac{\ell\pi}{2}(r+1))L_n(z)$ for (k,ℓ,n) in $\mathbb{N}^3$. We propose an initial discretization space $X_{M,N}$ spanned by $\varphi_{k,\ell,n}$ for $1 \le k,\ell \le M$, $0 \le n \le N$. In the second step we diagonalize over $X_{M,0}$ the 2D operators $-\partial_{RR} - \partial_{rr} + V(.,.,\zeta_i)$ for each i, $1 \le i \le N+1$; we call $\left(\Phi_{p,q,i}\right)_{1\le p,q\le M}$ and $\left(\Lambda_{p,q,i}\right)_{1\le p,q\le M}$ the L^2 associated normalized eigenvectors and corresponding eigenvalues respectively.

We define some Sobolev-type spaces associated with the kinetic operator $T_{R,r,z}$. More precisely let X_0^s be the closure of $C_0^1(]-1,1[^3) \cap C^\infty(]-1,1[^3)$ in the domain of $(T_{R,r,z})^{s/2}$ endowed with its canonical norm. Theorem

[2]This "adiabatic reduction method" has some similarities with the dimension reduction method used in mechanics. See [14] for a presentation of this method and for adapted error estimators. However the method and the analysis technique are different.

5.6 from [5] and Theorem 2.3 from [4] tome 1 p.19 allow to describe X_0^s. We obtain for instance:

$$X_0^2 = \{u \in H_0^1(]-1,1[^3); \partial_{RR}u, \partial_{rr}u, \partial_{Rr}u, \sqrt{1-z^2}\partial_{Rz}u, \sqrt{1-z^2}\partial_{rz}u, (1-z^2)\partial_{zz}u \in L^2(]-1,1[^3)\}. \quad (7)$$

Next we introduce the linear space $\mathcal{E}_\delta$ spanned by $\Phi_{p,q,i}(R,r)h_i(z)$ (3D functions) that correspond to eigenvalues $\Lambda_{p,q,i} \leq (1+\epsilon)E_{MAX}$. The final approximation of our problem then consists in searching in $\mathcal{E}_\delta$ the eigenfunctions of the operator H_δ defined for all $\psi, \varphi \in X_0^1$ as follows

$$\begin{aligned}(H_\delta\varphi,\psi) = \int_{]-1,1[^3} \partial_R\psi\partial_R\varphi + \partial_r\psi\partial_r\varphi + (1-z^2)\partial_z\psi\partial_z\varphi dRdrdz \\ + \int_{]-1,1[^2} \sum_{i=1}^{N+1} V(R,r,\zeta_i)(\psi\varphi)(R,r,\zeta_i)\rho_i dRdr, \end{aligned} \quad (8)$$

where $\{\rho_i\}_{1\leq i\leq N+1}$ are the weights of the Gauss-Legendre quadrature formula.

Remark 2.1. It is interesting to note that $\Phi_{p,q,j}(R,r)h_j(z)$, $1 \leq p,q \leq M$, $1 \leq j \leq N+1$ are the eigenfunctions on $X_{M,N}$ of the operator H_δ^r defined as follows

$$(H_\delta^r\varphi,\psi) = \int_{]-1,1[^2} \sum_{i=1}^{N+1} \left((\partial_R\psi\partial_R\varphi + \partial_r\psi\partial_r\varphi)(R,r,\zeta_i) + V(R,r,\zeta_i)(\psi\varphi)(R,r,\zeta_i) \right) \rho_i dRdr.$$

This operator is a kind of localized hamiltonian in the points ζ_i (chemists use to note it $H(R,r,z=\zeta_i), i = 1, N+1$) made up by contributions from each ζ_i point.

Remark 2.2. The method can be readily extended for the case of more than 3 variables by recursively applying the above procedure. In fact we consider some of them as adiabatic until we reach a matrix that can be easily diagonalized. See [11] for an example in the case of 6 variables.

We write our problem in the form:

$$find \;\; u = (\psi,\lambda) \in L^2(]-1,1[^3) \times \mathbb{R} \;\; such\ that\ F(u) = 0, \quad (9)$$

where F is the smooth (C^1) function from $L^2(]-1,1[^3) \times \mathbb{R}$ into the dual $(X_0^2)^* \times \mathbb{R}$ of $X_0^2 \times \mathbb{R}$ given by:

$$\begin{aligned} < F(\psi,\lambda), (\varphi,\mu) >_{(X_0^2)^*\times\mathbb{R}, X_0^2\times\mathbb{R}} &= \int_{]-1,1[^3} \psi(H\varphi - \lambda\varphi) + \mu\left(\int_{]-1,1[^3} \psi^2 - 1 \right) \\ &= \int_{]-1,1[^3} \psi(T_{R,r,z}\varphi + V\varphi - \lambda\varphi) + \mu\left(\int_{]-1,1[^3} \psi^2 - 1 \right). \end{aligned} \quad (10)$$

It is easy to see that $F(\psi,\lambda) = 0$ is equivalent to (2)-(3). Moreover if λ_0 is a simple (i.e. of multiplicity 1) eigenvalue of (2) corresponding to an eigenvector ψ_0 (chosen with L^2-norm equal to 1) and $V \in L^\infty$ (which is never a restriction in practice), then, applying the Fredholm alternative as proven in Appendix A we conclude that $DF(\psi_0,\lambda_0)$ is an isomorphism from $L^2(]-1,1[^3) \times \mathbb{R}$ to $(X_0^2)^* \times \mathbb{R}$. In order to avoid technical difficulties we will suppose, in what follows, that all eigenvalues under consideration are simple and $V \in L^\infty$.

Let Π_δ be the projector to $\mathcal{E}_\delta$ associated with $T_{R,r,z}$ that is for all $v \in X_0^2$, $\Pi_\delta v$ is the element of $\mathcal{E}_\delta$ that verifies

$$\forall\ u_\delta \in \mathcal{E}_\delta \ : \int_{]-1,1[^3} T_{R,r,z}(v - \Pi_\delta v)u_\delta = 0. \quad (11)$$

We define functions F_δ from $L^2 \times \mathbb{R}$ into $(X_0^2)^* \times \mathbb{R}$ by the formulas:

$$< F_\delta(\psi,\lambda),(\varphi,\mu) >_{(X_0^2)^*\times\mathbb{R},X_0^2\times\mathbb{R}} = \int_{]-1,1[^3} \psi(H_\delta - \lambda)(\Pi_\delta\varphi) + \mu\left(\int_{]-1,1[^3} \psi^2 - 1\right) + \int_{]-1,1[^3} \psi T_{R,r,z}(\varphi - \Pi_\delta\varphi). \tag{12}$$

Proposition 2.3. *The solutions of $F_\delta(\psi_\delta,\lambda_\delta) = 0$ are exactly eigenfunctions of H_δ on $\mathcal{E}_\delta$.*

Proof. Choose first φ orthogonal to $\mathcal{E}_\delta$ with respect to $T_{R,r,z}$ and $\mu = 0$ and obtain $\psi \in \mathcal{E}_\delta$; then choosing $\varphi = 0$ yields $\|\psi\|_{L^2} = 1$ and finally $\varphi \in \mathcal{E}_\delta$ and $\mu = 0$ proves that

$$(H_\delta\psi,\varphi) = (\lambda\psi,\varphi),\ \forall\varphi \in \mathcal{E}_\delta. \tag{13}$$

□

We are now applying Theorem 6.1 ([2] vol 5 p.530) to show that $\|F_\delta(\psi_0,\lambda_0)\|_{(X_0^2)^*\times\mathbb{R}}$ is an upper bound (modulo some constant) for the error between (ψ_0,λ_0) and $(\psi_\delta,\lambda_\delta)$. More precisely there exists a constant $C > 0$ that does not depend on M,N or E_{MAX} and a neighborhood V of δ_0 (defined as the "limit" value where $F_{\delta_0} = F$) such that for all $\delta \in V\backslash\{\delta_0\}$ and (ψ_0,λ_0) such that $F(\psi_0,\lambda_0) = 0$ there exists $(\psi_\delta,\lambda_\delta)$ solution of $F_\delta(\psi_\delta,\lambda_\delta) = 0$ such that:

$$\|\psi_0 - \psi_\delta\|_{L^2(\Omega)} + |\lambda_0 - \lambda_\delta| \le C\|F_\delta(\psi_0,\lambda_0)\|_{(X_0^2)^*\times\mathbb{R}}. \tag{14}$$

It remains to evaluate the right hand side of (14) in order to obtain the a priori upper bound for the error between the exact and the discrete solution.

Since (ψ_0,λ_0) is a solution to our problem and by the definition (11) of the projector Π_δ we obtain for all $(\varphi,\mu) \in (X_0^2) \times \mathbb{R}$:

$$< F_\delta(\psi_0,\lambda_0),(\varphi,\mu) >_{(X_0^2)^*\times\mathbb{R},X_0^2\times\mathbb{R}} = \int_{]-1,1[^3} \psi_0(H_\delta - H)(\Pi_\delta\varphi)\ + (\psi_0 - \Pi_\delta\psi_0)T_{R,r,z}(\varphi - \Pi_\delta\varphi). \tag{15}$$

Definition. We state that N,M and E_{MAX} are chosen in a coherent manner and denote $N^2 \simeq M^2 \simeq E_{MAX}$ if there exists 3 constants independent of the discretization such that $N^2 \le c_1M^2 \le c_2E_{MAX} \le c_3N^2$.

We will make use in the following of some (optimal) approximation properties of projector Π_δ :

Lemma 2.4. *Assume that $N^2 \simeq M^2 \simeq E_{MAX}$. Then for any $b \ge 1 \ge a \ge 0$ there exists a constant $c(a,b)$ such that:*

$$\forall v \in X_0^b :\ \|v - \Pi_\delta v\|_{X_0^a} \le c(a,b)(\epsilon_\delta)^{b-a}\|v\|_{X_0^b}. \tag{16}$$

where ϵ_δ is $\max\left\{\frac{1}{N},\frac{1}{M},\frac{1}{\sqrt{E_{MAX}}}\right\}$

Proof. See the appendix A. □

Using lemma 2.4 the optimality properties of the interpolation operator $\mathcal{I}_N$ (stated in (6)) we obtain from (14) and (15) the following a priori estimate :

Theorem 2.5. *Let (ψ_0, λ_0) be a simple eigenmode of (2)-(3) and $s \geq 1$, $t > \frac{1}{2}$ such that $\psi_0 \in X_0^s$ and $V\psi_0 \in L^2(]-1,1[^2; H^t(]-1,1[))$. Then there exists a constant $C(s,t) > 0$ such that for each δ there exists a solution of $F_\delta(\psi_\delta, \lambda_\delta) = 0$[3] such that:*

$$\|\psi_0 - \psi_\delta\|_{L^2} + |\lambda_0 - \lambda_\delta| \leq C(s,t) \left((\epsilon_\delta)^s \|(\psi_0, \lambda_0)\|_{X_0^s \times \mathbb{R}} + N^{-t} \|V\psi_0\|_{L^2(]-1,1[^2; H^t(]-1,1[))} \right). \tag{17}$$

Proof. Inserting in (14) the equality (15) and using the definition of the norm in $(X_0^2)^* \times \mathbb{R}$ one obtains

$$\begin{aligned} \|\psi_0 - \psi_\delta\|_{L^2(\Omega)} + |\lambda_0 - \lambda_\delta| &\leq C \sup_{\|\varphi\|_{X_0^2}=1} \int_{]-1,1[^3} \psi_0 (H_\delta - H)(\Pi_\delta \varphi) \ + (\psi_0 - \Pi_\delta \psi_0) T_{R,r,z}(\varphi - \Pi_\delta \varphi) \\ &\leq C \sup_{\|\varphi\|_{X_0^2}=1} \int_{]-1,1[^3} (V\psi_0 - (Id_{\mathbb{R}^2} \otimes \mathcal{J}_N) V\psi_0) \Pi_\delta \varphi + (\psi_0 - \Pi_\delta \psi_0) T_{R,r,z}(\varphi - \Pi_\delta \varphi) \\ &\leq \sup_{\|\varphi\|_{X_0^2}=1} \int_{]-1,1[^3} (V\psi_0 - (Id_{\mathbb{R}^2} \otimes \mathcal{J}_N) V\psi_0) \Pi_\delta \varphi + \sup_{\|\varphi\|_{X_0^2}=1} \int_{]-1,1[^3} (\psi_0 - \Pi_\delta \psi_0) T_{R,r,z}(\varphi - \Pi_\delta \varphi) \end{aligned} \tag{18}$$

By the definition of the projector Π_δ the second term in the right hand side of (18) equals

$$\sup_{\|\varphi\|_{X_0^2}=1} \int_{]-1,1[^3} (\psi_0 - \Pi_\delta \psi_0) T_{R,r,z} \varphi, \tag{19}$$

and can be upper bounded by

$$\sup_{\|\varphi\|_{X_0^2}=1} \|\psi_0 - \Pi_\delta \psi_0\|_{L^2} \|T_{R,r,z}(\varphi)\|_{L^2} \leq \|\psi_0 - \Pi_\delta \psi_0\|_{L^2} \leq c(0,s) \epsilon_\delta^s \|\psi_0\|_{X_0^s}. \tag{20}$$

Using (6) and the stability of the projector Π_δ one can now bound the first term in the right hand side of (18) and obtain the conclusion of the theorem. □

Remark 2.6. If V is smooth enough, it is obvious that the norms $\|\psi_0\|_{X_0^p}$, $\|V\psi_0\|_{L^2(]-1,1[^2; H^{2p}(]-1,1[))}$ and $\|V\psi_0\|_{H^{2p}(]-1,1[^2; L^2(]-1,1[))}$ are upper bounded by $c|\lambda_0|^p$ so that for the natural choice $N^2 \simeq M^2 \simeq E_{MAX}$ the convergence rate scales as $c(p) \left(\frac{\lambda_0}{N^2} \right)^p$.

3. A POSTERIORI ANALYSIS OF THE METHOD

Let us still focus on the case of the triatomic system (2) and (3), and let us consider now an *a posteriori* error analysis. The goal of such a tool is to asses the approximation once the computation is done. We are working as before on the formulation $F(u) = 0$ defined in (10).

The result (17) show that for any simple eigenmode $u_0 = (\psi_0, \lambda_0)$ of (2)-(3), there exists an eigenmode $(\psi_\delta, \lambda_\delta)$ which is close enough. To know more precisely how close they are, one uses results derived from [9] which allow to prove that under certain hypothesis, $F(u)$ is an estimator for the error between u_0 and u. We shall make use of this abstract result in the following form:

Theorem 3.1. *Let Z,Y be two Hilbert spaces and $F \in C^1(Z,Y)$. Let u_0 be a solution of $F(u) = 0$ such that $DF(u_0) \in Isom(Z,Y)$ and moreover assume DF satisfies a Lipschitz-type property*

$$\exists \epsilon_{u_0} > 0 : \quad \|[DF(u_0) \ - \ DF(u_0 + tU)] \ U\|_Y \leq \ ct\|U\|_Z^2, \ \forall \ 0 < t < \epsilon_{u_0}, \ \forall \ U \in Z, \ \|U\| < \epsilon_{u_0}. \tag{21}$$

[3] In fact since the eigenmode (ψ_0, λ_0) is simple for δ close enough to δ_0 the problem $F_\delta(\psi_\delta, \lambda_\delta) = 0$ will have only two solutions with corresponding eigenvalues close to λ_0 that is $(\psi_\delta, \lambda_\delta)$ and $(-\psi_\delta, \lambda_\delta)$.

Then there exists some $R > 0$ ($R = min\left\{\frac{1}{2}\|DF(u_0)^{-1}\|^{-1}_{\mathcal{L}(Y,Z)}\ ,\ \|DF(u_0)\|_{\mathcal{L}(Z,Y)}\right\}$) *such that for all* $u \in B(u_0, R)$*:*

$$\frac{1}{2}\|DF(u_0)\|^{-1}_{\mathcal{L}(Z,Y)} \cdot \|F(u)\|_Y \leq \|u - u_0\|_Z \leq 2\|DF(u_0)^{-1}\|_{\mathcal{L}(Y,Z)} \cdot \|F(u)\|_Y. \tag{22}$$

Choose $Z = L^2(]-1,1[^3) \times \mathbb{R}$ and $Y = (X_0^2)^* \times \mathbb{R}$ and note that DF obviously satisfies the hypothesis (21) of Theorem 3.1 ; recalling that $DF(\psi_0, \lambda_0) \in Isom(L^2(]-1,1[^3) \times \mathbb{R}, (X_0^2)^* \times \mathbb{R})$ we obtain from Theorem 3.1 :

$$c\|F(\psi_\delta, \lambda_\delta)\|_Y \leq \|\psi_0 - \psi_\delta\|_{L^2(]-1,1[^3)} + |\lambda_0 - \lambda_\delta| \leq C\|F(\psi_\delta, \lambda_\delta)\|_Y \tag{23}$$

for two positive constants c and C.

We write easily

$$\|F(\psi_\delta, \lambda_\delta)\|_Y = \sup_{(\varphi,\mu)\in X_0^2\times\mathbb{R}} \frac{\int_{]-1,1[^3}(T_{R,r,z}\psi_\delta + V\psi_\delta - \lambda_\delta\psi_\delta)\varphi}{\|(\varphi,\mu)\|_{X_0^2\times\mathbb{R}}}, \tag{24}$$

(note that μ does not enter in this estimate). Define π_M as the L^2-projection operator from $L^2(]-1,1[^3)$ to $X_{M,0}$; we will use the following approximation property of π_M (cf. [15] Ch.9, p.278): for any $\sigma \geq 0$ there exists a constant $c > 0$ depending only of σ such that

$$\forall v \in H^\sigma(]-1,1[^2; L^2(]-1,1[)) \quad \|v - \pi_M v\|_{L^2(]-1,1[^2;L^2(]-1,1[))} \leq cN^{-\sigma}\|v\|_{H^\sigma(]-1,1[^2;L^2(]-1,1[))} \tag{25}$$

By defining φ_{MN} as the L^2 projection of φ on X_{MN} we obtain

$$\begin{aligned}
\|F(\psi_\delta, \lambda_\delta)\|_Y &= \sup_{\|\varphi\|_{X_0^2}=1} \int_{]-1,1[^3}((V\psi_\delta - \pi_M \otimes \mathcal{J}_N(V\psi_\delta))\varphi + (T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta)\varphi \\
&= \sup_{\|\varphi\|_{X_0^2}=1} \int_{]-1,1[^3}((V\psi_\delta - \pi_M \otimes \mathcal{J}_N(V\psi_\delta))\varphi + (T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta)\varphi_{MN} \\
&\leq \sup_{\|\varphi\|_{X_0^2}=1} \int_{]-1,1[^3}((V\psi_\delta - \pi_M \otimes \mathcal{J}_N(V\psi_\delta))\varphi \\
&+ \sup_{\varphi\in X_0^2, \|\varphi\|_{X_0^2}=1} \int_{]-1,1[^3}(T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta)\varphi_{MN},
\end{aligned} \tag{26}$$

where we have used the fact that $T_{R,r,z}\psi_\delta \in X_{MN}$ between the first and second line. The **first contribution** in the right hand side measures the approximation resulting from the reduction of the action of V to X_{MN}. By (6) - (25) it can be bounded as follows

$$\begin{aligned}
&\sup_{\varphi\in X_0^2, \|\varphi\|_{X_0^2}=1} \left| \int_{]-1,1[^3}(V\psi_\delta - \pi_M \otimes \mathcal{J}_N(V\psi_\delta))\varphi\right| \\
&\leq c(N^{-s}\|V\psi_\delta\|_{L^2(]-1,1[^2;H^s(]-1,1[)} + M^{-\sigma}\|V\psi_\delta\|_{H^\sigma(]-1,1[^2;L^2(]-1,1[))}),
\end{aligned} \tag{27}$$

for all $\sigma \geq 0$ and $s > \frac{1}{2}$ such that

$$V\psi_\delta \in L^2(]-1,1[^2; H^s(]-1,1[)) \cap H^\sigma(]-1,1[^2; L^2(]-1,1[). \tag{28}$$

The **second contribution** in the right hand side of (26) represents the loss of information resulting from neglecting in X_{MN} the eigenmodes $\Phi_{p,q,i}h_i$ having energy larger than $(1+\epsilon)E_{MAX}$. It is *this* contribution that allows us to asses the adiabaticity of the chosen coordinate system since it measures the amount of energy

contained in the projection of $(T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta)$ on the rejected eigenmodes. Indeed its projection on all other eigenmodes is zero by the definition of ψ_δ . This leads us to

$$\begin{aligned}
&\sup_{\varphi\in X_0^2,\|\varphi\|_{X_0^2}=1}\int_{]-1,1[^3}(T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta)\varphi_{MN}\\
&= \sup_{\varphi\in X_0^2,\|\varphi\|_{X_0^2}=1}\int_{]-1,1[^3}(T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta)(\varphi_{MN} - \pi_{\mathcal{E}_\delta}(\varphi_{MN}))\\
&\leq \|T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta\|_{L^2}\sup_{\varphi\in X_0^2,\|\varphi\|_{X_0^2}=1}\|\varphi_{MN} - \pi_{\mathcal{E}_\delta}(\varphi_{MN})\|_{L^2}. \qquad (29)
\end{aligned}$$

In these estimates, $\pi_{\mathcal{E}_\delta}$ is the L^2 projection operator over the reduced space $\mathcal{E}_\delta$.

An upper bound for the last term is given by the

Lemma 3.2. *For any element φ_{MN} in $X_{M,N}$ the following estimate is true*

$$\|\varphi_{MN} - \pi_{\mathcal{E}_\delta}(\varphi_{MN})\|_{L^2}^2 \leq (\frac{1}{(1+\epsilon)E_{MAX}})^2\Big(\|(-\partial_{RR} - \partial_{rr})\varphi_{MN}\|_{L^2(]-1,1[^3)}^2 + \|V\|_{L^\infty}^2\|\varphi_{MN}\|_{L^2(]-1,1[^3)}^2\Big). \qquad (30)$$

Moreover for any $b \geq 0$ there exists a constant C independent of M,N,E_{MAX} such that

$$\|\varphi_{MN} - \pi_{\mathcal{E}_\delta}(\varphi_{MN})\|_{L^2} \leq C\Big(\frac{1}{\sqrt{E_{MAX}}}\Big)^b\|\varphi_{MN}\|_{X_0^b}. \qquad (31)$$

Proof. See the appendix A . □

From now on we suppose ϵ smaller than some fixed constant (usually less than 1). Using the stability of the L^2 projector on eigenmodes we obtain that there exists a constant $c > 0$ such that

$$\|\varphi_{MN} - \pi_{\mathcal{E}_\delta}(\varphi_{MN})\|_{L^2} \leq (\frac{c}{E_{MAX}})(1 + \|V\|_{L^\infty})\|\varphi\|_{X_0^2} \leq \frac{c(V)}{E_{MAX}}\|\varphi\|_{X_0^2}. \qquad (32)$$

This allows us to write first

$$\begin{aligned}
&\sup_{\|\varphi\|_{X_0^2}=1}\int_{]-1,1[^3}(T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta)\varphi_{MN}\\
&\leq \tfrac{c(V)}{E_{MAX}}\|T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta\|_{L^2(]-1,1[^3)}. \qquad (33)
\end{aligned}$$

Recalling the definition of ψ_δ, we have

$$\pi_{\mathcal{E}_\delta}(T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta) = 0. \qquad (34)$$

from the definition of the eigenmodes that span $\mathcal{E}_\delta$, we also have

$$(Id - \pi_{\mathcal{E}_\delta})((-\partial_{RR} - \partial_{rr})\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta) = 0, \qquad (35)$$

hence

$$T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta = (Id - \pi_{\mathcal{E}_\delta})(\partial_z(1-z^2)\partial_z\psi_\delta), \qquad (36)$$

so that

$$\sup_{\|\varphi\|_{X_0^2}=1}\int_{]-1,1[^3}(T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta)\varphi_{MN} \leq \frac{c(V)}{E_{MAX}}\|(Id - \pi_{\mathcal{E}_N})(\partial_z(1-z^2)\partial_z\psi_\delta)\|_{L^2(]-1,1[^3)} \qquad (37)$$

Combining this inequality with (27) allows us to state the following result:

Theorem 3.3. *Let $\sigma \geq 0$, $s > \frac{1}{2}$ be such that $V\psi_\delta \in L^2(]-1,1[^2; H^s(]-1,1[)) \cap H^\sigma(]-1,1[^2; L^2(]-1,1[))$. Then there exists two constants c and $c(V)$ such that*

$$\|\psi_0 - \psi_\delta\|_{L^2(]-1,1[^3)} + |\lambda_0 - \lambda_\delta| \leq \frac{c(V)}{E_{MAX}} \|(Id - \pi_{\mathcal{E}_N})(\partial_z(1-z^2)\partial_z\psi_\delta)\|_{L^2(]-1,1[^3)}$$
$$+c(M^{-\sigma}\|V\psi_\delta\|_{H^\sigma(]-1,1[^2;L^2]-1,1[)} + N^{-s}\|V\psi_\delta\|_{L^2(]-1,1[^2;H^s]-1,1[)}) \tag{38}$$

and

$$\frac{c}{\sup(M,N)^2} \quad \|(Id - \pi_{\mathcal{E}_\delta})(\partial_z(1-z^2)\partial_z\psi_\delta)\|_{L^2(]-1,1[^3)} \leq \Big(\|\psi_0 - \psi_\delta\|_{L^2(]-1,1[^3)} + |\lambda_0 - \lambda_\delta|\Big)$$
$$+c(M^{-\sigma}\|V\psi_\delta\|_{H^\sigma(]-1,1[^2;L^2]-1,1[)} + N^{-s}\|V\psi_\delta\|_{L^2(]-1,1[^2;H^s]-1,1[)}). \tag{39}$$

Proof. Only (38) has been proven, we are going to prove (39) after having noticed that the first term in the right hand side of (38) accounts for the reliability of the adiabatic variable reduction and the second accounts for the choice of the filtering frequency (M,N)[4]. All we have to prove is that the estimator in the right hand side of (38) is not too large. For φ in X_0^2 denote φ_{MN} as its projection on X_{MN} ; then for all $\mu \in \mathbb{R}$

$$\int_{]-1,1[^3} (T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta)\varphi_{MN} = < F(\psi_\delta,\lambda_\delta), (\varphi,\mu) > - \int_{]-1,1[^3} ((V\psi_\delta - \pi_M \otimes \mathcal{J}_N(V\psi_\delta))\varphi,$$

so that

$$\sup_{\|\varphi\|_{X_0^2}=1} \int_{]-1,1[^3} (T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta)\varphi_{MN}$$
$$\leq \sup_{\|\varphi\|_{X_0^2}=1} < F(\psi_\delta,\lambda_\delta), (\varphi,\mu) > + \sup_{\|\varphi\|_{X_0^2}=1} \int_{]-1,1[^3} ((V\psi_\delta - \pi_M \otimes \mathcal{J}_N(V\psi_\delta))\varphi. \tag{40}$$

Using the upper bound in (27) we obtain

$$\sup_{\|\varphi\|_{X_0^2}=1} \int_{]-1,1[^3} (T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta)\varphi_{MN}$$
$$\leq \|F(\psi_\delta,\lambda_\delta)\|_{(X_0^2)^* \times \mathbb{R}} + c(N^{-s}\|V\psi_\delta\|_{L^2(]-1,1[^2;H^s]-1,1[)} + M^{-\sigma}\|V\psi_\delta\|_{H^\sigma(]-1,1[^2;L^2]-1,1[)}) \tag{41}$$

The term $(T_{R,r,z}\psi_\delta + \pi_M \otimes \mathcal{J}_N(V\psi_\delta) - \lambda_\delta\psi_\delta)$ being in X_{MN} hence in X_0^2, we choose it as φ after proper normalization in the above supremum; recalling for $b = 2$, $a = 0$ the inverse inequality that is true for elements of X_{MN} ([1] p.256)

$$\forall\ b \geq a \geq 0,\ \forall \psi_{MN} \in X_{M,N} \quad \|\psi_{MN}\|_{X_0^b} \ \leq\ C\ \max(M,N)^{b-a}\|\psi_{MN}\|_{X_0^a}. \tag{42}$$

we obtain trivially from (36) and the first inequality in (23) the second estimate of the theorem. □

Remark 3.4. The estimator can be explicitely computed since it involves L^2 norms of discrete functions ; moreover its computation can be done in an fast manner as it will be seen in section 5, remark 5.1.

[4]When the functions involved are regular enough, the second term in the right hand side of (38) can be considered small enough to be neglected (see also [6,7]); this is the case for instance in formula (38) with $N^2 \simeq M^2 \simeq E_{MAX}$ as soon as regularity allows using $\sigma, s > 2$ (close enough to the solution).

4. FURTHER RESULTS

4.1. X_0^1 estimate

Although the L^2 norm seems the most natural when studying the convergence of the eigenfunctions, there are some remarkable situations (see below) where another norm, here the X_0^1 norm, is required to measure the error. Our approach lets us the freedom to analyze this cases as well, obtaining thus an estimator for the error expressed as $\|\psi_0 - \psi_\delta\|_{X_0^1} + |\lambda_0 - \lambda_\delta|$.

Indeed, denote by $H_*^s = D(A^{s/2})$ the domain in $L^2(]-1,1[)$ of the $s/2$-th power of the operator $A = \partial_z(1-z^2)\partial_z$ endowed with canonical norm; then, for any $\alpha > 0$ there exists some constant $c_\alpha > 0$ such that the following interpolation property be valid (use (6) and (5.9) p.256, like in Thm. 13.4, p.303 [1]):

$$\forall v \in H_*^\alpha(]-1,1[), \quad \|v - \mathcal{J}_N v\|_{H_*^1} \leq c_\alpha N^{1-\alpha}\|v\|_{H_*^\alpha}. \tag{43}$$

The result reads:

Theorem 4.1. *Let $\sigma \geq 0$, $s > \frac{1}{2}$ be such that $V\psi_\delta \in L^2(]-1,1[^2; H^s(]-1,1[)) \cap H^\sigma(]-1,1[^2; L^2(]-1,1[))$. There exists constants $c, C > 0$ and $c(V) > 0$ such that*

$$\begin{aligned} \|\psi_0 - \psi_\delta\|_{X_0^1} + |\lambda_0 - \lambda_\delta| \leq \frac{c(V)\max(M,N)}{E_{MAX}} \|(Id - \pi_{\mathcal{E}_N})(\partial_z(1-z^2)\partial_z\psi_\delta)\|_{L^2(]-1,1[^3)} \\ + \quad c(M^{1-\sigma}\|V\psi_\delta\|_{H^\sigma(]-1,1[^2;L^2]-1,1[)} + N^{1-s}\|V\psi_\delta\|_{L^2(]-1,1[^2;H_*^s)}) \end{aligned} \tag{44}$$

and

$$\begin{aligned} \frac{C}{\max(M,N)} \quad \|(Id - \pi_{\mathcal{E}_\delta})(\partial_z(1-z^2)\partial_z\psi_\delta)\|_{L^2(]-1,1[^3)} \leq \left(\|\psi_0 - \psi_\delta\|_{X_0^1} + |\lambda_0 - \lambda_\delta|\right) \\ + c(M^{1-\sigma}\|V\psi_\delta\|_{H^\sigma(]-1,1[^2;L^2]-1,1[)} + N^{1-s}\|V\psi_\delta\|_{L^2(]-1,1[^2;H_*^s)}). \end{aligned} \tag{45}$$

Proof. We follow the same lines of proof as in Theorem 3.3 making use of the abstract result for $Z = X_0^1 \times \mathbb{R}$, $Y = X_0^{1\,*} \times \mathbb{R}$. For the second part we are making use of (42) for $b = 1$, $a = 0$. □

Remark 4.2. From the a priori estimate (and the common sense) it is natural to choose $N^2 \simeq M^2 \simeq E_{MAX}$. Theorem 2 gives an *optimal* a posteriori estimate to judge on the adiabaticity of the variable.

4.2. Separate estimates for eigenvalues and eigenfunctions

The estimators obtained before do not provide separated indications on the convergence of the eigenvalues and the eigenfunctions alone; moreover they cannot account for well-known phenomena like super-convergence of eigenvalues when compared with the H^1 convergence of eigenfunctions.

It seems therefore legitimate to us to search for such tailored estimators. The framework is the following: suppose as can be hinted from Thm. 3.3 and 4.1 that our discretization of the problem allows for a better convergence of eigenfunctions in the L^2 norm when compared with H^1 norm[5]. Then we recall in what follows that the error for the eigenvalues behaves (asymptotically) like the square of the H^1 error for eigenfunctions. We use this to obtain an estimator for the error in the eigenvalues alone; it is that estimator that we illustrate next in numerical experiments.

[5] this is generally true for most approximation of nuclear structure computations while this may however not be the case for electronic structure when incomplete basis are used

Let $(\psi_\delta, \lambda_\delta)$ be an approximation of the eigenmode (ψ_0, λ_0) (ψ_δ and ψ_0 are L^2-normalized to 1). Then we can write:

$$\begin{aligned}\lambda_\delta - \lambda_0 &= (H\psi_\delta, \psi_\delta) - (H\psi_0, \psi_0) = (H(\psi_\delta - \psi_0), (\psi_\delta - \psi_0)) + 2(H\psi_0, (\psi_\delta - \psi_0)) \\ &= (H(\psi_\delta - \psi_0), (\psi_\delta - \psi_0)) + 2\lambda_0(\psi_0, \psi_\delta - \psi_0) \end{aligned} \tag{46}$$

Using the normalization of ψ_δ and ψ_0 we see that $2\lambda_0(\psi_0, \psi_\delta - \psi_0)$ equals $-\lambda_0 \int (\psi_\delta - \psi_0)^2$. By the definition of the space X_0^1 we obtain:

$$\lambda_\delta - \lambda_0 = \|\psi_\delta - \psi_0\|^2_{X_0^1} + \int (V - \lambda_0)(\psi_\delta - \psi_0)^2. \tag{47}$$

In what follows we need the following

HYPOTHESIS [A]: the $L^2 = X_0^0$ norm of the error for eigenfunctions converges faster than the X_0^1 norm. Assuming hypothesis [A] holds, then there exists c_1 and c_2 (close to 1) not depending on the parameter δ such that for δ small enough

$$c_1\|\psi_\delta - \psi_0\|^2_{X_0^1} \leq |\lambda_\delta - \lambda_0| \leq c_2\|\psi_\delta - \psi_0\|^2_{X_0^1}. \tag{48}$$

Let us now assume (to simplify) that $M^2 \simeq N^2 \simeq E_{MAX}$. From the discussion above we know that in the term $\|\psi_0 - \psi_\delta\|_{X_0^1} + |\lambda_0 - \lambda_\delta|$ the leading part is the first one (the second one behaving like the square of the first) so we obtain by Theorem 3 a new error estimator $\frac{c(V)}{\sqrt{E_{MAX}}}\|(Id - \pi_{\mathcal{E}_N})(\partial_z(1-z^2)\partial_z\psi_\delta)\|_{L^2(]-1,1[^3)}$ for $\|\psi_0 - \psi_\delta\|_{X_0^1}$ and of course, its square is an estimator for $|\lambda_0 - \lambda_\delta|$. We have therefore proven:

Corollary 4.3. *Under the hypothesis [A] and for the $M^2 \simeq N^2 \simeq E_{MAX}$ there exists two constants $c > 0$, $C > 0$ and $c(V) > 0$ such that*

$$\begin{aligned}\max\{\|\psi_0 - \psi_\delta\|_{X_0^1}, \sqrt{|\lambda_0 - \lambda_\delta|}\} \leq \tfrac{c(V)}{\sqrt{E_{MAX}}}\|(Id - \pi_{\mathcal{E}_N})(\partial_z(1-z^2)\partial_z\psi_\delta)\|_{L^2(]-1,1[^3)} + \\ c(M^{1-\sigma}\|V\psi_\delta\|_{H^\sigma(]-1,1[^2;L^2]-1,1[)} + N^{1-s}\|V\psi_\delta\|_{L^2(]-1,1[^2;H^s_*)})\end{aligned} \tag{49}$$

and

$$\begin{aligned}\tfrac{C}{\sqrt{E_{MAX}}}\|(Id - \pi_{\mathcal{E}_N})(\partial_z(1-z^2)\partial_z\psi_\delta)\|_{L^2(]-1,1[^3)} \leq \min\{\|\psi_0 - \psi_\delta\|_{X_0^1}, \sqrt{|\lambda_0 - \lambda_\delta|}\} \\ c(M^{1-\sigma}\|V\psi_\delta\|_{H^\sigma(]-1,1[^2;L^2]-1,1[)} + N^{1-s}\|V\psi_\delta\|_{L^2(]-1,1[^2;H^s_*)}).\end{aligned} \tag{50}$$

5. RESULTS AND CONCLUSIONS

In order to prove the efficiency of our error estimator we have considered some numerical experiments. The system of interest is the water molecule : the hydrogen atoms are located in A and C and the oxygene in B ; we are interested in finding the fundamental and the first 8 excited states.

Although the theory described so far was derived (for the sake of simplicity) only for some constant multiplication function $f(R, r) \equiv 1$ in the kinetic operator in the adiabatic variable $f(R, r)\partial_z(1 - z^2)\partial_z$ (see above) it can be easily extended in order to accommodate the most appropriate modelisation

$$f(R, r) = \frac{\mu_1}{R^2} + \frac{\mu_2}{r^2}, \quad r \in]r_{min}, r_{max}[, \ R \in]R_{min}, R_{max}[, \ r_{min}, R_{min} > 0 \tag{51}$$

where μ_1 and μ_2 are structural constants that depend on the system under consideration.

Remark 5.1. The explicit computation of the contribution

$$\|(Id - \pi_{\mathcal{E}_\delta})(\partial_z(1-z^2)\partial_z\psi_\delta)\|_{L^2(]-1,1[^3)} \tag{52}$$

can be done in a "fast" (i.e. less operation than for the evaluation of ψ_δ) manner as follows; let us note

$$\partial_z(1-z^2)\partial_z h_i = \sum_{j=1}^{M} \gamma_i^j \; h_j \quad for\ all\ i = 1, ..., N \tag{53}$$

$$\Phi_{p,q,i}(R,r) = \sum_{r,s=1}^{M} \alpha_{pqi}^{rs} \varphi_{r,s,0}(R,r). \tag{54}$$

Then we consider the following change of basis

$$\Phi_{p,q,i}(R,r) = \sum_{p',q'=1}^{M} \eta_{pqi}^{p'q'j} \Phi_{p',q',j}(R,r) \quad for\ all\ i,j = 1,..,N\ p,q = 1,...,M, \tag{55}$$

where, by the orthonormality of all basis involved (i.e. $(\Phi_{p,q,i})_{p,q=1}^M$ for every i and $(\varphi_{r,s,0})_{r,s=1}^M$), we have:

$$\eta_{pqi}^{p'q'j} = \sum_{r,s=1}^{M} \alpha_{pqi}^{rs} \alpha_{p'q'j}^{rs}, \tag{56}$$

hence

$$\Phi_{p,q,i}(R,r)(\partial_z(1-z^2)\partial_z h_i)(z) = \sum_{p',q'=1}^{M} \sum_{j=1}^{N} \gamma_i^j \eta_{pqi}^{p'q'j} \Phi_{p',q',j}(R,r) h_j(z). \tag{57}$$

From the formula $\psi_\delta = \sum_{p,q,i} \check{\psi}_{pqi} \Phi_{p,q,i} h_i$ given by the solution of the reduced problem we notice

$$A\psi_\delta := [\partial_z(1-z^2)\partial_z]\psi_\delta = \sum_{p',q',j} \left[\sum_{p,q,i} \gamma_i^j \eta_{pqi}^{p'q'j} \check{\psi}_{pqi} \right] \Phi_{p',q',j}(R,r) h_j(z). \tag{58}$$

This gives us the value of the coefficients $A\psi_\delta$ in the orthonormal basis $\Phi_{p',q',j}(R,r)h_j(z)$. By tensorization the computation (58) can be done in $c\max(M,N)^5$ operations, less than the number of operations required by the computation of ψ_δ (for instance, the diagonalization of $2D$ hamiltonians is of higher complexity) [8, 9, 11].

Indeed, our goal is to compute for $\{(p',q',j); |\Lambda_{p',q',j}| \geq (1+\epsilon)E_{MAX}\}$ the term:

$$\beta_{p',q',j} = \sum_{r,s,p,q,i} \check{\psi}_{pqi} \alpha_{pqi}^{rs} \alpha_{p'q'j}^{rs} \gamma_i^j, \quad p',q' = 1,...,M, \quad j = 1,...,N. \tag{59}$$

It is easy to check that summing first for p and q we obtain in $c\max(M,N)^5$ operations some coefficients

$$\theta_{rs}^i = \sum_{p,q} \check{\psi}_{pqi} \alpha_{pqi}^{rs}. \tag{60}$$

Next we sum up for the "i" index and note $\chi_{rs}^j = \sum_i \theta_{rs}^i \gamma_i^j$. Our quantity is:

$$\sum_{rs} \chi_{rs}^j \alpha_{p'q'j}^{rs} \tag{61}$$

and it is clear now that we can compute it for all values of $(p'q'j)$ needed in $c\max(M,N)^5$ operations. The L^2 norm of $\|(Id-\pi_{\mathcal{E}_N})(\partial_z(1-z^2)\partial_z\psi_\delta)\|_{L^2(]-1,1[^3)}$ is obtained by summing up the square of $\beta_{p',q',j}$ for all indices $\{(p',q',j);|\Lambda_{p',q',j}|<(1+\epsilon)E_{MAX}\}$. Note that only these coefficients have to be computed in (61) and that in (60) the $\check{\psi}_{pqi}$ all vanish for indices $\{(p,q,i);|\Lambda_{p,q,i}|\geq(1+\epsilon)E_{MAX}\}$. Taking this into account leads to a further reduction in CPU time [9].

The results are displayed in the figures 1-10. We choose discretization parameters M and N such that $N^2\simeq M^2\simeq E_{MAX}$. We are plotting the effectivity indexes, i.e. the quotient "true error over estimated error". Of course the ideal case would be "effectivity index = constant", but this never happens for discretization of non linear problems. Due to the intricate nature of the eigenvalue problem we cannot expect that. What we do expect is that our estimator be robust and rather insensitive to different discretization parameters (here E_{MAX}). The quotient " true error over estimated error" was computed with energy expressed in atomic units (Hartree, E_h): $1E_h=219474.63cm^{-1}$; the true error was computed with respect to a solution obtained with a very fine discretization.

The relative error was measured with respect to the first excitation of the system, that is the difference between the first and the second eigenvalue, and was found to be in the range $3\%-0.001\%$, which is typical for this kind of computations. This choice for measuring the relative error is suggested by the fact that the value of zero for the potential (or energy) is defined up to an additive constant, thus only relative variations are relevant. Other procedures for measuring the relative error on the i-th eigenvalue can be proposed (one may consider as basis for computations the difference between the "i"-th and "$i-1$"-th eigenvalues), the present choice was retained for the sake of uniformity. Finally, let us mention that in practice chemists are satisfied when the energies are known up to several cm^{-1} units, $1cm^{-1}=.455\cdot10^{-5}E_h$. The computations presented also comply with this requirement, as e.g. for the first eigenvalue, the error decreases from $24cm^{-1}$ to less than $1cm^{-1}$.

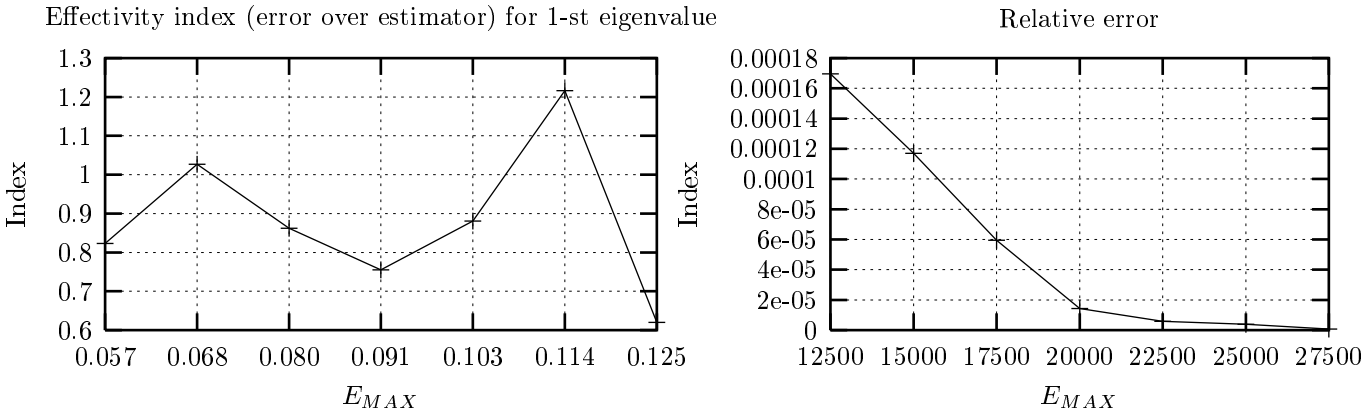


FIGURE 1. First eigenvalue

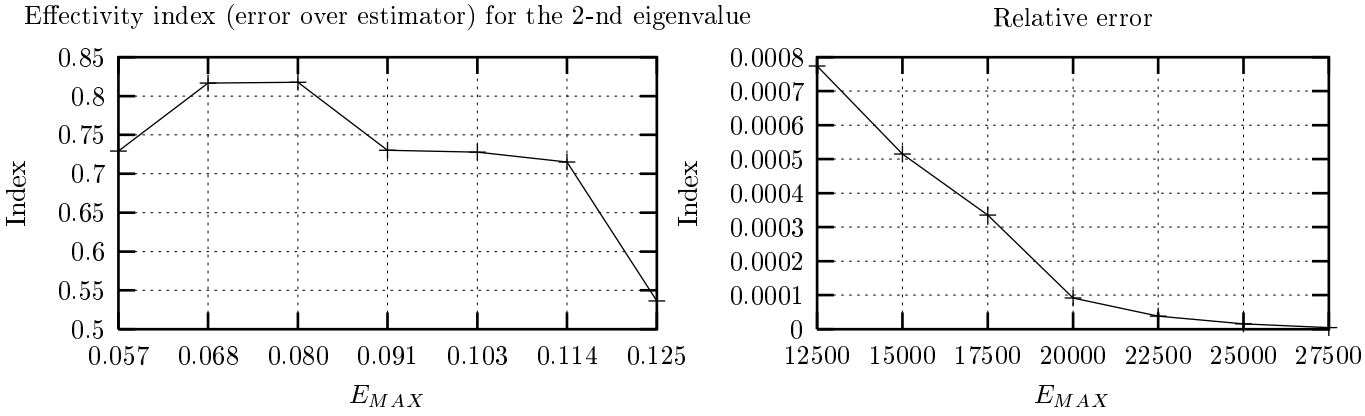


FIGURE 2. Second eigenvalue

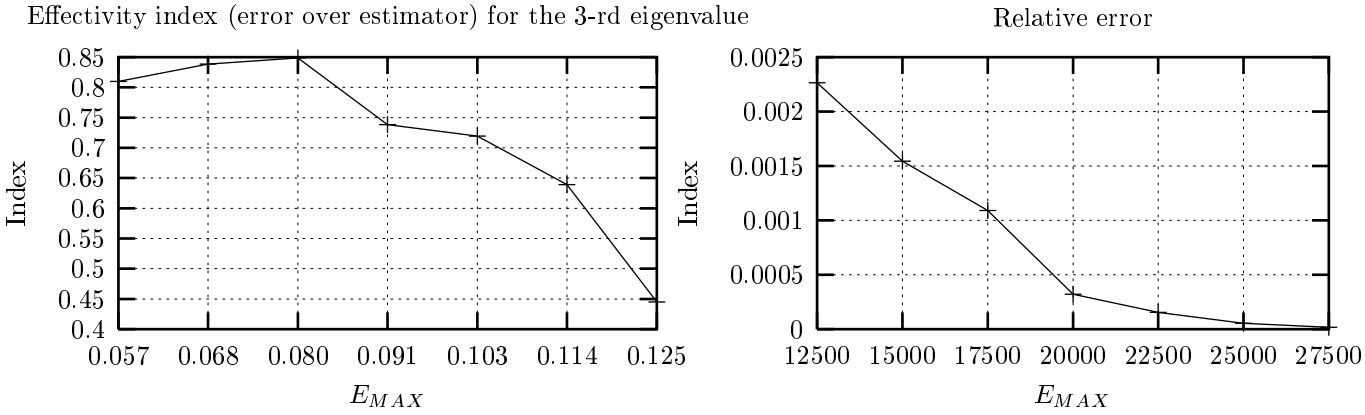


FIGURE 3. Third eigenvalue

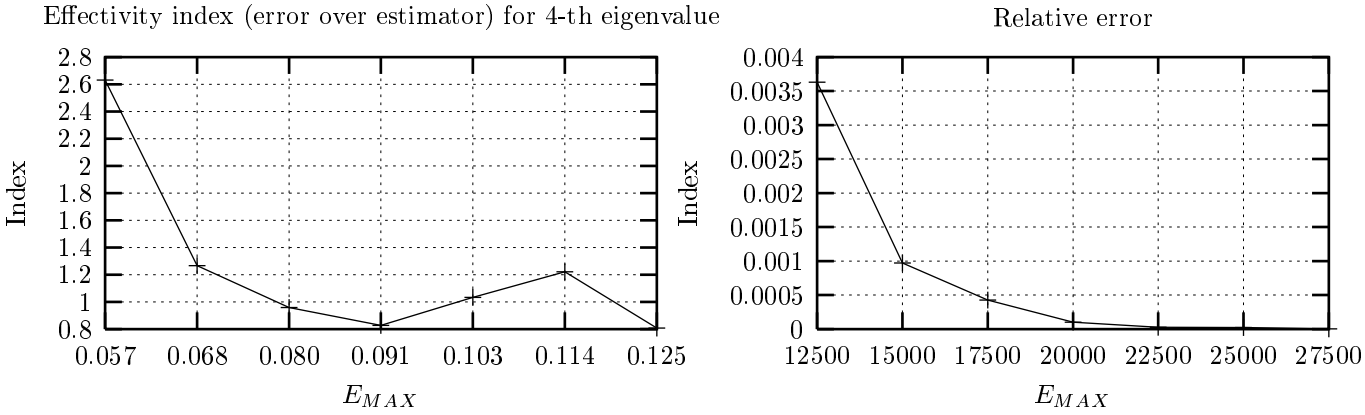


FIGURE 4. Fourth eigenvalue

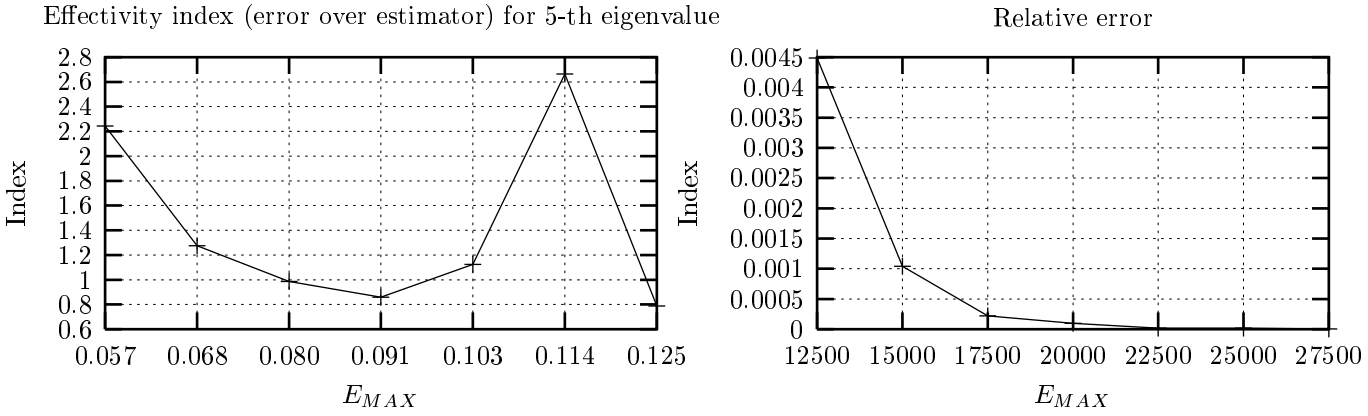


FIGURE 5. Fifth eigenvalue

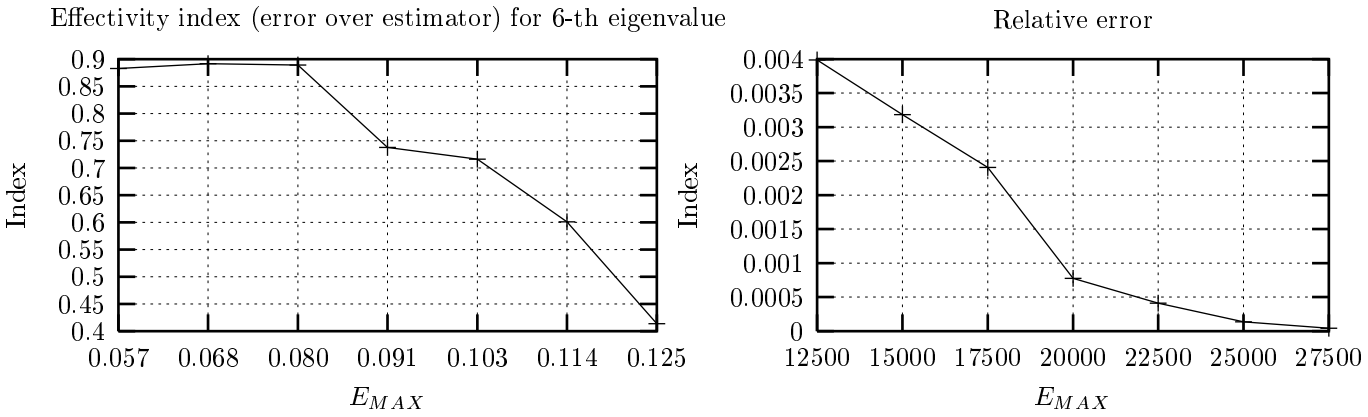


FIGURE 6. Sixth eigenvalue

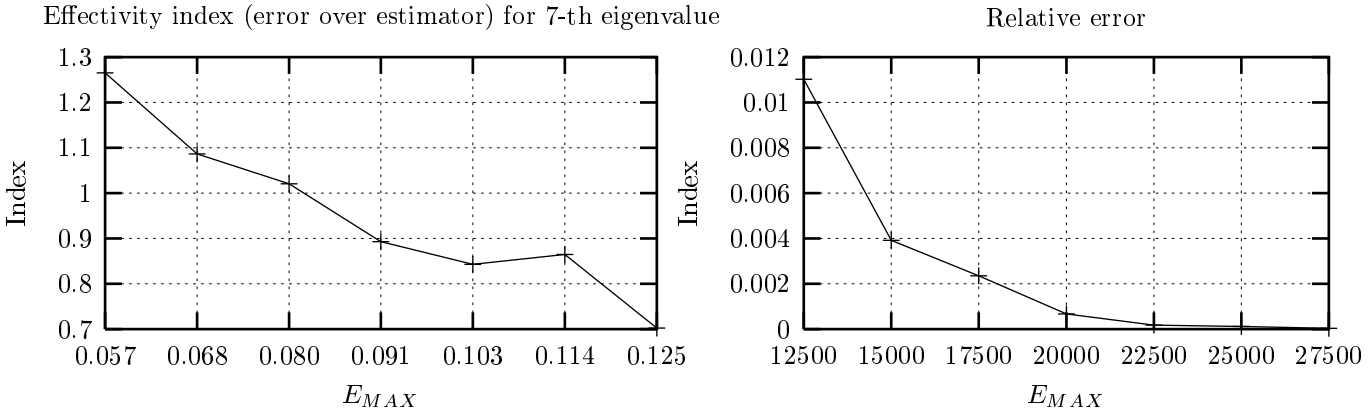


FIGURE 7. Seventh eigenvalue

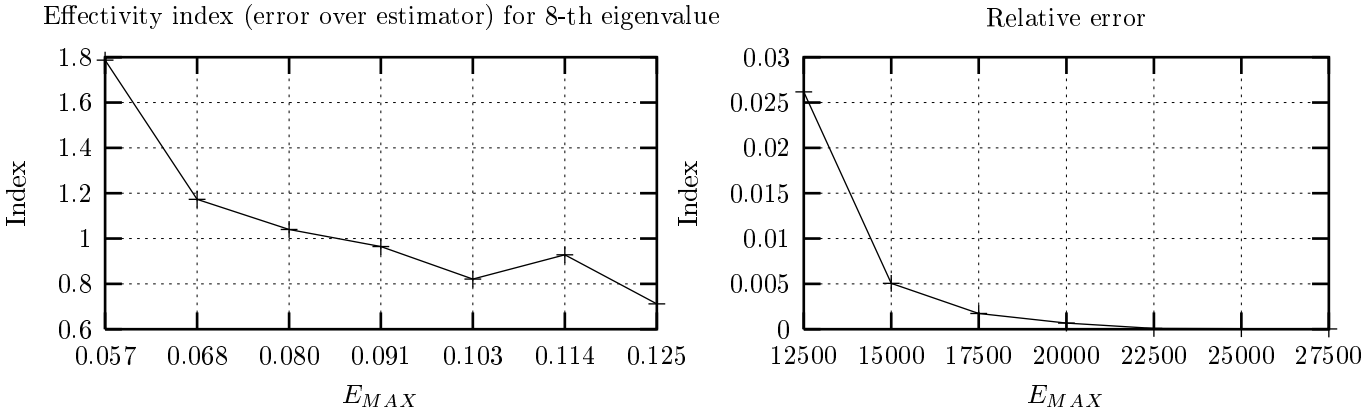


FIGURE 8. Eighth eigenvalue

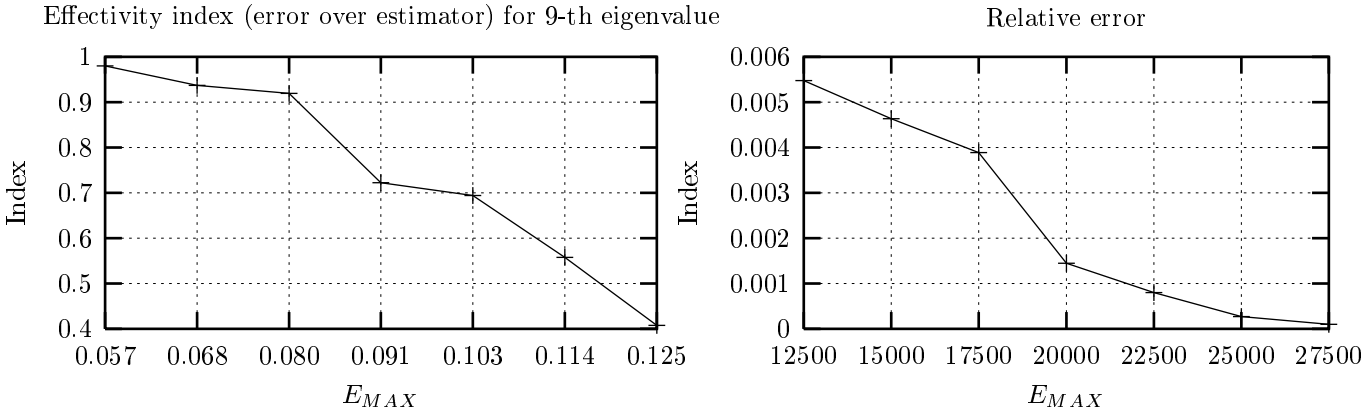


FIGURE 9. Ninth eigenvalue

We would also want that the estimator **quantitatively** describe the order of magnitude of the error. For the effectivity index this condition requires that the ratio between the extremal values of the effectivity index be no larger than 10. As we can see from the results displayed, all our indexes fulfill this requirement. In fact in our case this ratio is roughly 2 (except for eigenmodes 4,5 and 8 where it is closer to 3).

The index involves the norm of the operator $DF(\psi_0, \lambda_0)$ and its inverse mapping; it is surprising to notice that the range for the effectivity indexes is basically the same, **even for different eigenmodes**, which was not predicted by the theory. It seems that the various norms $DF(\psi_0, \lambda_0)$ vary slowly when calculated in different eigenmodes. The variation of the effectivity index for two values of $E_{MAX} = 17500cm^{-1}(0.0797E_h)$ and $E_{MAX} = 27500cm^{-1}(0.1253E_h)$ is plotted in figure 10 for all the nine eigenmodes.

Let us finally mention that the form of the estimator is not easy to find intuitively; other empirical combinations of, for instance, powers of E_{MAX} and the L^2 "residual" norm involved display divergence for the effectivity index.

Remark 5.2. It is of course natural to test the estimator on other types of molecules and also on other choices of adiabatic variables that might be less performant. This will allow to investigate the qualiy of the part of the estimator related to adiabaticity. This work is in proposition but requires heavy investment of our colleagues that have to change significantly their code.

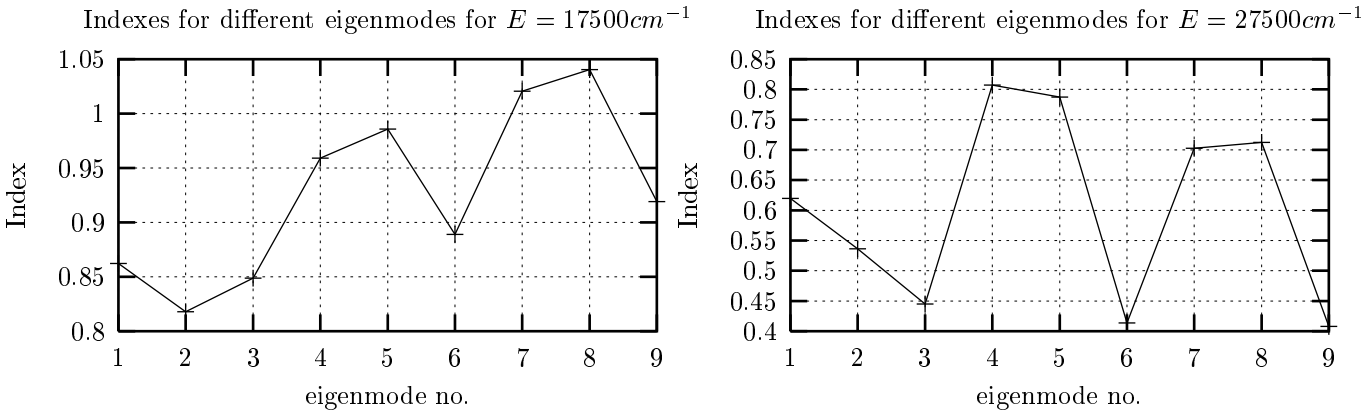


FIGURE 10. Efectivity indexes for different eigenmodes and cut-off E_{MAX} values.

The authors wish to thank C. Leforestier from *Laboratoire Structure et Dynamique des Systèmes Moléculaires et Solides, Université de Montpellier 2* for providing his basic code to use with our estimators and for the discussions on the subject.

REFERENCES

[1] Christine Bernardi and Yvon Maday "Spectral methods" in "Handbook of numerical analysis", Ph. G. Ciarlet and J.L. Lions (eds.), North-Holland, vol. V, Part 2, 1997
[2] Christine Bernardi and Yvon Maday "Approximations spectrales de problèmes aux limites elliptiques" Paris; Berlin ; NewYork NY : Springer, 1992.
[3] R.Dutray and J.L.Lions "Analyse mathématique et calcul numérique pour les sciences et les techniques" tome 5 , MASSON , CEA, 1984
[4] J.L.Lions and E.Magenes "Problèmes aux limites non-homogènes et applications", DUNOD, Paris 1968
[5] M. Azaiez, M. Dauge and Y. Maday "Méthodes spectrales et les éléments spectraux" Institut de Recherche Mathématique de Rennes
[6] R.Verfürth "A Review of A Posteriori Error Estimates and Adaptative Mesh-Refinement Techniques", Wiley-Teubner 1997
[7] R.Verfürth "A Posteriori Error Estimates For Non-Linear Problems. Finite Element Discretisations of Elliptic Equations" Math. of Comp. **62**, 206(1994), pp 445-475
[8] R.Friesner, J.Bentley, M.Menou and C.Leforestier, "Adiabatic pseudospectral methods for multimensional vibrational potentials " J.Chem.Phys., **99** , 324(1993).
[9] C.Leforestier "Grid representation of rotating triatomics", J.Chem.Phys., **94** , 6388(1991).
[10] K.Yamashita, K.Mokoruma and C.Leforestier "Theoretical study of the highly vibrationally excited states of FHF^-: Ab initio potential energy surface and hyperspherical formulation", J.Chem.Phys., **99** , 8848(1993).
[11] J.Antihainen, R.Friesner and C.Leforestier "Adiabatic pseudospectral calculation of the vibrational states of the four atom molecules: Application to hydrogen peroxide" J.Chem.Phys., **102** , 1270(1995).
[12] D. Kosloff and R.Kosloff "Fourier Method for the Time Dependent Schrödinger Equation as a Tool in Molecular Dynamics" J. Comp. Phys. **vol 52**, 35 (1983)
[13] R.Kosloff "Time-Dependent Quantum-Mecanical Methods for Molecular Dynamics" J. Chem. Phys., **92**, 2087(1988).
[14] I. Babuška and C. Schwab "A posteriori error estimation for hierarchic models of elliptic boundary value problems on thin domains" SIAM J. Numer. Anal. Vol **33** (1996), No.1, pp 241-246
[15] C.Canuto, M.Y.Hussaini, A.Quarteroni and T.A.Zang "Spectral Methods in Fluid Dynamics" (Springer, Berlin, 1987)

APPENDIX A.

Remark A.1. By the definition of the spaces X_0^s the operator $T_{R,r,z}$ is an isometry between X_0^2 and $X_0^0 = L^2(]-1,1[^3)$; for any $g \in L^2(]-1,1[^3)$ the equation

$$T_{R,r,z} f = g \tag{62}$$

has therefore an unique solution $f \in X_0^2$; moreover the mapping that to g associates the solution f of (62) is a compact mapping from $L^2(]-1,1[^3)$ into $L^2(]-1,1[^3)$ (because of the embedding $H_0^1(]-1,1[^3) \subset L^2(]-1,1[^3)$ which is compact). By the Lax-Milgram lemma, as soon as $V \in L^\infty$, $\alpha \geq \|V\|_{L^\infty}$ the same properties remain true for the equation

$$(H + \alpha Id)f = T_{R,r,z}f + Vf + \alpha f = g \tag{63}$$

Is is essential for the a posteriori analysis of the (2)-(3) to study the properties of the differential $DF(\psi_0, \lambda_0)$ of F in the solution (ψ_0, λ_0) of (2)-(3) ; more precisely, it will be proven that if λ_0 is a simple eigenmode (i.e. of multiplicity 1) of H and $V \in L^\infty$ then $DF(\psi_0, \lambda_0)$ is an isomorphism from $L^2(]-1,1[^3) \times \mathbb{R}$ into $(X_0^2)^* \times \mathbb{R}$.

A straightforward computation gives the following formula for $DF(\psi_0, \lambda_0)$:

$$\begin{aligned} &< DF(\psi_0,\lambda_0)(\psi,\lambda),(\varphi,\mu) >= \int_{]-1,1[^3} H\varphi\psi - \lambda_0\psi\varphi - \lambda\psi_0\varphi + 2\mu \int_{]-1,1[^3} \psi_0\psi \\ &= \int_{]-1,1[^3} (H\varphi - \lambda_0\varphi + 2\mu\psi_0)\cdot\psi - \lambda \int_{]-1,1[^3} \psi_0\varphi =< (\psi,\lambda), DF(\psi_0,\lambda_0)^*(\varphi,\mu) > \end{aligned} \tag{64}$$

where $DF(\psi_0, \lambda_0)^*$ is the adjoint of $DF(\psi_0, \lambda_0)$. To prove the bijectivity of $DF(\psi_0, \lambda_0)$ we check that $DF(\psi_0, \lambda_0)^*$ is bijective. This is equivalent to prove that for any $\beta \in \mathbb{R}$ and $w \in L^2(]-1,1[^3)$ there exists an (unique) couple (φ, μ) such that :

$$H\varphi + 2\mu\psi_0 - \lambda_0\varphi = w \tag{65}$$

$$\int_{]-1,1[^3} \psi_0\varphi = \beta \tag{66}$$

The equation (65) can be written $(H - \lambda_0)\varphi = w - 2\mu\psi_0$. If we suppose that λ_0 is a simple eigenvalue, then, by the remark A.1 and by the Fredholm alternative[6] (65) has a solution iff $w - 2\mu\psi_0 \perp \psi_0$ that is $\mu = \frac{<w,\psi_0>}{2}$; in this case the set of solutions is $\{\varphi_0 + \gamma\psi_0; \gamma \in \mathbb{R}\}$ where φ_0 is a particular **fixed** solution. By (66) we compute $\gamma = \beta - < \psi_0, \varphi_0 >$ and so we have found a couple $(\varphi = \varphi_0 + \gamma\psi_0, \mu)$ that satisfy (65) and (66). It is therefore natural to suppose that $V \in L^\infty$ and that all eigenvalues under study are simples.

A.1. **Proof of lemma 3.2.**

Let us remind that all element φ_{MN} in $X_{M,N}$ can be written as

$$\varphi_{MN}(R,r,z) = \sum_{p,q,i=1}^{N+1} c_{p,q,i}\Phi_{p,q,i}(R,r)h_i(z), \tag{67}$$

with

$$c_{p,q,i} = \int_R \int_r \varphi_{MN}(R,r,\zeta_i)\Phi_{p,q,i}(R,r)dRdr. \tag{68}$$

By the definition of eigenmodes $\Phi_{p,q,i}$ we have also (by use of integration by parts)

$$\begin{aligned} c_{p,q,i} &= \int_R \int_r \varphi_{MN}(R,r,\zeta_i)\tfrac{1}{\Lambda_{p,q,i}}\big((-\partial_{RR} - \partial_{rr} - V(R,r,\zeta_i))\Phi_{p,q,i}\big)dRdr \\ &= \tfrac{1}{\Lambda_{p,q,i}} \int_R \int_r \big((-\partial_{RR} - \partial_{rr} - V(R,r,\zeta_i))\varphi_{MN}\big)(R,r,\zeta_i)\Phi_{p,q,i}(R,r)dRdr. \end{aligned} \tag{69}$$

[6] we write $H - \lambda_0 = (H + \alpha Id) - (\alpha + \lambda_0)Id$ and we use, for α large enough, the Fredholm alternative ([3] p. 39) for the compact operator $(H + \alpha Id)^{-1}$ and the eigenvalue $\frac{1}{\alpha+\lambda_0} \neq 0$.

Moreover by the definition of the projector we have

$$\big(\varphi_{MN} - \pi_{\mathcal{E}_\delta}(\varphi_{MN})\big)(R,r,z) = \sum_{(p,q,i);|\Lambda_{p,q,i}|>(1+\epsilon)E_{MAX}} c_{p,q,i}\Phi_{p,q,i}(R,r)h_i(z), \tag{70}$$

so that

$$\|\varphi_{MN} - \pi_{\mathcal{E}_\delta}(\varphi_{MN})\|^2_{L^2} \leq \sum_{(p,q,i);|\Lambda_{p,q,i}|>(1+\epsilon)E_{MAX}} (c_{p,q,i})^2\rho_i$$
$$\leq \sum_{(p,q,i);|\Lambda_{p,q,i}|>(1+\epsilon)E_{MAX}} \frac{1}{\Lambda^2_{p,q,i}}\Big(\int_R\int_r\big((-\partial_{RR}-\partial_{rr}-V(R,r,\zeta_i))\varphi_{MN}(R,r,\zeta_i)\big)\Phi_{p,q,i}(R,r)dRdr\Big)^2\rho_i. \tag{71}$$

By the orthogonality of $\Phi_{p,q,i}$ we have

$$\|\varphi_{MN} - \pi_{\mathcal{E}_\delta}(\varphi_{MN})\|^2_{L^2} \leq (\tfrac{1}{(1+\epsilon)E_{MAX}})^2\Big(\|(-\partial_{RR}-\partial_{rr})\varphi_{MN}\|^2_{L^2(]-1,1[^3)}$$
$$+\|\textstyle\sum_i (V(.,.,\zeta_i)\varphi_{MN}(.,.,\zeta_i))^2\rho_i\|^2_{L^2(]-1,1[^2)}\Big) \tag{72}$$

which concludes the proof of the first part of the lemma.

To prove (31) note that it is trivially true for $b=0$ and by the argument above for $b=2$; using once more in (69) the definition of eigenmodes $\Phi_{p,q,i}$ and after one supplementary integration by parts we obtain

$$c_{p,q,i} = \tfrac{1}{\Lambda^2_{p,q,i}}\int_R\int_r\big((-\partial_{RR}-\partial_{rr}-V(R,r,\zeta_i))^2\varphi_{MN}\big)(R,r,\zeta_i)\Phi_{p,q,i}(R,r)dRdr \tag{73}$$

so, by the same line of reasoning as above, upper bound (31) is proved for $b=4$; by continuing the procedure for all even values of b and using classical interpolation arguments the conclusion will follows.

A.2. **Proof of lemma 2.4.**

Let $\Pi_{M,N}$ be the projector to $X_{M,N}$ associated with $T_{R,r,z}$ that is for all $v \in X^1_0$, $\Pi_{M,N}v$ is the element of $X_{M,N}$ that verifies

$$\forall\ u \in X_{M,N}\ :\ \int_{]-1,1[^3} T_{R,r,z}(v-\Pi_{M,N}v)u = 0. \tag{74}$$

Note that $\Pi_\delta\Pi_{M,N} = \Pi_\delta$. It is classical[7] to see that $\Pi_{M,N}$ has optimal approximation properties, that is, for any $b \geq 1 \geq a \geq 0$ there exists a constant c independent of M,N such that

$$\|v-\Pi_{M,N}v\|_{X^a_0} \leq c\Big(\frac{1}{\max(M,N)}\Big)^{b-a}\|v\|_{X^b_0}. \tag{75}$$

Write then :

$$\|v-\Pi_\delta v\|_{X^a_0} \leq \|v-\Pi_{M,N}v\|_{X^a_0} + \|\Pi_{M,N}v - \Pi_\delta\Pi_{M,N}v\|_{X^a_0}. \tag{76}$$

By (75) the first term in (76) is optimal, so only the second term remains to be (optimally) upper bounded. Denote $f = \Pi_{M,N}v$; recall the minimization property of Π_δ :

$$\Pi_\delta v = argmin\{\|v-u\|_{X^1_0}; u \in \mathcal{E}_\delta\}$$

[7] use for instance the reasoning in [1] p. 262

and write, for $a = 1$:

$$\|f - \Pi_\delta f\|_{X_0^1} \leq \|f - \pi_{\mathcal{E}_\delta} f\|_{X_0^1} \leq C \max(M,N) \|f - \pi_{\mathcal{E}_\delta} f\|_{L^2} \leq C \max(M,N) \Big(\frac{1}{\sqrt{E_{MAX}}} \Big)^b \|f\|_{X_0^b}, \qquad (77)$$

which ends the proof of the lemma for $a = 1$; the values of a in $[0,1[$ are treated by the duality technique of Aubin and Nitsche (see for instance [1] p. 274-275).

Chapitre 3

Étude des équations de Hartree-Fock (A posteriori numerical analysis for the Hartree-Fock equations and quadratically convergent methods)

Cette étude [27] a été effectuée en collaboration avec Yvon Maday et porte sur l'analyse a posteriori des équations de Hartree-Fock.

Les methodes a posteriori sont employées à la recherche des bornes sur des résultats en sortie tels que l'énergie de Hartree-Fock et conduisent dans un premier temps à l'identification des procédés constructifs pour le calcul d'un intervalle de confiance pour l'énergie de Hartree-Fock.

Le cadre particulier du problème variationnel est ensuite mis en valeur par des résultats qui portent sur l'accéleration de la convergence des algorithmes SCF utilisés. Des exemples numériques sont portés à l'appui des résultats théoriques.

A posteriori numerical analysis for the Hartree-Fock equations and quadratically convergent methods

Abstract

This paper presents an a posteriori error analysis of the discretization methods used in computational quantum chemistry on the Hartree-Fock equations. Bounds on the energy are obtained from any discrete approximation strategy of the solution and the estimator proposed is shown to possess further algorithmic virtues.

1 Introduction

The purpose of this paper is to present an a posteriori error analysis for the approximation of the Hartree-Fock equations. This analysis is designed to quantitatively asses the performance of an approximation strategy of a solution of the Hartree-Fock equations obtained by prior computation. In agreement with the general paradigm of the a posteriori analysis of [11, 13, 14, 15], a trust interval for an output such as the Hartree-Fock energy starting from the approximated solution at hand is proposed. In addition we will show that in some cases the a posteriori method may also be seen as an accelerator of the convergence of the primary algorithm used to compute the solution.

The time independent Schrödinger equation that models the behavior of a quantum molecular system deals with state functions $\psi(\underline{x})$, where $\underline{x}$ denotes the position of the particles (nuclei and electrons) hence is a variable that lives in $\mathbb{R}^{3K}$ where K is the number of particles[1]. This system is far too large to be directly tractable by numerical simulations for molecules larger than the hydrogen atom. The quantum chemists have thus introduced a series of simplified models. One of them (the Born Oppenheimer approximation) allows to separate the electron and the nuclei so as to consider first

[1] we will consider non relativistic models without spin variables

a system in which only the N electrons of the molecule move (thus are the only N variables of the state function) and the nuclei are fixed in $\overline{x}_j$ (and appear as parameters). For each configuration $(\overline{x}_1, ..., \overline{x}_m)$ of the m nuclei a complex electronic wavefunction $\Phi(x_1, ..., x_N) \in \mathbb{C},\ x_i \in \mathbb{R}^3,\ i = 1, ..., N$ is sought after that minimizes the energy of the system. This first simplification is nevertheless not sufficient to make the resulting equations accessible for computations for large molecules; another simplification is therefore introduced by considering that the state function is a N dimensional determinant of simple functions of $\mathbb{R}^3$, called *Slater determinant*:

$$\Phi(r_1, ..., r_N) = \frac{1}{\sqrt{N!}} \det(\Phi_i(r_j)), \tag{1}$$

where $\Phi_i,\ i = 1, ..., N$ are now functions of one variable in $\mathbb{R}^3$ chosen orthogonal with respect to the canonical scalar product $< \cdot, \cdot >$ on $L^2(\mathbb{R}^3)$.

Let us denote by $\mathcal{K}$ the subset of $(L^2(\mathbb{R}^2))^N$ defined by

$$\mathcal{K} = \{(\Phi_1, ..., \Phi_N) \in (L^2(\mathbb{R}^2))^N; < \Phi_i, \Phi_j >= \delta_{ij}\}. \tag{2}$$

Assuming that the molecule is isolated and only Coulombic forces are present, the description of the non-relativistic electrons where, for the sake of simplicity we have neglected the spin dependency, leads to the following expression of the Hartree-Fock energy :

$$\begin{aligned} \mathcal{E}^{HF}(\Phi_1, ..., \Phi_N) \quad &= \sum_{i=1}^{N} \int_{\mathbb{R}^3} \left(|\nabla \Phi_i|^2 + V\ |\Phi_i|^2 \right) + \frac{1}{2} \iint_{\mathbb{R}^3 \times \mathbb{R}^3} \frac{\rho_\Phi(x)\rho_\Phi(y)}{|x-y|} dxdy \\ &- \frac{1}{2} \iint_{\mathbb{R}^3 \times \mathbb{R}^3} \frac{|\rho_\Phi(x,y)|^2}{|x-y|} dxdy, \end{aligned} \tag{3}$$

where the *density matrix* $\rho_\Phi(x, y)$, the *electronic density* $\rho_\Phi(x)$ and the *potential* V are given by the formulaes :

$$\rho_\Phi(x, y) = \sum_{i=1}^{N} \Phi_i(x)\overline{\Phi_i(y)} \tag{4}$$

$$\rho_\Phi(x) = \rho_\Phi(x, x) \tag{5}$$

$$V(x) = -\sum_{j=1}^{m} \frac{Z_j}{|x - \overline{x}_j|}. \tag{6}$$

We have denoted here by $Z_j > 0$ the charge of the j-th nucleo.

In order to determine the ground state of the molecule that, by definition, minimizes the energy (3) under the constraint (2), the Euler Lagrange equations give rise to the Hartree-Fock problem :

Find a $L^2(\mathbb{R}^3)$-orthonormal system $\Phi = \{\Phi_i\}_{i=1,N}^t$ and an hermitian matrix $\Lambda = [\lambda_{i,j}]_{i,j=1,N}$ such that

$$\forall i,\ 1 \leq i \leq N, \quad \mathcal{F}_\Phi(\Phi_i) = -\sum_{j=1}^{N} \lambda_{i,j}\Phi_j, \tag{7}$$

where $\mathcal{F}_\Phi$ is the Fock operator. When acting on an element ψ regular enough of the variable $x \in \mathbb{R}^3$, this operator associates the following function of the $x \in \mathbb{R}^3$ variable:

$$\mathcal{F}_\Phi(\psi)(x) = -\Delta\psi(x) + V(x)\psi(x) + (\rho_\Phi \star \frac{1}{|x|})\psi(x) - \int_{\mathbb{R}^3} \frac{\rho_\Phi(x,y)}{|x-y|}\psi(y)dy. \tag{8}$$

Here $\star$ is the convolution product

$$(f \star g)(x) = \int_{\mathbb{R}^3} f(x-y)g(y)dy. \tag{9}$$

Remark 1 *It is standard to notice that the density matrix is invariant under unitary transforms, i.e. for any element U of the set of the $N \times N$ unitary matrices $\mathcal{U}(N)$:*

$$\forall (x,y) \in \mathbb{R}^3, \quad \rho_\Phi(x,y) = \rho_{U\Phi}(x,y) \tag{10}$$

Hence it follows that the unitary transform U can be chosen in such a way that the hermitian matrix Λ become diagonal: $\Lambda = [\lambda_i]_{i=1,N}$. The solution $\Psi = U\Phi = \{(U\Phi)_i\}_{i=1,N}$ satisfies indeed the more simple Hartree-Fock problem :

$$\forall i,\ 1 \leq i \leq N, \quad \mathcal{F}_\Psi(\psi_i) = -\lambda_i\psi_i \tag{11}$$

The problem then appears as a non linear eigenvalue problem.

This highly nonlinear problem is solved through iterations known as Self Consistent Field approximation; we refer to [4] for a very recent and complete analysis on the convergence of some of these algorithms (Roothaan algorithm and the level shifting algorithm). It is still a very expensive problem since the non linear contribution has a large computational complexity (we refer to [18, 6] for some example of tailored techniques to minimize this complexity). The numerical analysis of the method used typically by the chemists community is most often an open problem an in any case will not provide sound information since most of the numerical approximations are very often at the limit of the convergence. More interesting seems the concept of *a posteriori error*

estimators where, from the computed solution, it is possible to derive reliable information about the validity of the computation that has been done. The purpose of this paper is in this direction.

Denote by $\mathcal{H} = (H^1(\mathbb{R}^3))^N$ the natural space for the solutions of the Hartree-Fock equations and by F_{ij} the mapping $F_{ij} : \mathcal{H} \mapsto \mathbb{R}$ defined over any element $\Phi = (\Phi_i)_{i=1}^N$ by

$$F_{ij}(\Phi) =< \Phi_i, \Phi_j > -\delta_{ij}. \tag{12}$$

In all that follows any N-tuple element $\Phi = (\Phi_i)_{i=1}^N$ will be supposed to be a **column** $(N \times 1)$ vector of $\mathcal{H}$. Consider the minimization problem

$$\inf\{\mathcal{E}^{HF}(\Phi); \Phi \in \mathcal{H} \cap \mathcal{K}\} \tag{13}$$

Remark 2 *The analysis of problem (11) is not completely under control: we can cite the partial results obtained in [8, 9] about the existence of a ground state for positive or neutral molecules and non existence results for negative ions. The basic result of uniqueness of the density solution is still an open problem of outstanding difficulty. Under the hypothesis*

$$\sum_{j=1}^{m} Z_j > N - 1, \tag{14}$$

it has been proven in [9] that a minimum of the problem (13) exists and any such minimum is a solution of the Hartree-Fock equation (7). Moreover, when this problem is written in the form (11) additional information is available on λ_i, namely $\lambda_i > 0$, $i = 1, ..., N$. We will assume in all that follows that (14) is true.

In order to make the presentation easy, we will assume in all that follows that the electronic wavefunction is real and will work on real function spaces; trivial adaptations allow the treatment of complex valued wavefunctions.

2 Error decomposition

2.1 Error metrics

Let $\Phi_0 = (\Phi_{0i})_{i=1}^N \in \mathcal{H} \cap \mathcal{K}$ be a minimum of (13) and $\Phi = (\Phi_i)_{i=1}^N \in \mathcal{H} \cap \mathcal{K}$ an approximation of Φ_0 obtained as the solution of a minimization problem:

$$\inf\{\mathcal{E}^{HF}(\Phi); \Phi \in X^N \cap \mathcal{K}\} \tag{15}$$

where X is a finite dimensional subspace of $H^1(\mathbb{R}^3)$.

The a posteriori analysis on the one hand studies bounds for the difference $\mathcal{E}^{HF}(\Phi_0) - \mathcal{E}^{HF}(\Phi)$ and on the other hand proposes explicit trust intervals on the desired (but unknown) quantity $\mathcal{E}^{HF}(\Phi_0)$ using only the approximate solution at hand Φ ; of course, due to the variational setting, an upper bound on $\mathcal{E}^{HF}(\Phi_0)$ is $\mathcal{E}^{HF}(\Phi)$ itself ; the main focus will therefore be placed on finding lower bounds for $\mathcal{E}^{HF}(\Phi_0)$, which is a non-trivial problem that, to our knowledge, has not been addressed in the literature.

Before dwelling into the a posteriori analysis of (13) it is crucial to introduce the proper definition for the error between a minimizer Φ_0 and its approximation Φ. To this end one has to recall the invariance property of the Hartree-Fock energy:

$$\mathcal{E}^{HF}(\Psi) = \mathcal{E}^{HF}(U\Psi), \forall \Psi \in \mathcal{H} \cap \mathcal{K},\ \forall U \in \mathcal{U}(N) \tag{16}$$

From (16) it follows that if Φ_0 is a minimizer of (13), then for any $U \in \mathcal{U}(N)$, $U\Phi_0$ is also a minimizer and therefore a solution of (7). The same considerations remain true for the problem (15). It is therefore natural to consider the distance between the sets $\{U\Phi_0; U \in \mathcal{U}(N)\}$ and $\{V\Phi; V \in \mathcal{U}(N)\}$ as the most appropriate definition of the distance between Φ_0 and Φ. For reasons that will be made clear later on, we will use in fact an equivalent form (see section 2.3) of the above definition. For any $\Psi_1, \Psi_2 \in \mathcal{H}$ let

$$U_{\Psi_1,\Psi_2} = argmin\{\|U\Psi_1 - \Psi_2\|^2_{(L^2(\mathbb{R}^3))^N}; U \in \mathcal{U}(N)\}. \tag{17}$$

For a given norm $\|\cdot\|$ ($\|\cdot\|_{(L^2)^N}$, $\|\cdot\|_{(H^1)^N}$ ...) we will measure the distance between (sets represented by) Ψ_1 and Ψ_2 as:

$$\|\Psi_1 - \Psi_2\|_\star = \|U_{\Psi_1,\Psi_2}\Psi_1 - \Psi_2\| = \|\Psi_1 - U_{\Psi_2,\Psi_1}\Psi_2\|, \tag{18}$$

the last equality being motivated by the fact that $U_{\Psi_2,\Psi_1} = U^t_{\Psi_1,\Psi_2} \in \mathcal{U}(N)$.

Remark 3 *Note from (17) that U_{Ψ_2,Ψ_1} is intrinsically related to the norm of $(L^2)^N$; when $\|\cdot\| = \|\cdot\|_{(L^2)^N}$ we recover the distance between the sets $\{U\Psi_1; U \in \mathcal{U}(N)\}$ and $\{V\Psi_2; V \in \mathcal{U}(N)\}$.*

The properties of this metric are closely related to the following decomposition of $\mathcal{H}$:

$$\mathcal{H} = \mathcal{A}_\Phi \oplus \mathcal{S}_\Phi \oplus \Phi^{\perp\!\!\!\perp} \tag{19}$$

where for any $\Phi \in \mathcal{H} \cap \mathcal{K}$:

$$\mathcal{A}_\Phi = \{C\Phi; C \in \mathbb{R}^{N\times N}, C^t = -C\} \tag{20}$$

$$\mathcal{S}_\Phi = \{S\Phi; S \in \mathbb{R}^{N\times N}, S^t = S\} \tag{21}$$

$$\Phi^{\perp\!\!\!\perp} = \{\Psi = (\psi_i)_{i=1}^N \in \mathcal{H}; < \psi_i, \Phi_j >= 0; i,j = 1, ..., N\} \tag{22}$$

We will denote for any $\Psi_1, \Psi_2 \in (L^2)^N$: $\Psi_1 \perp\!\!\!\perp \Psi_2$ if for any $i, j = 1, N$: $< (\Psi_1)_i, (\Psi_2)_j >= 0$; then $\Phi^{\perp\!\!\!\perp}$ can be defined equivalently

$$\Phi^{\perp\!\!\!\perp} = \{\Psi \in \mathcal{H}; \Psi \perp\!\!\!\perp \Phi\}.$$

For any $\xi = (\xi_i)_{i=1}^N \in \mathcal{H}$ the decomposition (19) is obtained in the following manner: compute the matrix $M = (M_{ij})_{i,j=1}^N$ where for each $i, j = 1, ..., N$: $M_{ij} =< \xi_i, \Phi_j >$. Denote by S the symmetric part of M: $S = \frac{M+M^t}{2}$ and by C the antisymmetric part: $C = \frac{M-M^t}{2}$. Then $S\Phi$ will be the component of ξ in the space $\mathcal{S}_\Phi$ and $C\Phi$ the component of ξ in the space $\mathcal{A}_\Phi$; in addition it is easy to see that $(\xi - S\Phi - C\Phi) \perp\!\!\!\perp \Phi$, so the difference $\xi - S\Phi - C\Phi$ is in $\Phi^{\perp\!\!\!\perp}$.

Lemma 1 *Let $\Phi, \Psi \in \mathcal{H} \cap \mathcal{K}$. Then the matrix $U_{\Psi,\Phi}$ solution of (17) has the properties*

$$U_{\Psi,\Phi}\Psi - \Phi \in \mathcal{S}_\Phi \oplus \Phi^{\perp\!\!\!\perp}, \quad \Phi - U_{\Psi,\Phi}\Psi \in \mathcal{S}_{U_{\Psi,\Phi}\Psi} \oplus \Psi^{\perp\!\!\!\perp}. \tag{23}$$

In particular for $\Psi = \Phi_0$,

$$U_{\Phi_0,\Phi}\Phi_0 = \Phi + S\Phi + W, \ S \in \mathbb{R}^{N\times N} : S^t = S, \ W \in \Phi^{\perp\!\!\!\perp}. \tag{24}$$

Proof: Consider the decomposition

$$\Psi - \Phi = C\Phi + S\Phi + W, \ C\Phi \in \mathcal{A}_\Phi, \ S\Phi \in \mathcal{S}_\Phi, \ W \in \Phi^{\perp\!\!\!\perp}, \tag{25}$$

and denote $M = C + S$. Then we can write

$$\begin{aligned}
&U_{\Psi,\Phi} = argmin\{\|U\Psi - \Phi\|^2_{(L^2(\mathbb{R}^3))^N}; U \in \mathcal{U}(N)\} \\
&= argmin\{\|U((Id_N + M)\Phi + W) - \Phi\|^2_{(L^2(\mathbb{R}^3))^N}; U \in \mathcal{U}(N)\} \\
&= argmin\{\|(U(Id_N + M) - Id_N)\Phi\|^2_{(L^2(\mathbb{R}^3))^N}; U \in \mathcal{U}(N)\} \\
&= argmin\{\|U(Id_N + M) - Id_N)\|^2_{\mathbb{R}^{N\times N}}; U \in \mathcal{U}(N)\} \\
&= argmin\{\|(Id_N + M) - U^t\|^2_{\mathbb{R}^{N\times N}}; U \in \mathcal{U}(N)\}
\end{aligned} \tag{26}$$

The transformation from the second to the third line is a consequence of the fact that $W \perp\!\!\!\perp \Phi$ so therefore $U(Id_N + M)\Phi \perp\!\!\!\perp W$; the next equality is true because $\Phi \in \mathcal{K}$.
For any antisymmetric matrix $\tilde{C} \in \mathbb{R}^{N\times N}$ consider the path in $\mathcal{U}(N)$ given by $t \to e^{\tilde{C}t}U_{\Psi,\Phi}$. The tangent at $t = 0$ to this path is $\tilde{C}U_{\Psi,\Phi}$. Writing the first order conditions for the minimality in (26) we obtain:

$$\begin{aligned}
0 &= < (Id_N + M) - U^t_{\Psi,\Phi}, U^t_{\Psi,\Phi}\tilde{C}^t >_{\mathbb{R}^{N\times N}} \\
&= < U_{\Psi,\Phi}(Id_N + M) - Id_N, \tilde{C}^t >_{\mathbb{R}^{N\times N}}, \ \forall \tilde{C} \in \mathbb{R}^{N\times N} : \tilde{C}^t = -\tilde{C},
\end{aligned} \tag{27}$$

which shows that $U_{\Psi,\Phi}(Id_N + M)$ is a symmetric matrix ; and therefore $U_{\Psi,\Phi}\Psi \in \mathcal{S}_\Phi \oplus \Phi^{\perp\!\!\!\perp}$. To prove the second part of the equation (23) denote for any Ψ_1, Ψ_2 by C_{Ψ_1,Ψ_2} the antisymmetric matrix appearing in the decomposition $\Psi_1 - \Psi_2 = C_{\Psi_1,\Psi_2}\Psi_2 + S_{\Psi_1,\Psi_2}\Psi_2 + W_{\Psi_1,\Psi_2}$ with $C_{\Psi_1,\Psi_2}\Psi_2 \in \mathcal{A}\Psi_2$, $S_{\Psi_1,\Psi_2}\Psi_2 \in \mathcal{S}\Psi_2$ and $W_{\Psi_1,\Psi_2} \in \Psi_2^{\perp\!\!\!\perp}$; then one obtains by straightforward computations $C_{\Psi_1,\Psi_2} = -C_{\Psi_2,\Psi_1}$. $\square$

Remark 4 *In practice the representative of the class of isoenergy functions* $\{U\Phi_0; U \in \mathcal{U}(N)\}$ *is taken to be the one that solves equations (11), and the same is true for any of its approximations* Φ*. It is not clear whether a norm for which this practical choice give optimal approximations in the sense of (17) exists and to what extent this choice is also optimal in the* L^2 *norm.*

2.2 Order of the symmetric part of the error

Let $\Psi, \Phi \in \mathcal{H} \cap \mathcal{K}$ and let us consider the decomposition (25). We have seen that the antisymmetric part given by matrix C may be set to zero modulo some appropriate "rotation" on Ψ ; it is therefore natural to study the properties of the symmetric part $S\Phi$.

Lemma 2 *Let* $\Psi, \Phi \in \mathcal{H} \cap \mathcal{K}$ *with associated decomposition (25). Then there exists constants* C_1*,* C_2 *depending only of* N *such that:*

$$\|S\Phi\|_{(L^2(\mathbb{R}^3))^N} \leq C_1 \|\Psi - \Phi\|^2_{(L^2(\mathbb{R}^3))^N} \tag{28}$$

$$\|S\Phi\|_{\mathcal{H}} \leq C_2 \|\Psi - \Phi\|^2_{\mathcal{H}} \|\Phi\|_{\mathcal{H}} \tag{29}$$

Proof: Let us write $W = D\tilde{W}$ such that $< \tilde{W}_i, \tilde{W}_j >= \delta_{ij}$, $M = C + S$. Denote

$$\epsilon = \|\Psi - \Phi\|_{(L^2(\mathbb{R}^3))^N} = \sqrt{\sum_{i,j=1}^{N} M_{ij}^2 + D_{ij}^2}$$

Since $\Psi \in \mathcal{K}$, $F_{ij}(\Psi) = 0$, $i, j = 1, ..., N$. For $j = i$ we obtain:

$$1 = (1 + M_{ii})^2 + \sum_{j \neq i} M_{ij}^2 + \sum_{j=1}^{N} D_{ij}^2,$$

or equivalently:

$$S_{ii} = M_{ii} = -\frac{\sum_{j=1}^{N} M_{ij}^2 + \sum_{j=1}^{N} D_{ij}^2}{2},$$

which proves that $M_{ii} \leq \epsilon^2$, $i = 1, ..., N$. For $i \neq j$ one obtains:

$$0 = \sum_{k \neq i, k \neq j} M_{ik} M_{jk} + (M_{ii} + 1) M_{ji} + M_{ij}(M_{jj} + 1) + \sum_{k=1}^{N} D_{ki} D_{kj},$$

which gives after straightforward manipulations $S_{ij} = \frac{M_{ij} + M_{ji}}{2} \leq \epsilon^2$; this concludes the proof of (28). For (29) one denotes first that $\|\Psi - \Phi\|_{(L^2(\mathbb{R}^3))^N} \leq \|\Psi - \Phi\|_{\mathcal{H}}$ and apply (28) to conclude that $S_{ij} \leq \|\Psi - \Phi\|_{\mathcal{H}}^2$, $i, j = 1, ..., N$. The conclusion follows then by the definition of the norm $\| \cdot \|_{\mathcal{H}}$. $\square$

2.3 Optimality in H^1 norm

We have proposed in section 2.1 that for any norm $\| \cdot \|$ the error $\Phi_0 - \Phi$ be computed as $\|U_{\Phi_0, \Phi} \Phi_0 - \Phi\|$. Since the definition $U_{\Phi_0, \Phi}$ is closely related to the L^2 norm it is natural to ask whether this definition is still appropriate when norms other than L^2 are used, for instance the canonical norm of $\mathcal{H}$. The situation is settled by the following

Lemma 3 *Let $\Psi = (\psi_1, ..., \psi_N) \in \mathcal{H} \cap \mathcal{K}$ and $\Phi \in \mathcal{H} \cap \mathcal{K}$ and denote*

$$U^1_{\Psi, \Phi} = argmin\{\|U\Psi - \Phi\|_{\mathcal{H}}; U \in \mathcal{U}(N)\}$$

There exists a constant c depending only of N and Ψ such that

$$c\|U_{\Psi, \Phi}\Psi - \Phi\|_{\mathcal{H}} \leq \|U^1_{\Psi, \Phi}\Psi - \Phi\|_{\mathcal{H}} \leq \|U_{\Psi, \Phi}\Psi - \Phi\|_{\mathcal{H}} \tag{30}$$

Proof: The inequality

$$\|U_{\Psi, \Phi}\Psi - \Phi\|_{\mathcal{H}} \geq \|U^1_{\Psi, \Phi}\Psi - \Phi\|_{\mathcal{H}}$$

follows as a consequence of the definition of $U^1_{\Psi, \Phi}$.

Denote by F the linear space generated by $\{\psi_1, ..., \psi_N\}$ and define:

$$M = \{\zeta \in H^1(\mathbb{R}^3); < \zeta, \chi >_{L^2, L^2} = 0, \ \forall \chi \in F\}.$$

For any $\chi \in H^1(\mathbb{R}^3)$ denote by χ_F the L^2 projection of χ on F and $\chi_M = \chi - \chi_F$. We define a norm $\| \cdot \|_d$ on $H^1(\mathbb{R}^3)$ as follows:

$$\|\chi\|_d^2 = \|\chi_F\|_{L^2}^2 + \|\chi_M\|_{H^1(\mathbb{R}^3)}^2.$$

We will prove that this norm is equivalent to the canonical norm of $H^1(\mathbb{R}^3)$ (with constants depending only N and Ψ). Write for any $\chi \in H^1(\mathbb{R}^3)$:

$$\|\chi\|_{H^1(\mathbb{R}^3)} \leq \|\chi - \chi_F\|_{H^1(\mathbb{R}^3)} + \|\chi_F\|_{H^1(\mathbb{R}^3)} \leq \|\chi\|_d + \|\chi_F\|_{H^1(\mathbb{R}^3)} \leq C\|\chi\|_d$$

where we have used the fact that the norms $\|\cdot\|_{L^2}$ and $\|\cdot\|_{H^1(\mathbb{R}^3)}$ are equivalent on the **finite dimensional** space F. It follows that there exists a constant C (depending only N and Ψ) such that for any $\chi \in H^1(\mathbb{R}^3)$

$$\|\chi\|_{H^1(\mathbb{R}^3)} \leq C\|\chi\|_d.$$

We will prove next that the norm $\|\cdot\|_{H^1(\mathbb{R}^3)}$ can also be lower bounded by the norm $\|\cdot\|_d$ modulo some constant depending only N and Ψ. Assume on the contrary that this is not true. Then there exists a sequence $(\chi_n)_{n\geq 1} \subset H^1(\mathbb{R}^3)$ such that $\|\chi_n\|_d = 1$ and $\|\chi_n\|_{H^1(\mathbb{R}^3)} \to 0$ as $n \to \infty$. It follows that the sequence χ_n converges to zero in L^2 and in particular the sequence $(\chi_{nF})_{n\geq 1}$ of L^2 projections to F is also converging to zero: $\|\chi_{nF}\|_{L^2} \to 0$ $(n \to \infty)$; by the same argument as above we obtain $\|\chi_{nF}\|_{H^1(\mathbb{R}^3)} \to 0$ $(n \to \infty)$. Then

$$\|\chi_{nM}\|_{H^1(\mathbb{R}^3)} = \|\chi_n - \chi_{nF}\|_{H^1(\mathbb{R}^3)} \leq \|\chi_n\|_{H^1(\mathbb{R}^3)} + \|\chi_{nF}\|_{H^1(\mathbb{R}^3)}$$

and it follows that $\|\chi_{nM}\|_{H^1(\mathbb{R}^3)} \to 0$ $(n \to \infty)$. Together with $\|\chi_{nF}\|_{L^2} \to 0$ $(n \to \infty)$ we conclude that $\|\chi_n\|_d \to 0$ $(n \to \infty)$, in contradiction with the initial assumption. We have therefore proved that there exists constants c, C (depending only N and Ψ) such that for any $\chi \in H^1(\mathbb{R}^3)$

$$c\|\chi\|_d \leq \|\chi\|_{H^1(\mathbb{R}^3)} \leq C\|\chi\|_d.$$

The above equivalence imply that canonical norm $\|\cdot\|_{d,N}$ of $(H^1(\mathbb{R}^3), \|\cdot\|_d)^N$ is equivalent (with constants depending only on N and Ψ) to the canonical norm of $\mathcal{H}$:

$$c_1\|\zeta\|_{\mathcal{H}} \leq \|\zeta\|_{d,N} \leq C_1\|\zeta\|_{\mathcal{H}},\ \forall \zeta \in \mathcal{H}.$$

Since $\Psi \in \mathcal{K}$, the functions $\{\psi_1, ..., \psi_N\}$ are orthonormal with respect to the scalar product of $L^2(\mathbb{R}^3)$ and also with respect to the scalar product $< \cdot, \cdot >_d$ associated to the norm $\|\cdot\|_d$. It follows by (26) that

$$U_{\Psi,\Phi} = argmin\{\|U\Psi - \Phi\|_{\|\cdot\|_{d,N}}; U \in \mathcal{U}(N)\},$$

as both solve the same minimization problem on $\mathcal{U}(N)$. But then

$$\|U^1_{\Psi,\Phi}\Psi - \Phi\|_{\mathcal{H}} \geq \frac{1}{C_1}\|U^1_{\Psi,\Phi}\Psi - \Phi\|_{d,N} \geq \frac{1}{C_1}\|U_{\Psi,\Phi}\Psi - \Phi\|_{d,N} \geq \frac{c_1}{C_1}\|U_{\Psi,\Phi}\Psi - \Phi\|_{\mathcal{H}}.$$

which concludes the proof. □

3 Optimality conditions and coercivity

We will begin this section with some elementary information about the geometry of the manifolds $\mathcal{K}$ and $\mathcal{H} \cap \mathcal{K}$:

Lemma 4 *Let $\Phi \in \mathcal{H} \cap \mathcal{K}$. The tangent space in Φ to the manifold $\mathcal{H} \cap \mathcal{K}$ is $\mathcal{A}_\Phi \oplus \Phi^{\perp\!\!\!\perp}$.*

Proof: Let $\Phi(t) :]-\epsilon, \epsilon[\to \mathcal{H} \cap \mathcal{K}$, $\epsilon > 0$, $\Phi(0) = \Phi$ be a C^1 path in $\mathcal{H} \cap \mathcal{K}$. Consider the decomposition $\Phi'(0) = S\Phi + C\Phi + W$, $S\Phi \in \mathcal{S}_\Phi$, $C\Phi \in \mathcal{A}_\Phi$, $W \in \Phi^{\perp\!\!\!\perp}$. By differentiating the condition $F_{ij}(\Phi(t)) = 0$ we obtain $< \Phi_i, \Phi_j'(0) > + < \Phi_i'(0), \Phi_j >= 0$ which proves that $S_{ij} = 0$. Since this is true for any $i, j = 1, ..., N$ we conclude $S = 0$ i.e. $\Phi'(0) \in \mathcal{A}_\Phi \oplus \Phi^{\perp\!\!\!\perp}$.

To prove that any $\Psi = C\Phi + W \in \mathcal{A}_\Phi \oplus \Phi^{\perp\!\!\!\perp}$ may be seen as the tangent in Φ of a C^1 path in $\mathcal{H} \cap \mathcal{K}$, choose $\Phi(t) :]-\epsilon, \epsilon[\to \mathcal{H} \cap \mathcal{K}$, $0 < \epsilon < 1$, $\Phi(t) = \sqrt{1-t^2}e^{Ct}\Phi + tW$ and note that $\Phi'(0) = \Psi$. $\square$

Remark 5 *The Hartree-Fock equations (7) can be "symbolically" derived as a corollary of lemma 4. Indeed, the first order minimality conditions associated to (13) read*

$$< D\mathcal{E}^{HF}(\Phi_0), \Psi >_{(L^2(\mathbb{R}^3))^N} = 0, \ \forall \Psi \in \mathcal{A}_{\Phi_0} \oplus \Phi_0^{\perp\!\!\!\perp}$$

which is the same as writing $D\mathcal{E}^{HF}(\Phi_0) = S\Phi_0$, ($S$ being a symmetric matrix) which are exactly equations (7) since $D\mathcal{E}^{HF}(\Phi_0)$ can be identified with $(\mathcal{F}_{\Phi_0}, ..., \mathcal{F}_{\Phi_0})$.

The second order optimality conditions for the minimization problem (13) will be seen to be very useful within our approach. Let $\Phi_0 \in \mathcal{H} \cap \mathcal{K}$ be a minimizer of (13) and Λ^0 be the hermitian matrix corresponding to Φ_0 in equations (7). We will write the second order conditions in the form:

$$D^2\mathcal{E}^{HF}(\Phi_0)(\Psi, \Psi) + < \Lambda^0\Psi, \Psi >_{(L^2(\mathbb{R}^3))^N} \geq 0, \ \forall \Psi \in \mathcal{A}_{\Phi_0} \oplus \Phi_0^{\perp\!\!\!\perp}. \tag{31}$$

Denote for any $\Phi \in \mathcal{H} \cap \mathcal{K}$: $\mathcal{E}^\Phi(\cdot) = \mathcal{E}^{HF}(\cdot) + \sum_{i,j=1}^N \Lambda_{ij}F_{ij}(\cdot)$ where $\Lambda_{ij} =< \mathcal{F}_\Phi\Phi_i, \Phi_j >$, $i, j = 1, ..., N$. Denote also by $a_\Phi(\cdot, \cdot)$ the bilinear form $D^2\mathcal{E}^\Phi(\Phi)(\cdot, \cdot)$ and remark that $a_{\Phi_0}(\cdot, \cdot) = D^2\mathcal{E}^{HF}(\Phi_0)(\cdot, \cdot) + < \Lambda^0\cdot, \cdot >_{(L^2(\mathbb{R}^3))^N}$. In order to obtain an explicit formula for a_{Φ_0} we need the expression of $D^2\mathcal{E}^{HF}(\Phi_0)$. Let $\Phi, \Psi^1, \Psi^2 \in \mathcal{H} \cap \mathcal{K}$. Then

$$\begin{aligned}
D^2\mathcal{E}^{HF}(\Phi)(\Psi^1, \Psi^2) &= 2 \cdot \sum_{i=1}^N \int_{\mathbb{R}^3} \left(\nabla\Psi_i^1 \cdot \nabla\Psi_i^2 + V\Psi_i^1\Psi_i^2\right) \\
&+ \frac{1}{2} \iint_{\mathbb{R}^3 \times \mathbb{R}^3} \frac{8\rho_{\Phi,\Psi^1}(x)\rho_{\Phi,\Psi^2}(y) + 4\rho_{\Psi^1,\Psi^2}(x)\rho_\Phi(y)}{|x-y|} dxdy \\
&- \frac{1}{2} \iint_{\mathbb{R}^3 \times \mathbb{R}^3} \frac{2\rho_\Phi(x,y)(\rho_{\Psi^1,\Psi^2}(x,y) + \rho_{\Psi^1,\Psi^2}(y,x))}{|x-y|} dxdy \\
&- \frac{1}{2} \iint_{\mathbb{R}^3 \times \mathbb{R}^3} \frac{+4\rho_{\Phi,\Psi^1}(x,y)(\rho_{\Phi,\Psi^2}(x,y) + \rho_{\Phi,\Psi^2}(y,x))}{|x-y|} dxdy,
\end{aligned} \tag{32}$$

with the definitions

$$\rho_{\Psi^1,\Psi^2}(x,y) = \sum_{i=1}^N \Psi_i^1(x)\Psi_i^2(y), \tag{33}$$

$$\rho_{\Psi^1,\Psi^2}(x) = \rho_{\Psi^1,\Psi^2}(x,x). \tag{34}$$

We obtain therefore:

$$\begin{aligned} D^2\mathcal{E}^{HF}(\Phi_0)(\Psi,\Psi) &= 2\cdot\sum_{i=1}^N \int_{\mathbb{R}^3} \left(|\nabla\Psi_i|^2 + V\Psi_i^2\right) \\ &+\frac{1}{2}\iint_{\mathbb{R}^3\times\mathbb{R}^3} \frac{8\rho_{\Phi_0,\Psi}(x)\rho_{\Phi_0,\Psi}(y) + 4\rho_\Psi(x)\rho_{\Phi_0}(y)}{|x-y|}dxdy \\ &-\frac{1}{2}\iint_{\mathbb{R}^3\times\mathbb{R}^3} \frac{4\rho_{\Phi_0}(x,y)\rho_\Psi(x,y)}{|x-y|}dxdy \\ &-\frac{1}{2}\iint_{\mathbb{R}^3\times\mathbb{R}^3} \frac{4\rho_{\Phi_0,\Psi}(x,y)(\rho_{\Phi_0,\Psi}(x,y)+\rho_{\Phi_0,\Psi}(y,x))}{|x-y|}dxdy. \end{aligned} \tag{35}$$

We will study in the following the coercivity properties of the bilinear form a_{Φ_0}. Note that for any $\Psi \in \mathcal{H}\cap\mathcal{K}$: $\mathcal{E}^{HF}(\Psi) = \mathcal{E}^{\Phi_0}(\Psi)$ and in addition $a_{\Phi_0} = D^2\mathcal{E}^{\Phi_0}(\Phi_0)$. By differentiating the invariance property (16) we obtain in particular (cf. lemma 4):

$$D\mathcal{E}^{\Phi_0}(\Psi)(C\Psi) = 0, \forall\Psi\in\mathcal{H}\cap\mathcal{K},\ \forall C\Psi\in\mathcal{A}_\Psi. \tag{36}$$

Differentiating now (36) in $\Psi=\Phi_0$ and taking into account the fact that Φ_0 is a solution of (7) we obtain:

$$< D^2\mathcal{E}^{\Phi_0}(\Phi_0)(C\Phi_0, \tilde{C}\Phi_0 + W) = 0,\ \forall C\Phi_0, \tilde{C}\Phi_0 \in \mathcal{A}_{\Phi_0},\ \forall W\in\Phi_0^{\perp\!\!\!\perp}. \tag{37}$$

Then it follows that a_{Φ_0} vanishes on $\mathcal{A}_{\Phi_0}$ thus cannot be coercive there ; the coercivity properties of a_{Φ_0} are described by the following two lemmata.

Lemma 5 *Let V_{Φ_0} be the closure of span$\{\Psi\in\mathcal{A}_{\Phi_0}\oplus\Phi_0^{\perp\!\!\!\perp} : a_{\Phi_0}(\Psi,\Psi)=0\}$ with respect to the canonical topology of $\mathcal{H}$. Then a_{Φ_0} is null on $V_{\Phi_0}\times V_{\Phi_0}$.*

Proof: Let $\Psi^1,\Psi^2\in\mathcal{A}_{\Phi_0}\oplus\Phi_0^{\perp\!\!\!\perp}$ be such that $a_{\Phi_0}(\Psi^i,\Psi^i)=0$, $i=1,2$. Then since $a_{\Phi_0}\geq 0$ on $\mathcal{A}_{\Phi_0}\oplus\Phi_0^{\perp\!\!\!\perp}$ by a standard Cauchy-Schwartz inequality for the positive bilinear form a_{Φ_0} we obtain $2|a_{\Phi_0}(\Psi^1,\Psi^2)| \leq a_{\Phi_0}(\Psi^1,\Psi^1) + a_{\Phi_0}(\Psi^2,\Psi^2)$ and therefore $a_{\Phi_0}(\Psi^1,\Psi^2)=0$. It follows then that for any $\Psi=\mu_1\Psi^1+\mu_2\Psi^2$ such that $\mu_1,\mu_2\in\mathbb{R}$ we have $a_{\Phi_0}(\Psi,\Psi)=0$ which, together with the continuity of a_{Φ_0} concludes the proof. □

Lemma 6 *Let X_{Φ_0} be a closed subspace of $\Phi_0^{\perp\!\!\!\perp}$ $(\mathcal{H})$ such that*

$$\forall \Psi \in X_{\Phi_0}, \Psi \neq 0 : \ a_{\Phi_0}(\Psi, \Psi) > 0. \tag{38}$$

Then a_{Φ_0} is coercive on X_{Φ_0}.

Proof: The proof makes use of the following auxiliary result

Lemma 7 *The mapping*

$$\begin{aligned}\Psi \mapsto & \frac{1}{2}\iint_{\mathbb{R}^3\times\mathbb{R}^3} \frac{8\rho_{\Phi_0,\Psi}(x)\rho_{\Phi_0,\Psi}(y) + 4\rho_\Psi(x)\rho_{\Phi_0}(y)}{|x-y|}dxdy \\ & -\frac{1}{2}\iint_{\mathbb{R}^3\times\mathbb{R}^3} \frac{4\rho_{\Phi_0}(x,y)\rho_\Psi(x,y)}{|x-y|}dxdy \\ & -\frac{1}{2}\iint_{\mathbb{R}^3\times\mathbb{R}^3} \frac{4\rho_{\Phi_0,\Psi}(x,y)(\rho_{\Phi_0,\Psi}(x,y) + \rho_{\Phi_0,\Psi}(y,x))}{|x-y|}dxdy \end{aligned} \tag{39}$$

is sequentially weakly lower semicontinuous with respect to the canonic topology of $\mathcal{H}$.

Proof: Let us recall the Hardy inequality ([9] p.42) which holds for all $y \in \mathbb{R}^3$, $\varphi \in H^1(\mathbb{R}^3)$:

$$\int_{\mathbb{R}^3} \frac{|\varphi(x)|^2}{|x-y|}dx \leq C\|\varphi\|_{L^2(\mathbb{R}^3)}\|\nabla\varphi\|_{L^2(\mathbb{R}^3)} \tag{40}$$

with a constant C independent of y and φ. Note that if $u, v \in H^1(\mathbb{R}^3)$ $\frac{u(x)v(y)}{\sqrt{|x-y|}} \in L^2(\mathbb{R}^3 \times \mathbb{R}^3)$. Indeed:

$$\begin{aligned}&\iint_{\mathbb{R}^3\times\mathbb{R}^3} \frac{u^2(x)v^2(y)}{|x-y|}dxdy = \int_{\mathbb{R}^3}\Big(\int_{\mathbb{R}^3} \frac{u^2(x)}{|x-y|}dx\Big)v^2(y)dy \\ &\leq C\|u\|_{L^2(\mathbb{R}^3)}\|\nabla u\|_{L^2(\mathbb{R}^3)}\int_{\mathbb{R}^3} v^2(y)dy \leq C\|u\|_{L^2(\mathbb{R}^3)}\|\nabla u\|_{L^2(\mathbb{R}^3)}\|v\|^2_{L^2(\mathbb{R}^3)}\end{aligned}$$

Let Ψ^m be a sequence weakly convergent in $\mathcal{H}$ to Ψ; this sequence is bounded in $\mathcal{H}$; without loss of generality it can be supposed that $\|\Psi^m\|_{\mathcal{H}} \leq 1$. Consider a term of the form

$$\iint_{\mathbb{R}^3\times\mathbb{R}^3} \frac{f(x)g(y)\Psi_i^m(x)\Psi_j^m(y)}{|x-y|}dxdy \tag{41}$$

where $f, g \in \{(\Phi_0)_1, ..., (\Phi_0)_N\}$. We have seen that $\frac{f(x)g(y)}{\sqrt{|x-y|}}, \frac{\Psi_i^m(x)\Psi_j^m(y)}{\sqrt{|x-y|}} \in L^2(\mathbb{R}^3\times\mathbb{R}^3)$; since $\|\Psi^m\|_{\mathcal{H}} \leq 1$, it follows that $\frac{\Psi_i^m(x)\Psi_j^m(y)}{\sqrt{|x-y|}}$ is weakly convergent

in $L^2(\mathbb{R}^3\times\mathbb{R}^3)$ to[2] $\frac{\Psi_i(x)\Psi_j(y)}{\sqrt{|x-y|}}$ so any term of the form (41) is weakly continuous (so also lower weakly semicontinuous), and of course the same is true for any sum of terms of this type, in particular $\frac{\rho_{\Phi_0,\Psi^m}(x)\rho_{\Phi_0,\Psi^m}(y)}{|x-y|}$, $\frac{\rho_{\Phi_0}(x,y)\rho_{\Psi^m}(x,y)}{|x-y|}$, $\frac{\rho_{\Phi_0,\Psi^m}(x,y)\rho_{\Phi_0,\Psi^m}(y,x)}{|x-y|}$.

The only term that remains to be analyzed in (39) is

$$4\iint_{\mathbb{R}^3\times\mathbb{R}^3}\frac{\rho_\Psi(x)\rho_{\Phi_0}(y)-\rho_{\Phi_0,\Psi}(x,y)^2}{|x-y|}dxdy \tag{42}$$

We transform the numerator of the above fraction as follows:

$$\begin{aligned}
&\rho_\Psi(x)\rho_{\Phi_0}(y)-(\rho_{\Phi_0,\Psi}(x,y))^2=\sum_{i=1}^N(\Psi_i)^2(x)(\Phi_0)_i^2(y)\\
&+\sum_{i<j}(\Psi_i)^2(x)(\Phi_0)_j^2(y)+(\Psi_j)^2(x)(\Phi_0)_i^2(y)\\
&-\sum_{i=1}^N(\Psi_i)^2(x)(\Phi_0)_i^2(y)-\sum_{i<j}\Psi_i(x)(\Phi_0)_i(y)\Psi_j(x)(\Phi_0)_j(y)\\
&=\sum_{i<j}\Big(\Psi_i(x)(\Phi_0)_j(y)-\Psi_j(x)(\Phi_0)_i(y)\Big)^2
\end{aligned}\tag{43}$$

It is easy to see from this equality that $\rho_\Psi(x)\rho_{\Phi_0}(y)-(\rho_{\Phi_0,\Psi}(x,y))^2$ is a convex function of Ψ and therefore, by a classical functional analysis argument, is sequentially weakly lower semicontinuous. □

Let us proceed with the proof of lemma 6. Suppose on contrary that the conclusion is not true. Then there exists a sequence $\{\Psi^m\}_{m\geq1}\in X_{\Phi_0}$ such that $\|\Psi^m\|_{\mathcal{H}}=1$, and $\lim_{m\to\infty}a_{\Phi_0}(\Psi^m,\Psi^m)=0$; extracting if necessary a subsequence out of it, we may suppose that $\{\Psi^m\}_{m\geq1}$ is weakly convergent in $\mathcal{H}$ to $\Psi\in X_{\Phi_0}$. We first write:

$$\begin{aligned}
&a_{\Phi_0}(\Psi^m,\Psi^m)=2\cdot\sum_{i=1}^N\int_{\mathbb{R}^3}|\nabla\Psi_i^m|^2+2\cdot\sum_{i,j=1}^N\Lambda_{ij}^0\int_{\mathbb{R}^3}\Psi_i^m\Psi_j^m+\\
&2\cdot\sum_{i=1}^N\int_{\mathbb{R}^3}V(\Psi_i^m)^2+\frac{1}{2}\iint_{\mathbb{R}^3\times\mathbb{R}^3}\frac{8\rho_{\Phi_0,\Psi^m}(x)\rho_{\Phi_0,\Psi^m}(y)+4\rho_{\Psi^m}(x)\rho_{\Phi_0}(y)}{|x-y|}dxdy\\
&-\frac{1}{2}\iint_{\mathbb{R}^3\times\mathbb{R}^3}\frac{4\rho_{\Phi_0}(x,y)\rho_{\Psi^m}(x,y)}{|x-y|}dxdy\\
&-\frac{1}{2}\iint_{\mathbb{R}^3\times\mathbb{R}^3}\frac{4\rho_{\Phi_0,\Psi^m}(x,y)(\rho_{\Phi_0,\Psi^m}(x,y)+\rho_{\Phi_0,\Psi^m}(y,x))}{|x-y|}dxdy
\end{aligned}\tag{44}$$

[2]In order to rigorously identify the weak limit one uses appropriate test functions $\sqrt{|x-y|}\alpha(x)\beta(y)\mathbf{1}_{|x|\leq R}\mathbf{1}_{|y|\leq R}$ for any $\alpha,\beta\in L^2(\mathbb{R}^3)$, $R>0$.

Recall that ([9] p.42) that $\int_{\mathbb{R}^3} V\psi^2 dx$ is weakly lower semicontinuous on $H^1(\mathbb{R}^3)$ ([9] p.42). By the lemma 7 the integrals on $\mathbb{R}^3 \times \mathbb{R}^3$ in (44) also have weakly lower semicontinuity properties. Since the matrix Λ^0 has **strictly positive** eigenvalues (remark 2) the first two terms on the right hand side of (44) define a norm so this part is also weakly lower semicontinuous ; we obtain

$$a_{\Phi_0}(\Psi, \Psi) \leq \lim_{m\to\infty} a_{\Phi_0}(\Psi^m, \Psi^m) = 0$$

which together with (38) imply $\Psi = 0$. We will use now this information for a finer analysis of the sequence $a_{\Phi_0}(\Psi^m, \Psi^m)$; by the argument above there exists a constant $c_0 > 0$ depending on Φ_0 such that for any $\Psi \in \mathcal{H}$:

$$\sum_{i=1}^{N} \int_{\mathbb{R}^3} 2|\nabla\Psi_i|^2 + \sum_{i,j=1}^{N} \Lambda^0_{ij} \int_{\mathbb{R}^3} \Psi_i\Psi_j \geq c_0 \|\Psi\|_{\mathcal{H}}.$$

Using again the lower semicontinuity of the remaining terms we obtain:

$$0 = \lim_{m\to\infty} a_{\Phi_0}(\Psi^m, \Psi^m) \geq 0 \ + \ \liminf_{m\to\infty} \sum_{i=1}^{N} \int_{\mathbb{R}^3} 2|\nabla\Psi_i^m|^2 +$$

$$\sum_{i,j=1}^{N} \Lambda^0_{ij} \int_{\mathbb{R}^3} \Psi_i^m\Psi_j^m \geq c_0 \liminf_{m\to\infty} \|\Psi^m\|_{\mathcal{H}} = c_0 > 0,$$

which is impossible. □

Motivated by the above analysis, we will introduce the following hypothesis:

$$\forall \Psi \in \Phi_0^{\perp\!\!\!\perp}, \Psi \neq 0 : \ a_{\Phi_0}(\Psi, \Psi) > 0. \tag{45}$$

which, by lemma 6, assures the existence of a "coercivity constant" $c_{\Phi_0} > 0$ such that

$$\forall \Psi \in \Phi_0^{\perp\!\!\!\perp}, \Psi \neq 0 : \ a_{\Phi_0}(\Psi, \Psi) \geq c_{\Phi_0} \|\Psi\|^2_{\mathcal{H}}. \tag{46}$$

Remark 6 *Using the lemma 5 a posteriori analysis may still be carried out without the hypothesis 45 ; some aspects of a more general analysis are presented in remark 12.*

4 Error estimators, bounds and convergence acceleration

Let $\Phi_0, \Phi \in \mathcal{H} \cap \mathcal{K}$ be as in section 2.1: Φ_0 a minimizer of (13) (which is thus a solution of (7)) and $\Phi \in \mathcal{H} \cap \mathcal{K}$ a known discrete approximation of Φ_0 obtained by a previous computation.

Let us denote by $\epsilon = \|U_{\Phi_0,\Phi}\Phi_0 - \Phi\|_{\mathcal{H}} = \|U_{\Phi,\Phi_0}\Phi - \Phi_0\|_{\mathcal{H}}$ the distance between Φ and Φ_0. Even if the wavefunction Φ_0 may be intrinsically interesting (e.g. when the form of the molecular orbitals is studied), the main result of a Hartree-Fock computation is the Hartree-Fock energy $\mathcal{E}^{HF}(\Phi_0)$.

We will suppose in all that follows that Φ is close enough to Φ_0 such that e.g. in the development of the error $\mathcal{E}^{HF}(\Phi) - \mathcal{E}^{HF}(\Phi_0)$ with respect to powers of ϵ: $\mathcal{E}^{HF}(\Phi) - \mathcal{E}^{HF}(\Phi_0) = c_k\epsilon^k + o(\epsilon^k)$ the second term $o(\epsilon^k)$ is indeed smaller than $c_k\epsilon^k$ (due to the asymptotic properties of the decomposition this is certain to happen when ϵ is small enough).

4.1 Error estimators

The a posteriori analysis method presented in this section is connected to the works of Babuška [1], Bernardi [3], Ladevèze [7], Oden [12], Pousin and Rappaz [16], Verfürth [19, 20] and is aimed at giving quantitative indications on the form of the error, through **bilateral** estimates. Even when not all constants are explicitly known, this method may prove interesting when only relative error estimates are needed (as in adaptative procedures) or when the estimator is shown to possess further properties that allow to estimate those constants.

Let us recall (see also (24)) that $U_{\Phi,\Phi_0}\Phi - \Phi_0 \in \mathcal{S}_{\Phi_0} \oplus \Phi_0^{\perp\!\!\!\perp}$ and denote $U_{\Phi,\Phi_0}\Phi - \Phi_0 = S\Phi_0 + W$, $S\Phi_0 \in \mathcal{S}_{\Phi_0}$, $W \in \Phi_0^{\perp\!\!\!\perp}$. Then one can write

$$\begin{aligned}
&\mathcal{E}^{HF}(\Phi) - \mathcal{E}^{HF}(\Phi_0) = \mathcal{E}^{HF}(U_{\Phi,\Phi_0}\Phi) - \mathcal{E}^{HF}(\Phi_0) = \mathcal{E}^{\Phi_0}(\Phi_0 + S\Phi_0 + W) - \mathcal{E}^{\Phi_0}(\Phi_0) \\
&= D\mathcal{E}^{\Phi_0}(\Phi_0)(S\Phi_0 + W) + D^2\mathcal{E}^{\Phi_0}(\Phi_0)(S\Phi_0 + W, S\Phi_0 + W) + O(\epsilon^3) \\
&= 0 + D^2\mathcal{E}^{\Phi_0}(\Phi_0)(W, W) + O(\epsilon^3) = a_{\Phi_0}(W, W) + O(\epsilon^3) \qquad (47)
\end{aligned}$$

where we have used firstly the fact that Φ_0 is the solution of (7) (cf. also remark 5) and secondly the lemma 2 for $(U_{\Phi,\Phi_0}\Phi, \Phi_0) \to (\Psi, \Phi)$. From the continuity of a_{Φ_0} and (46) one concludes that $\|W\|_{\mathcal{H}}^2$ is a third order estimator of the energy error $\mathcal{E}^{HF}(\Phi) - \mathcal{E}^{HF}(\Phi_0)$.

Remark 7 *It easy to see by (29) that* $\|W\|_{\mathcal{H}} = \epsilon + O(\epsilon^2)$.

Unfortunately direct computation of W (and then of $\|W\|_{\mathcal{H}}^2$) assumes knowledge of Φ_0 which is not available. However good approximations of $\|W\|_{\mathcal{H}}^2$ that require only the knowledge of Φ can be found. Indeed, let us set $F = D\mathcal{E}^{HF}$, $\Psi = U_{\Phi,\Phi_0}\Phi$ and study the norm of $F(\Psi)$ in the dual space $\Psi^{\perp\!\!\!\perp *}$

of $\Psi^{\perp\!\!\!\perp}$

$$\|F(\Psi)\|_{\Psi^{\perp\!\!\!\perp *}} = \sup_{\xi\in\Psi^{\perp\!\!\!\perp}} \frac{< D\mathcal{E}^{HF}(\Psi),\xi >}{\|\xi\|_{\mathcal{H}}} = \sup_{\xi\in\Psi^{\perp\!\!\!\perp}} \frac{< D\mathcal{E}^{\Phi_0}(\Psi),\xi >}{\|\xi\|_{\mathcal{H}}}$$
$$= \sup_{\xi\in\Psi^{\perp\!\!\!\perp}} \frac{< D\mathcal{E}^{\Phi_0}(\Psi) - D\mathcal{E}^{\Phi_0}(\Phi_0),\xi >}{\|\xi\|_{\mathcal{H}}} = \sup_{\xi\in\Psi^{\perp\!\!\!\perp}} \frac{D^2\mathcal{E}^{\Phi_0}(\Phi_0)(\Psi-\Phi_0,\xi)}{\|\xi\|_{\mathcal{H}}} + O(\epsilon^2) \quad (48)$$

We show now that we can replace in the above supremum the space $\Psi^{\perp\!\!\!\perp} = (U_{\Phi,\Phi_0}\Phi)^{\perp\!\!\!\perp} = \Phi^{\perp\!\!\!\perp}$ by $\Phi_0^{\perp\!\!\!\perp}$. Let $\xi \in \Psi^{\perp\!\!\!\perp}$ be written as $\xi = M\Phi_0 + \tilde{\xi}$, $\tilde{\xi} \in \Phi_0^{\perp\!\!\!\perp}$. Note that $|M_{ij}| = | < \xi_i, \Phi_{0j} > | = | < \xi_i, \Phi_{0j} - \Psi_j > | \leq \|\xi\|_{(L^2(\mathbb{R}^3))^N} \|\Phi_0 - \Psi\|_{(L^2(\mathbb{R}^3))^N}$ so one can write

$$|\frac{a_{\Phi_0}(\Psi-\Phi_0, M\Phi_0)}{\|\xi\|_{\mathcal{H}}}| \leq \frac{C_{\Phi_0}\|\Psi-\Phi_0\|_{\mathcal{H}}\|\xi\|_{\mathcal{H}}\|\Phi_0-\Psi\|_{(L^2(\mathbb{R}^3))^N}}{\|\xi\|_{\mathcal{H}}} \leq C_{\Phi_0}\epsilon^2, \quad (49)$$

where C_{Φ_0} is the continuity constant of a_{Φ_0}. Since $\frac{\|\tilde{\xi}\|_{\mathcal{H}}}{\|\xi\|_{\mathcal{H}}} = 1 + O(\epsilon)$ one concludes that

$$\|F(\Psi)\|_{\Psi^{\perp\!\!\!\perp *}} = \sup_{\xi\in\Psi^{\perp\!\!\!\perp}} \frac{a_{\Phi_0}(\Psi-\Phi_0,\tilde{\xi})}{\|\tilde{\xi}\|_{\mathcal{H}}} + O(\epsilon^2)$$
$$= \sup_{\tilde{\xi}\in\Psi_0^{\perp\!\!\!\perp}} \frac{a_{\Phi_0}(\Psi-\Phi_0,\tilde{\xi})}{\|\tilde{\xi}\|_{\mathcal{H}}} + O(\epsilon^2) = \sup_{\tilde{\xi}\in\Psi_0^{\perp\!\!\!\perp}} \frac{a_{\Phi_0}(S\Phi_0+W,\tilde{\xi})}{\|\tilde{\xi}\|_{\mathcal{H}}} + O(\epsilon^2)$$
$$= \sup_{\tilde{\xi}\in\Psi_0^{\perp\!\!\!\perp}} \frac{a_{\Phi_0}(W,\tilde{\xi})}{\|\tilde{\xi}\|_{\mathcal{H}}} + O(\epsilon^2) = \|W\|_{\mathcal{H}} + O(\epsilon^2). \quad (50)$$

We have shown above that $\|F(\Psi)\|_{\Psi^{\perp\!\!\!\perp *}}$ is a second order approximation of $\|W\|_{\mathcal{H}}$ and therefore $\|F(\Psi)\|^2_{\Psi^{\perp\!\!\!\perp *}}$ will be a third order estimator of the energy error $\mathcal{E}^{HF}(\Phi) - \mathcal{E}^{HF}(\Phi_0)$. We next prove that $\|F(\Psi)\|_{\Psi^{\perp\!\!\!\perp *}}$ is invariant with respect to the multiplication of Ψ by unitary matrices and therefore equal to $\|F(\Phi)\|_{\Phi^{\perp\!\!\!\perp *}}$, so it can be computed (a posteriori) using only available data (i.e. Φ). Let us now compute for ζ in $\mathcal{H}\cap\mathcal{K}$ the function $F(U\zeta)$. By the definition of F this equals $D\mathcal{E}^{HF}(U\zeta)$ which can be written:

$$D\mathcal{E}^{HF}(U\zeta) = \Big(\mathcal{F}_{U\zeta}((U\zeta)_i)\Big)_{i=1}^N = \Big((-\frac{1}{2}\Delta + V)((U\zeta)_i)\Big)_{i=1}^N +$$
$$\Big((\rho_{U\zeta}\star\frac{1}{|x|})(U\zeta)_i - \int_{\mathbb{R}^3}\frac{\rho_{U\zeta}(x,y)}{|x-y|}(U\zeta)_i(y)dy\Big)_{i=1}^N$$
$$= U\Big((-\frac{1}{2}\Delta+V)(\zeta_i)\Big)_{i=1}^N + U\Big((\rho_\zeta\star\frac{1}{|x|})\zeta_i - \int_{\mathbb{R}^3}\frac{\rho_\zeta(x,y)}{|x-y|}\zeta_i(y)dy\Big)_{i=1}^N, \quad (51)$$

where we have used the invariance property (10). It was therefore proven that

$$F(U\zeta) = UF(\zeta), \forall \zeta \in \mathcal{H} \cap \mathcal{K}, \tag{52}$$

and therefore $\|F(\Psi)\|_{\Psi^{\perp\!\!\!\perp *}} = \|F(U_{\Phi,\Phi_0}\Phi)\|_{\Psi^{\perp\!\!\!\perp *}} = \|F(\Phi)\|_{\Phi^{\perp\!\!\!\perp *}}$ We will summarize the results obtained in this section in the following

Theorem 8 *Let Φ_0 be a minimizer of (13), $\Phi \in \mathcal{H} \cap \mathcal{K}$ a (known) discrete approximation of Φ_0 obtained by a previous computation as described in section 2.1 (15), and denote $\epsilon = \|U_{\Phi_0,\Phi}\Phi_0 - \Phi\|_{\mathcal{H}}$ the quotient distance between Φ and Φ_0. Then, under the assumption (45),*

$$\|D\mathcal{E}^{HF}(\Phi)\|_{\Phi^{\perp\!\!\!\perp *}} = \epsilon + O(\epsilon^2). \tag{53}$$

Moreover there exists constants c_1, c_2 depending only on Φ_0 such that

$$c_1\|D\mathcal{E}^{HF}(\Phi)\|^2_{\Phi^{\perp\!\!\!\perp *}} + O(\epsilon^3) \leq \mathcal{E}^{HF}(\Phi) - \mathcal{E}^{HF}(\Phi_0) \leq c_2\|D\mathcal{E}^{HF}(\Phi)\|^2_{\Phi^{\perp\!\!\!\perp *}} + O(\epsilon^3). \tag{54}$$

4.2 Explicit bounds for the Hartree-Fock energy and convergence acceleration

The purpose of this section is to propose methods to find explicit bounds for the Hartree-Fock energy. The method belongs to the more general paradigm [11, 13, 14, 15] of definition of explicit lower and upper bounds for outputs depending on the solution of a partial differential equation. The output of interest will be taken to be the Hartree-Fock energy ; this choice will be seen (cf. thm. 11 and remark 10) to posses particularities that in fact allow to design an improvement of the solution itself, although this is not expected to be the case for general outputs.

We will begin this section with some remarks on the coercivity properties of the bilinear forms a_{Φ_0} and a_Φ.

Lemma 9 *Under the hypothesis (45) there exists a constant $\gamma > 0$ depending only on Φ_0 such that for any $U \in \mathcal{U}(N)$ the bilinear form $a_{U\Phi_0}$ is coercive on $(U\Phi_0)^{\perp\!\!\!\perp} = \Phi_0^{\perp\!\!\!\perp}$ with coercivity constant γ.*

Proof: Note that for any $\Psi_1 \in \mathcal{H}\cap\mathcal{K}$, $\Psi_2 \in \mathcal{H}, U \in \mathcal{U}(N)$: $a_{U\Psi_1}(U\Psi_2, U\Psi_2) = a_{\Psi_1}(\Psi_2, \Psi_2)$, so by (45) and lemma 6 we obtain the conclusion. □

Lemma 10 *Under the assumption (45) there exists a constant $\eta > 0$ depending only on Φ_0 such that for all $\Phi \in \mathcal{H} \cap \mathcal{K}$ with $\|\Phi - \Phi_0\|_{\mathcal{H}} \leq \eta$ the bilinear form a_Φ is coercive on $\Phi^{\perp\!\!\!\perp}$ with a coercivity constant depending only of Φ_0.*

Proof: Let $\xi \in \Phi^{\perp\!\!\!\perp}$, $\|\xi\|_{\mathcal{H}} \leq 1$ be written as $\xi = M\Phi_0 + \tilde{\xi}$, $\tilde{\xi} \in \Phi_0^{\perp\!\!\!\perp}$. We will generically denote by C various constants depending only on Φ_0. Recall that $|M_{ij}| \leq \|\xi\|_{(L^2(\mathbb{R}^3))^N} \|\Phi_0 - \Psi\|_{(L^2(\mathbb{R}^3))^N}$, so for $\|\Phi_0 - \Psi\|_{\mathcal{H}}$ small enough

$$a_\Phi(\xi,\xi) = a_\Phi(\tilde{\xi} + M\Phi_0, \tilde{\xi} + M\Phi_0) \geq a_\Phi(\tilde{\xi},\tilde{\xi}) - C\|\xi\|_{\mathcal{H}}\|\tilde{\xi}\|_{\mathcal{H}}\|\Phi_0 - \Psi\|_{\mathcal{H}}.$$

But for $\|\Phi_0 - \Psi\|_{\mathcal{H}}$ small enough we can also write

$$\|\xi\|_{\mathcal{H}}\|\tilde{\xi}\|_{\mathcal{H}}\|\Phi_0 - \Psi\|_{\mathcal{H}} \leq \|\xi\|_{\mathcal{H}}(\|\xi\|_{\mathcal{H}} + \|\xi\|_{\mathcal{H}}\|\Phi_0 - \Psi\|_{\mathcal{H}})\|\Phi_0 - \Psi\|_{\mathcal{H}} \leq C\|\xi\|_{\mathcal{H}}^2\|\Phi_0 - \Psi\|_{\mathcal{H}}.$$

Since $|\Lambda_{ij} - \Lambda_{ij}^0| \leq C\|\Phi - \Phi_0\|_{\mathcal{H}}$ it follows that $|a_\Phi(\tilde{\xi},\tilde{\xi}) - a_{\Phi_0}(\tilde{\xi},\tilde{\xi})| \leq C\|\tilde{\xi}\|_{\mathcal{H}}^2\|\Phi_0 - \Psi\|_{\mathcal{H}}$ so in fact

$$\begin{aligned} a_\Phi(\xi,\xi) &\geq a_{\Phi_0}(\tilde{\xi},\tilde{\xi}) - C(\|\tilde{\xi}\|_{\mathcal{H}}^2 + \|\xi\|_{\mathcal{H}}^2)\|\Phi_0 - \Psi\|_{\mathcal{H}} \\ &\geq \gamma\|\tilde{\xi}\|_{\mathcal{H}}^2 - C(\|\tilde{\xi}\|_{\mathcal{H}}^2 + \|\xi\|_{\mathcal{H}}^2)\|\Phi_0 - \Psi\|_{\mathcal{H}}. \end{aligned}$$

It suffices now to use a last time $|\|\xi\|_{\mathcal{H}} - \|\tilde{\xi}\|_{\mathcal{H}}| \leq \|\xi\|_{\mathcal{H}}\|\Phi_0 - \Psi\|_{\mathcal{H}}$ to conclude. □

We will begin in the following the presentation of the construction of (lower) bounds for the Hartree-Fock energy. As it was seen in lemma 9, under the assumption (45) we have uniform coercivity properties for bilinear forms a_{Φ_0} with respect to the multiplication of Φ_0 by unitary matrices $U \in \mathcal{U}(N)$; for this reason we can replace Φ_0 with any $U\Phi_0$ that fits better our needs; we will therefore suppose in agreement with lemma 1 that Φ_0 is such that $\Phi_0 - \Phi = S\Phi + W \in \mathcal{S}_\Phi \oplus \Phi^{\perp\!\!\!\perp}$.

The construction of (lower) bounds for the Hartree-Fock energy is based on the following development:

$$\begin{aligned} \mathcal{E}^{HF}(\Phi_0) - \mathcal{E}^{HF}(\Phi) &= \mathcal{E}^\Phi(\Phi_0) - \mathcal{E}^\Phi(\Phi) = \mathcal{E}^\Phi(\Phi + S\Phi + W) - \mathcal{E}^\Phi(\Phi) \\ &= D\mathcal{E}^\Phi(\Phi)(S\Phi + W) + \frac{1}{2}D^2\mathcal{E}^\Phi(\Phi)(S\Phi + W, S\Phi + W) + O(\epsilon^3) \end{aligned}$$

Note first that by the properties of Φ as described in section 2.1 eq. (15) $D\mathcal{E}^\Phi(\Phi)$ is null on the dual space of the discretization space so in particular $D\mathcal{E}^\Phi(\Phi)(S\Phi) = 0$; recall also the fact that $S\Phi$ is of order ϵ^2 and W of order ϵ to obtain

$$\mathcal{E}^{HF}(\Phi_0) - \mathcal{E}^{HF}(\Phi) = D\mathcal{E}^\Phi(\Phi)(W) + \frac{1}{2}D^2\mathcal{E}^\Phi(\Phi)(W,W) + O(\epsilon^3) \qquad (55)$$

Consider now the problem: find the *reconstructed error* $\hat{W} \in \Phi^{\perp\!\!\!\perp}$ such that

$$D^2\mathcal{E}^\Phi(\Phi)(\hat{W},\Psi) + D\mathcal{E}^\Phi(\Phi)(\Psi) = 0,\ \forall \Psi \in \Phi^{\perp\!\!\!\perp}. \qquad (56)$$

By the coercivity of a_Φ it follows that (56) has an unique solution $\hat{W} \in \Phi^{\perp\!\!\!\perp}$.

Remark 8 *Note that in order to compute $\hat{W}$ one solves a* **direct** *(i.e. not eigenvalue) problem on the solution space ; moreover all operators involved depend only on Φ.*

Using the definition of $\hat{W}$ one can rewrite (55):

$$\mathcal{E}^{HF}(\Phi_0) = \mathcal{E}^{HF}(\Phi) - D^2\mathcal{E}^{\Phi}(\Phi)(\hat{W}, W) + \frac{1}{2}D^2\mathcal{E}^{\Phi}(\Phi)(W, W) + O(\epsilon^3)$$
$$= \mathcal{E}^{HF}(\Phi) - \frac{1}{2}D^2\mathcal{E}^{\Phi}(\Phi)(\hat{W}, \hat{W}) + \frac{1}{2}D^2\mathcal{E}^{\Phi}(\Phi)(W - \hat{W}, W - \hat{W}) + O(\epsilon^3). \quad (57)$$

But since a_Φ is positive on $\Phi^{\perp\!\!\!\perp}$ it follows that $\frac{1}{2}D^2\mathcal{E}^{\Phi}(\Phi)(W - \hat{W}, W - \hat{W}) \geq 0$ so in fact we obtain an **explicit lower bound on the Hartree-Fock energy**:

$$\mathcal{E}^{HF}(\Phi_0) \geq \mathcal{E}^{HF}(\Phi) - \frac{1}{2}D^2\mathcal{E}^{\Phi}(\Phi)(\hat{W}, \hat{W}) + O(\epsilon^3), \quad (58)$$

which together with the inequality $\mathcal{E}^{HF}(\Phi_0) \leq \mathcal{E}^{HF}(\Phi)$ gives an interval for the **exact** value of the Hartree-Fock energy.

Remark 9 *A natural question is to study the order in ϵ of the length of the confidence interval found above. Let us recall that the error in energy is of order ϵ^2 ; we will prove that this interval is optimal in a sense that its length is also of order ϵ^2 ; indeed the distance between the upper and lower bound is $\frac{1}{2}D^2\mathcal{E}^{\Phi}(\Phi)(\hat{W}, \hat{W}) + O(\epsilon^3)$ which is equivalent to $\|\hat{W}\|_{\mathcal{H}}$; all that remains to be proven is that $\|\hat{W}\|_{\mathcal{H}} \leq C\epsilon$ (with a constant not depending on Φ_0). Indeed:*

$$\|\hat{W}\|_{\mathcal{H}} \leq C\|D\mathcal{E}^{\Phi}(\Phi)\|_{\Phi^{\perp\!\!\!\perp *}} \leq C\|D\mathcal{E}^{\Phi}(\Phi) - D\mathcal{E}^{\Phi}(\Phi_0)\|_{\Phi^{\perp\!\!\!\perp *}}$$
$$+C\|D\mathcal{E}^{\Phi}(\Phi_0) - D\mathcal{E}^{\Phi_0}(\Phi_0)\|_{\Phi^{\perp\!\!\!\perp *}} \leq C\epsilon$$

where we have used the fact that $D\mathcal{E}^{\Phi_0}(\Phi_0)$ is null on $\Phi_0^{\perp\!\!\!\perp}$.

The nomination of $\hat{W}$ as "reconstructed error" is best explained by the following property:

$$\hat{W} = W + O(\epsilon^2). \quad (59)$$

In order to prove (59) we will prove that W has the following property:

$$|D^2\mathcal{E}^{\Phi}(\Phi)(W, \Psi) + D\mathcal{E}^{\Phi}(\Phi)(\Psi)| \leq C\epsilon^2, \ \forall \Psi \in \Phi^{\perp\!\!\!\perp}, \ \|\Psi\|_{\mathcal{H}} = 1. \quad (60)$$

with a constant C independent of Φ, Ψ. Suppose (60) is true then jointly with (56) one obtains:

$$|D^2\mathcal{E}^{\Phi}(\Phi)(W - \hat{W}, \Psi)| \leq C\epsilon^2, \forall \Psi \in \Phi^{\perp\!\!\!\perp}, \ \|\Psi\|_{\mathcal{H}} = 1.$$

Let $\Psi = \frac{W-\hat{W}}{\|W-\hat{W}\|_{\mathcal{H}}}$; from the coercivity of $a_\Phi = D^2\mathcal{E}^\Phi(\Phi)$ we deduce:

$$\frac{1}{\|W-\hat{W}\|_{\mathcal{H}}} \cdot c\|W-\hat{W}\|_{\mathcal{H}}^2 \leq C\epsilon^2,$$

and (59) follows.

Recall that, from lemma 2, $\|\Phi_0 - \Phi - W\|$ is of order ϵ^2. In order to prove (60) it is thus sufficient to prove it for $\Phi_0 - \Phi$ instead of W: let us write

$$D\mathcal{E}^\Phi(\Phi)(\Psi) = D\mathcal{E}^\Phi(\Phi_0)(\Psi) + D^2\mathcal{E}^\Phi(\Phi_0)(\Phi-\Phi_0,\Psi) + O(\epsilon^2).$$

Besides we have

$$|D^2\mathcal{E}^\Phi(\Phi_0)(\Phi-\Phi_0,\Psi) - D^2\mathcal{E}^\Phi(\Phi)(\Phi-\Phi_0,\Psi)| \leq C\epsilon^2\|\Psi\|_{\mathcal{H}},$$

(with a constant C depending only of Φ_0), so

$$D\mathcal{E}^\Phi(\Phi)(\Psi) = D\mathcal{E}^\Phi(\Phi_0)(\Psi) + D^2\mathcal{E}^\Phi(\Phi)(\Phi-\Phi_0,\Psi) + O(\epsilon^2)$$

and therefore

$$D^2\mathcal{E}^\Phi(\Phi)(\Phi_0-\Phi,\Psi) + D\mathcal{E}^\Phi(\Phi)(\Psi) = D\mathcal{E}^\Phi(\Phi_0)(\Psi) + O(\epsilon^2).$$

It suffices now to prove that $D\mathcal{E}^\Phi(\Phi_0)(\Psi) = O(\epsilon^2)$. By the definition of $\mathcal{E}^\Phi$,

$$\begin{aligned} D\mathcal{E}^\Phi(\Phi_0)(\Psi) &= D\mathcal{E}^{\Phi_0}(\Phi_0)(\Psi) + \textstyle\sum_{i,j=1}^N(\Lambda_{ij}-\Lambda_{ij}^0)DF_{ij}(\Phi_0)(\Psi) \\ &= 0 + \textstyle\sum_{i,j=1}^N(\Lambda_{ij}-\Lambda_{ij}^0)DF_{ij}(\Phi_0)(\Psi). \end{aligned}$$

Note firstly that $\Lambda_{ij} - \Lambda_{ij}^0 \leq C\epsilon$ (C depending only of Φ_0). Moreover

$$\begin{aligned} DF_{ij}(\Phi_0)(\Psi) &= <\Phi_{0i},\Psi_j> + <\Phi_{0j},\Psi_i> \\ &= <\Phi_{0i}-\Phi_i,\Psi_j> + <\Phi_{0j}-\Phi_j,\Psi_i> \end{aligned}$$

thus $|DF_{ij}(\Phi_0)(\Psi)|$ can be upper bounded by $C\epsilon$ (we used the fact that $\Psi \in \Phi^{\perp\!\!\!\perp}$), which concludes the proof of (59).

Combining (57) and (59) we can give a better version of (58):

$$\mathcal{E}^{HF}(\Phi_0) = \mathcal{E}^{HF}(\Phi) - \frac{1}{2}D^2\mathcal{E}^\Phi(\Phi)(\hat{W},\hat{W}) + O(\epsilon^3), \tag{61}$$

so instead of a lower bound we have obtained an **improvement** of the Hartree-Fock energy ; note that this improvement is of a **strictly** higher order in ϵ since the best approximation known before the computation of $\hat{W}$ was $\mathcal{E}^{HF}(\Phi)$ which is exact to the order ϵ^2.

Although (61) may represent in itself the conclusion of the a posteriori analysis, further progress is possible. To this end note that an improvement for the wavefunction Φ has also been found, namely $\tilde{\Phi} = \Phi + \hat{W}$. However we cannot propose $\tilde{\Phi}$ as a legitimate solution of (7) since it is not certain to be in $\mathcal{K}$. We will see in the following that it is possible to find a correction to add to $\Phi + \hat{W}$ which not only gives an admissible solution of (7) but also **improves with another order** the approximation (61) of the Hartree-Fock energy $\mathcal{E}^{HF}(\Phi_0)$.

The principle is to add to $\tilde{\Phi}$ a term $\hat{S}\Phi$ ($\hat{S} \in \mathbb{R}^{N\times N}$, $\hat{S} = \hat{S}^t$) such that $\hat{\Phi} = \tilde{\Phi} + \hat{S}\Phi = \Phi + \hat{W} + \hat{S}\Phi \in \mathcal{K}$. We will also see in the process that $\hat{S}\Phi$ can be interpreted as a "reconstruction" of symmetrical part $S\Phi$ of the error $\Phi_0 - \Phi$.

Consider the equality $\Phi_0 = \Phi + W + S\Phi$. Since both Φ_0 and Φ are in $\mathcal{K}$ we can write

$$\begin{aligned}\delta_{ij} &=< \Phi_{0i}, \Phi_{0j} >=< \Phi_i + \sum_{k=1}^{N} S_{ik}\Phi_k + W_i, \Phi_j + \sum_{l=1}^{N} S_{jl}\Phi_l + W_j > \\ &= \delta_{ij} + < W_i, W_j > + \sum_{k=1}^{N} S_{ik}\delta_{kj} + \sum_{k=1}^{N} S_{jl}\delta_{il} + O(\epsilon^4) \qquad (62)\end{aligned}$$

because we know that $S_{ij} = O(\epsilon^2)$. We obtain

$$0 =< W_i, W_j > + S_{ij} + S_{ji} + O(\epsilon^4) =< \hat{W}_i, \hat{W}_j > + S_{ij} + S_{ji} + O(\epsilon^3)$$

so denoting $\tilde{S}_{ij} = -\frac{1}{2} < \hat{W}_i, \hat{W}_j >$, we obtain that $\tilde{S}\Phi$ is a order ϵ^3 approximation of $S\Phi$: $\tilde{S}\Phi = S\Phi + O(\epsilon^3)$. Note that by remark 8 that the computation of $\tilde{S}$ requires knowledge of Φ only.

We will prove in the following that having an approximation $\hat{W}$ of W to the order ϵ^2 and an approximation $\tilde{S}$ of S to the order ϵ^3 is enough to have an approximation of the Hartree-Fock energy to the order ϵ^4. Indeed, write

$$\begin{aligned}&\mathcal{E}^{HF}(\Phi_0) - \mathcal{E}^{HF}(\Phi) = \mathcal{E}^{\Phi}(\Phi_0) - \mathcal{E}^{\Phi}(\Phi) = \mathcal{E}^{\Phi}(\Phi + S\Phi + W) - \mathcal{E}^{\Phi}(\Phi) = \\ &D\mathcal{E}^{\Phi}(\Phi)(S\Phi + W) + \frac{1}{2}D^2\mathcal{E}^{\Phi}(\Phi)(S\Phi + W, S\Phi + W) \\ &+\frac{1}{3!}D^3\mathcal{E}^{\Phi}(\Phi)(S\Phi + W, S\Phi + W, S\Phi + W) + O(\epsilon^4) \\ &= D\mathcal{E}^{\Phi}(\Phi)(W) + \frac{1}{2}D^2\mathcal{E}^{\Phi}(\Phi)(W, W) + D^2\mathcal{E}^{\Phi}(\Phi)(S\Phi, W) \\ &+\frac{1}{3!}D^3\mathcal{E}^{\Phi}(\Phi)(W, W, W) + O(\epsilon^4)\end{aligned}$$

$$= -\frac{1}{2}D^2\mathcal{E}^{\Phi}(\Phi)(\hat{W},\hat{W}) + \frac{1}{2}D^2\mathcal{E}^{\Phi}(\Phi)(W-\hat{W},W-\hat{W}) +$$
$$D^2\mathcal{E}^{\Phi}(\Phi)(\tilde{S}\Phi,\hat{W}) + \frac{1}{3!}D^3\mathcal{E}^{\Phi}(\Phi)(\hat{W},\hat{W},\hat{W}) + O(\epsilon^4)$$
$$= -\frac{1}{2}D^2\mathcal{E}^{\Phi}(\Phi)(\hat{W},\hat{W}) + D^2\mathcal{E}^{\Phi}(\Phi)(\tilde{S}\Phi,\hat{W}) +$$
$$\frac{1}{3!}D^3\mathcal{E}^{\Phi}(\Phi)(\hat{W},\hat{W},\hat{W}) + O(\epsilon^4),$$

so we have obtained

$$\mathcal{E}^{HF}(\Phi_0) = \mathcal{E}^{HF}(\Phi) - \frac{1}{2}D^2\mathcal{E}^{\Phi}(\Phi)(\hat{W},\hat{W}) + D^2\mathcal{E}^{\Phi}(\Phi)(\tilde{S}\Phi,\hat{W}) +$$
$$\frac{1}{3!}D^3\mathcal{E}^{\Phi}(\Phi)(\hat{W},\hat{W},\hat{W}) + O(\epsilon^4). \tag{63}$$

where all terms involved in the right hand side can be computed from Φ.

One problem remains though, our best approximation for the solution Φ_0, namely $\tilde{\tilde{\Phi}} = \Phi + \hat{W} + \tilde{S}\Phi$ is not certain to be in $\mathcal{K}$; in fact it can be proved that there exists and $\hat{S}$ that depends only of Φ that has the property $\hat{S}\Phi = S\Phi + O(\epsilon^3)$ and such that $\hat{\Phi} = \Phi + \hat{W} + \hat{S}\Phi \in \mathcal{K}$. Moreover, using the above arguments, we will also have $\mathcal{E}^{HF}(\Phi_0) = \mathcal{E}^{HF}(\hat{\Phi}) + O(\epsilon^4)$. The existence and properties of $\hat{S}$ follows by considering as in (62) the equations satisfied by $(\hat{S} - \tilde{S})_{ij}$.

We will summarize the results obtained in this section in the following theorem:

Theorem 11 *Let Φ_0 be a minimizer of (13), $\Phi \in \mathcal{H} \cap \mathcal{K}$ a (known) discrete approximation of Φ_0 obtained by a previous computation as described in section 2.1 (15). Then, under the assumption (45), there exists an $\eta > 0$ such that for any $\Phi \in \mathcal{H} \cap \mathcal{K}$ with $\|U_{\Phi_0,\Phi}\Phi_0 - \Phi\| \leq \eta$ there exists $\hat{W} \in \Phi^{\perp\!\!\!\perp}$ and $\hat{S}\Phi \in \mathcal{S}_\Phi$ whose computation requires only knowledge of Φ such that $\hat{\Phi} = \Phi + \hat{S}\Phi + \hat{W} \in \mathcal{H} \cap \mathcal{K}$ has the following properties:*

$$\|\hat{\Phi} - \Phi\|_{\mathcal{H}} \leq c_1\|\Phi - \Phi\|^2_{\mathcal{H}}, \tag{64}$$
$$|\mathcal{E}^{HF}(\Phi_0) - \mathcal{E}^{HF}(\hat{\Phi})| \leq c_2|\mathcal{E}^{HF}(\Phi_0) - \mathcal{E}^{HF}(\Phi)|^2. \tag{65}$$

with constants c_1,c_2 depending only of Φ_0.

Remark 10 *The initial bounds found for the Hartree-Fock energy have proven to be* **exact** *at the order ϵ^4 due to the minimization setting which is particular. We do not expect this convergence acceleration to hold true when bounds on more general functionals (polarizability, electronic density, ...) depending*

on the wavefunction are searched for; this functional (the energy) is particular in that the wavefunction is characterized by an Euler-Lagrange equation associated associated to it, giving therefore a kind of entangled feedback: energy can be computed from the wavefunction but at the same time the (ground state) wavefunction depends of the energy (that it minimizes).

Remark 11 *It follows from the minimization properties of* $\mathcal{E}^{HF}(\Phi_0)$ *that* $\mathcal{E}^{HF}(\hat{\Phi}) \geq \mathcal{E}^{HF}(\Phi_0)$ *; however, if one is interested in a (order* ϵ^2*) lower bound for* $\mathcal{E}^{HF}(\Phi_0)$ *it suffices to consider the symmetrical* $2 \cdot \mathcal{E}^{HF}(\hat{\Phi}) - \mathcal{E}^{HF}(\Phi)$ *of* $\mathcal{E}^{HF}(\Phi)$ *with respect to* $\mathcal{E}^{HF}(\hat{\Phi})$*.*

Remark 12 *The approach described in this section can be developed under more general assumptions than (45). Denote by* X_{Φ_0} *the closed subspace of* $\Phi_0^{\perp\!\!\!\perp}$ *where (38) holds so that, in agreement with lemma 6* a_{Φ_0} *is coercive on* X_{Φ_0} *; using the same arguments as in lemma 10 one proves for* $\|\Phi_0 - \Phi\|_{\mathcal{H}}$ *small enough coercivity for* a_Φ *on* $X_{\Phi_0} \cap \Phi^{\perp\!\!\!\perp}$ *; this shows that the problem (56) has an unique solution on* $X_{\Phi_0} \cap \Phi^{\perp\!\!\!\perp}$ *and this solution is then shown to posses the same property (59) as* $\hat{W}$*. A "reconstructed symmetrical" part is then computed by the same method as above and we obtain thus an improvement for the energy and for the wavefunction. The only computational impediment to this program is that one cannot really identify the space* $X_{\Phi_0} \cap \Phi^{\perp\!\!\!\perp}$ *where problem (56) is to be solved ; one chooses then the largest subspace of* $\Phi^{\perp\!\!\!\perp}$ *where* a_Φ *is positive (therefore coercive), which will contain* $X_{\Phi_0} \cap \Phi^{\perp\!\!\!\perp}$*, and proves that the solution of (56) on this space is an order* ϵ^2 *approximation of the solution of (56) on* $X_{\Phi_0} \cap \Phi^{\perp\!\!\!\perp}$*. In practice (cf. section 5) there was no need to implement this procedure as (45) seems to be satisfied.*

Remark 13 *Numerical computation of* $\hat{W}$ *involves the resolution of equation (56) discrete subspace* $\Phi_\delta^{\perp\!\!\!\perp}$ *of* $\Phi^{\perp\!\!\!\perp}$*; the corresponding solution* $\hat{W}_\delta$ *will be an approximation of* $\hat{W}$ *which converges to* $\hat{W}$ *when the discretization parameter* δ *is such that* $\Phi_\delta^{\perp\!\!\!\perp}$ *converges to space* $\Phi^{\perp\!\!\!\perp}$*.*

5 Numerical simulations

The theory presented in the previous sections was tested in two categories of numerical experiments.

In the experiments of the first category we checked on simple cases (hydrogen molecule, helium) that the methodology proposed above is coherent with available results when the problem (56) that provides $\hat{W}$ is solved on a very fine discretization of $\mathcal{H}$.

In a second stage more complex molecules were studied and the method was implemented in a Hartree-Fock quantum chemistry code.

Before presenting the results let us remark that the partial differential equation (PDE) (56) is, for N large, very difficult to discretize with classical tools from the PDE equations (finite elements, finite volumes, ...) due to the high dimensionality of the linear spaces involved. Moreover a good discretization has also to take into account some specific quantum chemical effects as the singularities of the electronic wavefunction around nuclei; in conclusion, only very small quantum systems are thus available for study using classical tools in solving PDEs ; such systems are for example the hydrogen molecule (H_2) and the helium atom (He).

For all the numerical experiments we placed ourselves into the *Restricted (closed) shell Hartree-Fock* (Lewis electron pair) approximation that states that when the number of electrons in a molecule is even, one can group together the electrons 2 by 2; the two electrons in each such pair will share a common spatial wavefunction but will have opposite spin. Within this approximation, for a bi-electronic system as the hydrogen molecule or Helium atom, the search of the electronic wavefunction of the system reduces to the search of a function u of 3 variables such that

$$-\Delta u + Vu + \left(|u|^2 \star \frac{1}{|x|}\right) + \lambda u = 0 \quad in\ \mathbb{R}^3. \tag{66}$$

The space to be discretized is therefore $\mathbb{R}^3$; in fact using classical localization arguments it can be reduced to a brick of $\mathbb{R}^3$ that contains the nuclei of the system ; in the case of the Helium atom this brick was taken to be a cube centered around the nucleus.

We will present in the following the results obtained for the Helium atom; each axis of a cube centered in the nucleus mentioned above was discretized with the same number of points that varied between 60 and 120 depending on the singularities of the initial solutions considered; precise results were obtained for about 100 points per dimension and corresponding vectors of size $100^3 = 10^6$.

Several initial approximations Φ_i, $i = 2, 3, 4, 5, 6$ of the electronic wavefunction were considered; each correspond to a quantum chemical computation that used specific quantum basis sets denominated as STOnG, $n = 2, 3, 4, 5, 6$; the larger the parameter n, the finer the basis used; in each case the linear problem (56) was solved on the chosen grid as indicated in Remark 13 and then the symmetric part of the error was reconstructed as indicated in previous section. In order to solve (56) an iterative algorithm was employed, the matrix associated to $D^2\mathcal{E}^\Phi(\Phi)(\cdot,\cdot)$ (typically $10^6 \times 10^6$) being too large for direct inversion; finally in order to take advantage of the

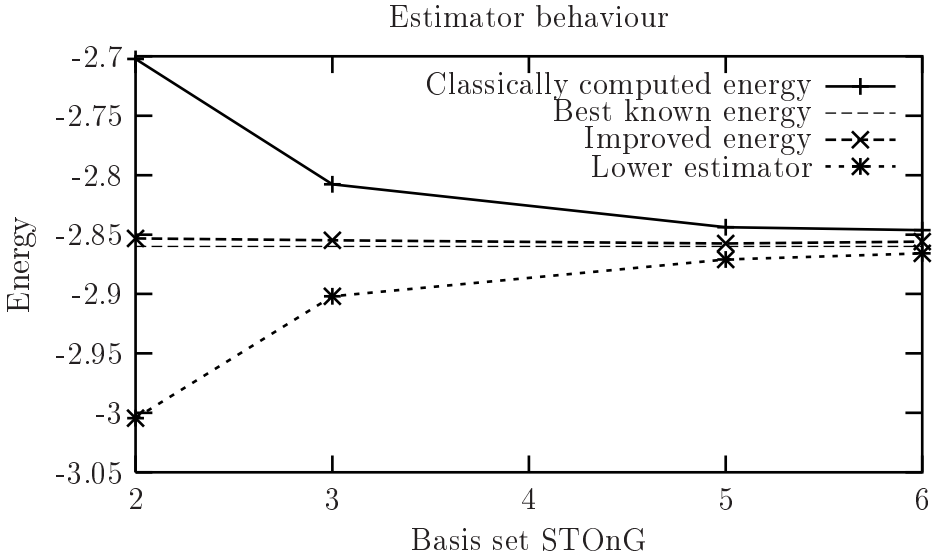


Figure 1: A posteriori improvements for the energy obtained with the basis sets STOnG.

tensor-product-like discretization the computation of convolution products was done by means of fast Fourier transforms.

The figure 1 shows the energy of the initial wavefunction Φ ("Classically computed energy"), the best known approximation of the energy Helium atom, the improved energy obtained as in thm. 11 and then the order ϵ^2 lower bound as described in remark 11 ("Lower estimator") ; agreement with the theoretical results is obtained.

Motivated by the success of the first series of experiments, a second approach towards testing the theoretical results was undertaken; this time the molecules considered were larger, as is for instance the case of the carbyn molecule $Cr(CO)_4ClCH$, with 52 electron pairs (104 electrons) ; the model chosen was again the Restricted Hartree Fock model; in this setting the energy to minimize is

$$\begin{aligned}\mathcal{E}^{HF}(\Phi_1,...,\Phi_N) &= \sum_{i=1}^{N}\int_{\mathbb{R}^3}\left(|\nabla\Phi_i|^2 + V\ |\Phi_i|^2\right) + \iint_{\mathbb{R}^3\times\mathbb{R}^3}\frac{\rho_\Phi(x)\rho_\Phi(y)}{|x-y|}dxdy \\ &\quad -\frac{1}{2}\iint_{\mathbb{R}^3\times\mathbb{R}^3}\frac{|\rho_\Phi(x,y)|^2}{|x-y|}dxdy \qquad (67)\end{aligned}$$

with the same formal definitions (cf. Eq. (4 , 5) for $\rho_\Phi(x)$, $\rho_\Phi(x,y)$). The Euler-Lagrange equations associated to the minimization of $\mathcal{E}^{HF}$ on $\mathcal{H}\cap\mathcal{K}$ are completely similar to (11) (only some multiplicative factors before the

last two terms in (8) are changed).

Due to concerns about computation complexity and efficiency and also for realistic verification we have chosen to implement the a posteriori procedure (and the "convergence acceleration" version) in a quantum computational chemistry code named Asterix [5, 17, 21]. As a consequence, the evaluation of the performances of the a posteriori procedure is to be compared with the performances of quantum chemistry ab initio codes. An introduction to the complexity of the algorithms used is given in the following.

One particularity of computational quantum chemistry codes (especially at the Hartree-Fock level) is the presence of very special Galerkin discretization basis. This basis contains in general functions on $\mathbb{R}^3$ which are centered in the nuclei of the system and are sum of Gaussian type functions; it is beyond the scope of this paper to give a rigorous presentation of the basis involved, let us just say that they all satisfy an important requirement: for any elements h_α, h_β, h_γ and h_δ of the discretization basis, the quantity

$$(\alpha\beta||\gamma\delta) = \iint_{\mathbb{R}^3\times\mathbb{R}^3} \frac{h_\alpha(x)h_\beta(x)h_\gamma(y)h_\delta(y)}{|x-y|}dxdy \tag{68}$$

can be computed in $O(1)$ time[3].

Let us denote by n the number of basis functions used when computing the Hartree-Fock energy of a molecule with N electron pairs ($2N$ electrons); in general n is taken to depend linearly on N.

In order to solve the nonlinear eigenvalue equations (11) iterative (also named selfconsistent - SCF) algorithms are used. The most straightforward idea is to start from an initial guess Φ^1 for the wavefunction and then, for any $i \geq 1$, construct the Fock operator $\mathcal{F}^i = \mathcal{F}_{\Phi^i}$ associated to Φ^i, diagonalize $\mathcal{F}^i$ and take its first N eigenfunctions as the next guess Φ^{i+1} for the wavefunction (Roothaan algorithm) ; ideally this fixed point algorithm will converge and the solution will be the solution of equations (11). Numerical reality does not however always validate this choice, we refer to [4] for a mathematical description of the phenomena involved. In order to cure the convergence deficiencies, various other methods have been proposed [4]: the basic level shift method, DIIS,...

During the SCF resolution of the Hartree-Fock equations, the most time consuming part is the construction of the Fock operator $\mathcal{F}_{\Phi^i}$; we will see in the following that this is an $O(N^4)$ operation, one order of magnitude larger than the diagonalization of the Fock operator itself (under assumption that

[3]Using the fact that the product of two gaussian functions is also a gaussian function, analytical formulas may be provided for the computation of the integral (68).

n is linear in N). Let

$$B = \{h_\alpha; \alpha = 1, ..., n\} \tag{69}$$

be a discretization basis and $\Phi = (\sum_{\alpha=1}^{n} \Phi_{i\alpha} h_\alpha)_{i=1}^{N}$ be an element in the discretized space $X = (span(B))^N$ and also in $\mathcal{K}$. The matrix of the operators $-\Delta$ and V take $O(N^2)$ time to compute, supposing that finite constant time to compute $\int_{\mathbb{R}^3} \nabla h_\alpha \cdot \nabla h_\beta$ and $\int_{\mathbb{R}^3} V h_\alpha h_\beta$ is needed. The situation is very different for the matrices of the operators $(\rho_\Phi \star \frac{1}{|x|})$ and $\psi \mapsto \int_{\mathbb{R}^3} \frac{\rho_\Phi(x,y)}{|x-y|}\psi(y)dy$. Let us take for instance the last operator. To compute the matrix of this operator it is necessary to compute for all h_β, $h_\gamma \in B$:

$$\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho_\Phi(x,y)h_\beta(y)}{|x-y|} dy h_\gamma(x)dx = \sum_{i=1}^{N} \iint_{\mathbb{R}^3\times\mathbb{R}^3} \frac{\sum_{\alpha=1}^{n} \Phi_{i\alpha} h_\alpha(x)}{|x-y|}$$
$$\cdot \sum_{\delta=1}^{n} \Phi_{i\delta} h_\delta(y) h_\beta(y) h_\gamma(x) dxdy = \sum_{i=1}^{N}\sum_{\alpha=1}^{n}\sum_{\delta=1}^{n} \Phi_{i\alpha}\Phi_{i\delta}(\alpha\gamma||\beta\delta).$$

Even if formally this is a $O(N^5)$ computation (summation over three indices for each of the N^2 required terms), it is easy to see that precomputing in $O(N^3)$ for any $\alpha, \delta = 1, ..., n$: $D^\Phi{}_{\alpha,\delta} = \sum_{i=1}^{N} \Phi_{i\alpha}\Phi_{i\delta}$ the computation reduces to order N^4 ; unfortunately no further reductions are possible so the matrix of the operator $\psi \mapsto \int_{\mathbb{R}^3} \frac{\rho_\Phi(x,y)}{|x-y|}\psi(y)dy$ is obtained by computing $(D^\Phi{}_{\alpha,\delta})_{\alpha,\delta=1}^{n}$, then obtain in $O(N^4)$ the desired matrix $\left(\sum_{\alpha,\delta=1}^{n} D^\Phi{}_{\alpha,\delta}(\alpha\gamma||\beta\delta)\right)_{\beta,\gamma=1}^{n}$. The computational complexity of a SCF Hartree-Fock computation is therefore $N_I \star N^4$ where N_I is the number of iterations required by the SCF method, usually in the range $10 - 50$. We shall apply the bound procedure and the improvement strategy to qualify the (known) solution obtained from the previous iterative procedure far from convergence.

Let us now present the complexity issues related to the computation of the reconstructed error $\hat{W}$. The problem (56) is approximated on a product of N dimensional spaces so the solution will be an $n \times N$ vector (considering the **same** discretization X of $\mathcal{H}$ as the one used to solve the Hartree-Fock problem)[4]; we will denote by P the matrix of the projector from X to $X \cap$

[4]Since only one discretization is used for the entire computation, the bounds thus obtained refer to the energy of the solution of the Hartree-Fock problem on *discrete* space X. When the discretization X is fine enough, one can consider to obtain bounds for the Hartree-Fock energy. In any situation, bounds are usefull e.g. as stopping criteria for the iterative SCF procedure (and eventually to accelerate convergence); then, in order to obtain bounds on the Hartree-Fock energy, correction need to be solved on a grid fine enough to be considered exact as is the case of the computation presented in Fig. 1.

$\Phi^{\perp\!\!\!\perp}$; it is easy to see that P is block diagonal so projecting an element $\Psi = (\sum_{\alpha=1}^{n} \Psi_{i\alpha} h_\alpha)_{i=1}^{n}$ of X to $X \cap \Phi^{\perp\!\!\!\perp}$ will be an $O(N^3)$ operation. Let us denote by A_Φ the matrix of the second differential in Φ of the energy with respect to this discretization, and by b_Φ the "vector" corresponding to the first differential in Φ of the energy, interpreted as an element of the dual X'. The problem (56) has then the following discretization: find $w \in \mathbb{R}^{n \times N}$ such that $w = Pw$ and

$$(PA_\Phi P)w + (Pb_\Phi) = 0. \tag{70}$$

The matrix A_Φ of the linear system (70) is full and impossible to completely invert in practice due to the high computational complexity $O(N^6)$ required. However, using the same argument as above, applying the matrix A_Φ to a vector $v \in \mathbb{R}^{n \times N}$ can be done in $O(N^4)$ operations. The problem (70) is then solved iteratively ; finally let us remark that the total cost of the reconstruction of the symmetric part is an $O(N^3)$ process.

The a posteriori method was tested in the computation of the Hartree-Fock energy of the carbyn $Cr(CO)_4ClCH$ molecule. For each iteration step of the SCF algorithm the order ϵ^4 exact energy estimations were constructed, and also the corresponding lower bounds as described in remark 11. The convergence of the SCF method is presented in Fig. 2 and 3. Remark the presence of quadratically convergence periods (iterations 10-50), the presence of "jumps" (55-65) and slow convergence periods (70-90). In order to avoid the last regime, in practice one only uses the SCF algorithm for a small number of iterations 10-40 and then enlarges discretization basis, or tries to empirically optimize other parameters (DIIS).

The results obtained by the a posteriori procedure are presented in the Fig. 4 and 5. For some approximate solution obtained during the SCF iterations, the method described in previous section was applied to improve the energy and obtain a lower bound (initial data corresponding to more than 65 iterations is interpreted as converged due to numerical round-off errors); we do not attach special meaning to the good properties of the reconstructed error for $N = 30$ (cf. Fig. 5). As the results show, the method gives nearly converged results as soon as the initial approximation is as good as the one from the 10^{th} iteration of the SCF procedure.

Remark 14 *The number of iterations required to solve the linear system (70) was of the order of* 10*, which makes this method more efficient than the SCF cycles; for instance finding the improvement from the* 10^{th} *SCF cycle needs* 10 *iterations to solve (70) and is as good as the result of the* 60^{th} *SCF iteration.*

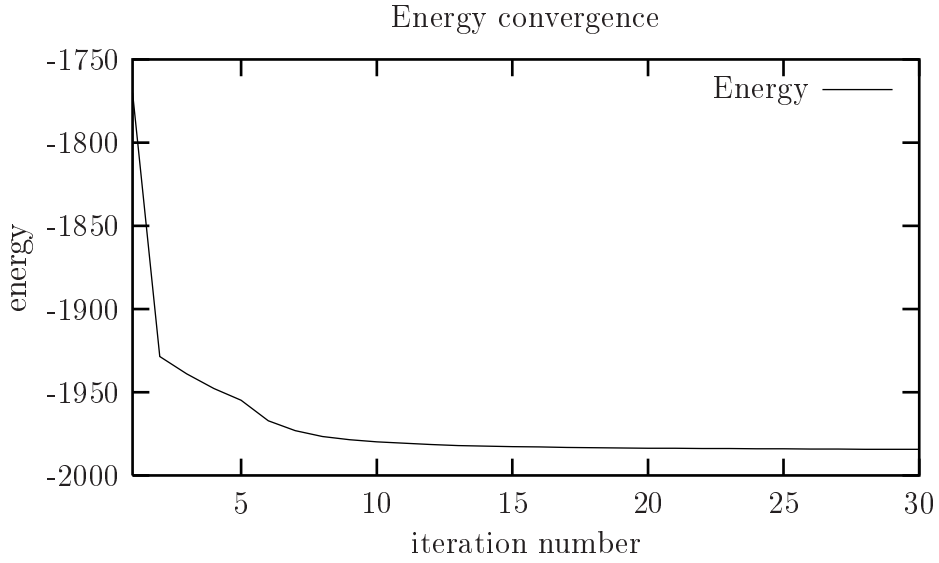


Figure 2: The convergence of the energy computed by the SCF algorithm in the form used by Chemists. The number of SCF cycles (iterations) ranges between 1 and 30. No a posteriori improvements are made.

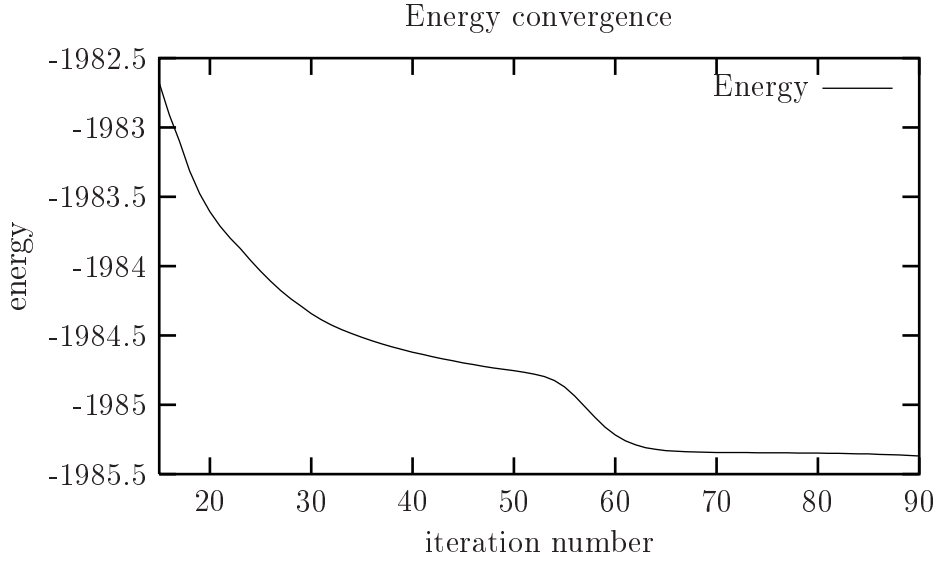


Figure 3: The convergence of the energy computed by the SCF algorithm in the form used by Chemists. The number of SCF cycles (iterations) ranges between 15 and 90. No a posteriori improvements are made.

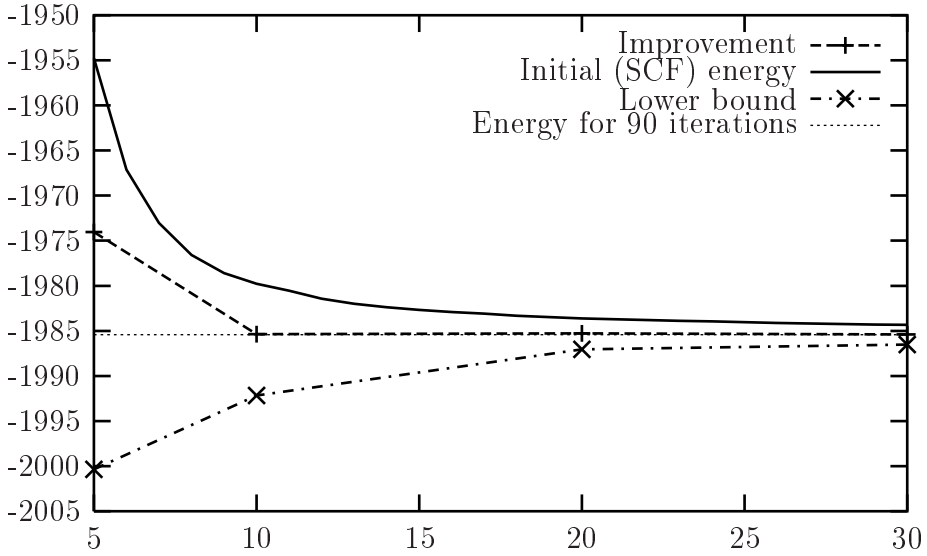


Figure 4: A posteriori error bounds and improvements are computed for the results of the SCF procedure. In each case we plot the energy of the initial (SCF) approximation, the energy of the wavefunction as computed by the a posteriori improvement procedure and the lower bound as described in Remark 11. The reference value of the energy is the result of the SCF algorithm after 90 iterations. The initial approximations to improve are the results of the SCF procedure for a number of cycles between 5 and 30.

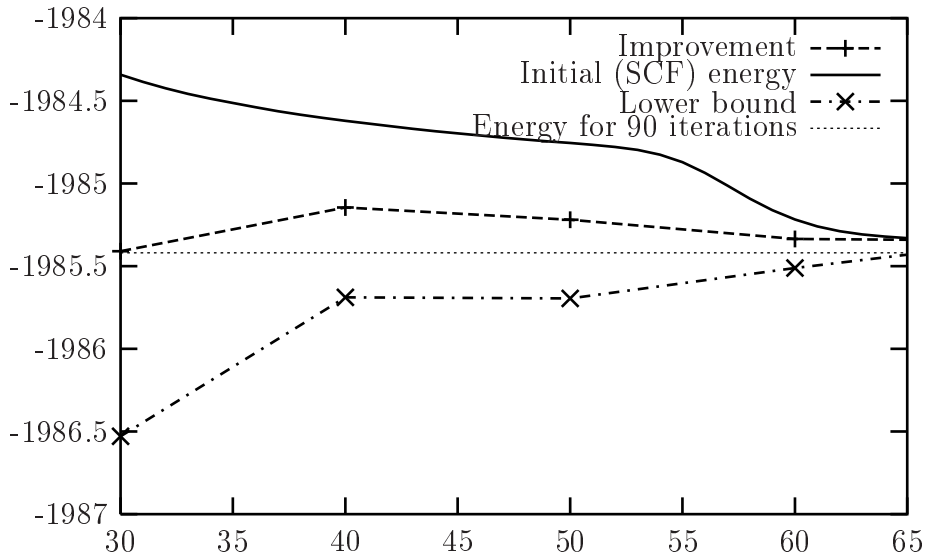


Figure 5: See Fig. 4 for details. The initial approximations to improve are the results of the SCF procedure for a number of cycles between 30 and 65.

Remark 15 *Applying the matrix A_Φ to a vector $v \in \mathbb{R}^{n\times N}$ in (70) requires at most $O(N^4)$ operations. The method is however compatible with the a priori introduction of further localization properties (as domain decomposition methods) of the electronic wavefunction as it is usually the case when more efficient Hartree-Fock computations are searched for [18], which results in the application of the matrix A_Φ being a $O(N^3)$ process (or even less); combining with classical convergence acceleration tools from the linear system solving (preconditioning ...) and with theorem 11, this method can be also seen as another approach towards to the design of Hartree-Fock computations of lesser algorithmic complexity.*

References

[1] I. Babuška "A posteriori error estimation for the finite element method", Internat. J. Numer. Methods Engrg. **12** (1978), 1597-1615

[2] I. Babuška and C. Schwab "A posteriori error estimation for hierarchic models of elliptic boundary value problems on thin domains" SIAM J. Numer. Anal. Vol **33** (1996), No.1, pp 241-246

[3] C. Bernardi, B. Metivet, "Indicateurs d'erreur pour l'equation de la chaleur. (Error indicators for the heat equation)" Rev. Eur. Elem. Finis 9, No.4, 425-438 (2000).

[4] E.Cancès and C.LeBris, "On the convergence of SCF algorithms for the Hartree-Fock equations", M2AN, Vol. 34, No. 4, 2000, pp. 749-774

[5] R. Ernenwein, M-M. Rohmer and M. Bénard, "A program system for ab initio MO calculations on vector and parallel processing machines, I. Evaluation of integrals", Comput. Phys. Comm. 58(1990), p.305-328

[6] S. Goedecker "Linear scaling electronic structure methods", Rev. Mod. Phys vol. 71, No.4, July 1999, p 1085-1123

[7] P.Ladevèze and D.Leguillon "Error estimate procedure in the finite element method and applications" SIAM J.Numer Anal. Vol 20,1991, no 5, 485-504

[8] E.H.Lieb and B.Simon " The Hartree-Fock theory for Coulomb systems" Commun. Math. Phys. 53, 185-194 (1977)

[9] P.L.Lions "Solutions of Hartree-Fock Equations for Coulomb systems" Commun. Math. Phys. 109, 33-97(1987)

[10] Y. Maday, A.T. Patera "Numerical analysis of a posteriori finite element bounds for linear-functional outputs", Math. Models Methods Appl. Sci. 10 (2000), no. 5, 785–799.

[11] Y. Maday, A. T. Patera and J. Peraire, "A general formulation for a posteriori bounds for output functionals of partial differential equations; application to the eigenvalue problem", Comptes Rendus de l'Académie des Sciences - Serie I - Mathematiques (328) 9 (1999) pp. 823-828

[12] J.T.Oden and Y.Feng, "Local and pollution error estimation for finite element approximations of elliptic boundary value problems" J.Comput. Appl. Math, **74**, 245-293 (1996)

[13] M. Paraschivoiu, A.T. Patera, "A hierarchical duality approach to bounds for the outputs of partial differential equations", Computer Methods in Applied Mechanics and Engineering 158 (3-4) (1998) pp. 389-407.

[14] M. Paraschivoiu, J. Peraire, A.T. Patera, "A posteriori finite element bounds for linear-functional outputs of elliptic partial differential equations", Computer Methods in Applied Mechanics and Engineering 150 (1-4) (1997) pp. 289-312.

[15] J. Peraire, A.T. Patera, "Asymptotic a posteriori finite element bounds for the outputs of noncoercive problems: the Helmholtz and Burgers equations", Computer Methods in Applied Mechanics and Engineering 171 (1-2) (1999) pp. 77-86.

[16] J. Pousin et J. Rappaz, "Consistency, stability, a priori and a posteriori errors for Petrov-Galerkin methods applied to nonlinear problems" (Report), EPFL, Lausanne, 1992

[17] M-M. Rohmer, J. Demuynck, M. Bénard, R. Wiest, C. Bachmann, C. Henriet and R. Ernenwein "A program system for ab initio MO calculations on vector and parallel processing machines, II. SCF closed-shell and open shell iterations", Comput. Phys. Comm. 60(1990), p.127-144

[18] M. C. Strain, G. E. Scuseria and M. J. Frisch, "Achieving linear scaling for the electronic quantum Coulomb problem." Science 271, 51 (1996).

[19] R.Verfürth "A Review of A Posteriori Error Estimates and Adaptative Mesh-Refinement Techniques", Wiley-Teubner 1997

[20] R.Verfürth "A Posteriori Error Estimates For Non-Linear Problems. Finite Element Discretisations of Elliptic Equations" Math. of Comp. **62**, 206(1994), pp 445-475

[21] R. Wiest, J. Demuynck, M. Bénard, M-M. Rohmer and R. Ernenwein "A program system for ab initio MO calculations on vector and parallel processing machines, III. Integral reordering and four-index transformation", Comput. Phys. Comm. 62(1991), p.107-124

Troisième partie

Contrôle en chimie quantique

Le contrôle quantique des processus chimiques est un thème de recherche très actif aujourd'hui [43, 44, 45, 46, 48, 49, 50, 51, 53, 54, 55, 58, 62, 69]. Outre l'obtention de réactions chimiques ou de produits nouveaux, cet effort est motivé par la possibilité de comprendre et de prédire les phénomènes en jeu à l'échelle atomique ainsi que par un large panorama d'applications dépassant le cadre de la chimie [58].

L'action qui est censée réaliser le contrôle est l'interaction de la matière avec un faisceau laser. Les premières expériences, qui ont tenté d'utiliser une impulsion laser construite selon l'intuition physique, ont donné des résultats très mitigés. C'est assez récemment que des outils venant de la théorie du contrôle d'une part et des développements de la technique des lasers femtoseconde d'autre part ont commencé à donner des résultats satisfaisants dans quelques cas particuliers (expériences à boucle fermée, [41, 53]). Néanmoins, des études expérimentales et théoriques sont encore nécessaires pour comprendre la nature subtile de tels processus de contrôle.

En France, des équipes d'horizons très divers travaillent dans ce domaine comme par exemple dans l'"Action Concertée Incitative" [73], animée par Claude Le Bris de l'École Nationale des Ponts et Chaussées. Elle a permis de réunir des chercheurs de cultures et compétences variées autour de quelques sujets d'étude communs.

Toujours dans le cadre des collaborations interdisciplinaires, une étude sur la contrôlabilité, les applications et les simulations numériques concernant le contrôle des systèmes quantiques a été menée conjointement avec Herschel Rabitz du Département de Chimie de l'Université de Princeton. Cette étude est présentée dans les chapitres suivants.

La contrôlabilité des systèmes quantiques de dimension finie est étudiée au chapitre 1 et est présentée sous la forme de trois articles, [66] (section 1.1), [67] (section 1.2) et [68] (section 1.3).

La problématique de la contrôlabilité en dimension infinie[1] ainsi que quelques techniques utiles dans la suite pour l'obtention des résultats en dimension finie sont présentés dans la section 1.1. L'étude dans la section 1.2 a pour but de donner des critères facilement calculables pour décider de la contrôlabilité d'un système discret (de dimension finie) donné. Ce travail est une généralisation de la Note [64] qui traitait de la contrôlabilité de la *population* des états propres dans le même cadre.

Afin d'énoncer le résultat central de cette étude, considérons un système quantique décrit par les équations (4) et (5), section 1.2, ou A et B sont les matrices de l'hamiltonien interne et de l'opérateur moment dipolaire (qui modélise l'interaction du système avec le faisceau laser), C est l'état

[1] voir aussi [42, 52] pour des discussions sur les résultats disponibles en dimension infinie

et $\epsilon(t) \in \mathbb{R}$ est le contrôle (intensité du laser). Les matrices A et B sont réelles symétriques; on peut donc considérer sans perte de généralité que A est diagonale et on note par λ_1,...,λ_N ses éléments diagonaux. Il est classique de voir que sous ces hypothèses le système évolue sur la sphère unité $S_M(0,1)$ de l'espace discret, donnée par les équations (6) section 1.2. On associe à ce système un graphe défini par (7) section 1.2 qu'on suppose connexe. Supposons aussi que ce graphe n'a pas des "transitions dégénérées", c'est-à-dire que pour tout $(i,j) \neq (a,b)$, $i \neq j$, $a \neq b$ tel que $B_{ij} \neq 0$, $B_{ab} \neq 0$: $\lambda_i - \lambda_j \neq \lambda_a - \lambda_b$.

Sous cettes hypothèses (voir aussi section A.1) on montre le résultat de contrôlabilité suivant:

Théorème. *Le système décrit par les équations (4) et (5) section 1.2 est contrôlable, c'est à dire pour tout $C \in S_M(0,1)$, l'ensemble des états atteignables à partir de C est $S_M(0,1)$.*

Quelques extensions de ce résultat sont présentées en annexe A.

Les applications de ce théorème et quelques extensions pour les systèmes qui ont des "transitions dégénérées" sont étudiées en section 1.3; des simulations numériques sont présentées à titre d'illustration. On conclut cette section par quelques considérations sur les relations qui existent entre les lois de conservation et les "défauts" de contrôlabilité.

Les résultats théoriques de contrôlabilité n'étant pas constructifs, on décrit dans le chapitre 2 une approche utilisée pour le calcul des champs qui réalisent le contrôle. Il s'agit d'un code numérique parallèle qui implémente des méthodes d'optimisation connues sous le nom d'*algorithmes génétiques*(AG). Après une introduction générale aux AG on présente quelques simulations numériques. On montre par la suite comment mettre au point des techniques de *filtrage* afin de sélectionner, dans la multitude des solutions, celles qui nous seront les plus utiles pour comprendre les mécanismes de contrôle en jeu.

Chapitre 1.
Contrôlabilité des systèmes quantiques
1.1 Considérations générales

Controllable quantities for bilinear quantum systems

Gabriel TURINICI
ASCI-CNRS Laboratory, Bat. 506, Université Paris Sud, 91405 Orsay Cedex
turinici@asci.fr

Abstract

This paper is dedicated to the search of tailored controllability concepts for quantum systems interacting with lasers. A negative result for infinite dimensional spaces serves as motivation for a finite dimensional analysis. We show that under physically reasonable hypothesis we can locally control sets of observables. As a remarkable particular case global exact controllability is proven for the population of the eigenstates.

1 Introduction

Controlling chemical reactions at the quantum level was a long-lasting goal [1, 3, 4, 5, 9, 11, 13, 14, 16, 17] from the very beginning of the laser technology. Indeed, due to the subtle nature of the interactions involved, this kind of manipulation is expected to allow on the one hand for much efficient and finer control than classical tools and on the other hand for new phenomena to be revealed.

The first experiments have shown that designing the laser pulse able to ensure the desired properties of the system is a non-trivial task that physical intuition alone cannot accomplish. It is only recently that tools coming from the control theory began to give satisfactory results in some particular cases.

A legitimate question arises in this context: what are the new controllability concepts that best fit this framework and which are the quantum quantities that can be exactly controlled using such an external field ? Some answers are given below.

2 Infinite dimensional controllability

The problem under consideration is controlling the time evolution of quantum systems. Let us consider such an independent system with internal Hamiltonian H_0 and let $\Psi_0(x)$ be its initial state. Denoting by $\Psi(x,t)$ the state at time t the evolution equations (**T**ime **D**ependent **S**chrödinger **E**quations) for the free system read:

$$\begin{cases} i\hbar\frac{\partial}{\partial t}\Psi(x,t) = H_0\Psi(x,t) \\ \Psi(x,t=0) = \Psi_0(x),\ \|\Psi_0\|_{L^2(\mathbf{R}^\gamma)} = 1. \end{cases} \tag{2.1}$$

The external action expected to allow for control is a laser field modeled by a laser intensity $\epsilon(t) \in \mathbf{R}$ and by a certain time independent dipole moment operator $\mathcal{B}$[1]. The new Hamiltonian is $H = H_0 - \epsilon(t)\mathcal{B}$ and the dynamical equations read:

$$\begin{cases} i\hbar\frac{\partial}{\partial t}\Psi_\epsilon(x,t) = H\Psi_\epsilon(x,t) \\ \Psi_\epsilon(x,t=0) = \Psi_0(x). \end{cases} \tag{2.2}$$

In a first approximation the goal may be formalized as to find (if any) a final time T and a finite energy laser pulse $\epsilon(t)$, $\epsilon(t) \in L^2([0,T])$ able to steer the system from $\Psi_0(x)$ to some predefined target $\Psi_\epsilon(x,T) = \Psi_{target}(x)$.

Note that the L^2 norm of Ψ_ϵ is conserved throughout the evolution:

$$\|\Psi_\epsilon(x,t)\|_{L^2_x(\mathbf{R}^\gamma)} = \|\Psi_0\|_{L^2(\mathbf{R}^\gamma)},\ \forall t > 0. \tag{2.3}$$

In general, for any selfadjoint operator O such that $[H_0, O]$ and $[\mathcal{B}, O]$ are both zero[2] one obtains

$$< \Psi_\epsilon(x,t)|O|\Psi_\epsilon(x,t) >=< \Psi_0|O|\Psi_0 >,\ \forall t > 0, \tag{2.4}$$

[1] Depending on the problem, one may choose to go beyond this first-order, bilinear term when describing the interaction between the laser and the system [7, 8].

[2] For two operators/matrices T_1 and T_2 we define $[T_1, T_2] = -T_1T_2 + T_2T_1$.

with the usual notation

$$< \Psi|O|\Psi >=< \Psi, O\Psi >_{L^2}=< O\Psi, \Psi >_{L^2} .$$

One remarkable class of operators are L^2-projections to closed subspaces. Let P be a projection to a closed subspace X of $L^2(\mathbf{R}^\gamma)$. Then $[H_0, P] = [\mathcal{B}, P] = 0$ means in particular that X and its orthogonal complement $X^\perp$ are involutive for H_0 and $\mathcal{B}$, i.e.

$$\begin{cases} \forall \Psi \in X: \ H_0\Psi \in X, \ \mathcal{B}\Psi \in X \\ \forall \Psi \in X^\perp: \ H_0\Psi \in X^\perp, \ \mathcal{B}\Psi \in X^\perp \end{cases} \tag{2.5}$$

The system can then be viewed as decomposed into two independent subsystems with corresponding wavefunctions equal to the projections of the total wavefunction to X and $X^\perp$. Of course this decomposition can be further refined for any additional projection operator that commutes with H_0 and $\mathcal{B}$. In order not to introduce unnecessary complications, we will suppose in all that follows that the system has only a **finite** number of independent subsystems (although the theory can be accommodated to fit a countable number of subsystems which can be proved to be the general case), each being associated its L^2-projector P_1,...,P_K such that:

$$[H_0, P_i] = [\mathcal{B}, P_i] = 0, \ \forall i = 1, ..., K \tag{2.6}$$

Moreover one can prove that the projectors can be chosen to fulfill the following conditions:

$$\sum_{i=1}^{K} P_i = I, \ P_iP_j = 0, \ \forall i \neq j, \ i, j = 1, ...K \tag{2.7}$$

Denote by S_{Ψ_0} the product of hyper-spheres: $S_{\Psi_0} = \{f \in L^2(\mathbf{R}^\gamma); \|P_i f\|_{L^2(\mathbf{R}^\gamma)} = \|P_i\Psi_0\|_{L^2(\mathbf{R}^\gamma)}, i = 1, ..., K\}$ By using 2.5 for the projectors P_1,...,P_k one can prove that the system evolves on S_{Ψ_0}.

Let us point out that due to the quantum nature of the system it follows by the uncertainty principle that one will never be able to experimentally verify, neither fully exploit, the **exact** controllability. In fact even if one obtains exactly the desired target state Ψ_{target} the free evolution (i.e. when laser is switched off $\epsilon(t) = 0, t \geq T$) of the quantum system **instantaneously modifies** this state (by a time dependent phase shift if Ψ_{target} is an eigenfunction of H_0 and by the formula (2.1) in general). In this context a negative controllability result is therefore not really restrictive. In fact using arguments as in [2] we may prove[3]

[3] For a different view on this issue we refer to [10], where generic negative controllability results are presented together with some simple cases where controllability is proved. Let us point out that their analysis is done on piecewise constant functions which may not always carry physical meaning for the problem considered; in particular one may prove controllability in this class and notice (by the theorem 2.1) that this controllability requires infinite L^2 norm and therefore infinite laser energy.

Theorem 2.1 *Let $\mathcal{B}$ be a bounded operator from $H^2_x(\mathbf{R}^\gamma)$ into itself and let H_0 generate a C^0 semigroup of bounded linear operators on $H^2_x(\mathbf{R}^\gamma)$. Denote by $\Psi_\epsilon(x,t)$ the solution of (2.2). Then the set of attainable states from Ψ_0 defined by*

$$\mathcal{AS} = \cup_{T>0}\{\Psi_\epsilon(x,T); \epsilon(t) \in L^2([0,T])\} \tag{2.8}$$

is contained in a countable union of compact subsets of $H^2_x(\mathbf{R}^\gamma) \cap S_{\Psi_0}$. In particular its complement with respect to S_{Ψ_0}: $\mathcal{N} = S_{\Psi_0} \setminus \mathcal{AS}$ is everywhere dense on S_{Ψ_0}. The same holds true for the complement with respect to $S_{\Psi_0} \cap H^2_x(\mathbf{R}^\gamma)$.

Proof: To prove the first part of the theorem one applies Thm. 3.6 from [2] on the space $H^2_x(\mathbf{R}^\gamma)$ for the operators $-iH_0$ and $-i\mathcal{B}$ (and restricts $\epsilon(t)$ to L^2 functions). Denote for any set A:

$$A_{r_1,...,r_K} = \{\sum_{i=1}^{K} s_iP_if; \ 0 \leq s_i \leq r_i, \ f \in A\}$$

Then for any compact subset C of X $C_{r_1,...,r_K}$ is also compact. Applying this to the compact components C of $\mathcal{AS}$ one notes that

$$\cup_{r_1\geq 0,...,r_K\geq 0}\mathcal{AS}_{r_1,...,r_K} = \cup_{n\in\mathbf{N}^*}\mathcal{AS}_{n,...,n}$$

is also a countable union of compacts subsets of $H^2_x(\mathbf{R}^\gamma)$. It follows by the Baire category theorem that $\cup_{r_1\geq 0,...,r_K\geq 0}\mathcal{AS}_{r_1,...,r_K}$ has dense complement in $H^2_x(\mathbf{R}^\gamma)$; in particular the complement of $\mathcal{AS}$ with respect to $S_{\Psi_0} \cap H^2_x(\mathbf{R}^\gamma)$ has to be everywhere dense on $S_{\Psi_0} \cap H^2_x(\mathbf{R}^\gamma)$. ∎

Given this result the search for exactly controllable quantities has to be directed to the finite dimensional setting.

3 Finite dimensional controllability

Let $D = \{\Psi_i(x); i = 1, .., N\}$ be an orthonormal basis for a finite dimensional sub-space[4] F of $L^2(\mathbf{R}^\gamma)$ and A and B be the matrices of the operators H_0 and $\mathcal{B}$ with respect to this base:

$$A_{ij} =< \Psi_i, H_0\Psi_j >, \ B_{ij} =< \Psi_i, \mathcal{B}\Psi_j >, \ i, j = 1, ..., N.$$

Denote $C = (c_i)_{i=1}^N$ as the coefficients of $\Psi_i(x)$ in the formula of the evolving state $\Psi(x,t) = \sum_{i=1}^N c_i(t)\Psi_i(x)$. From now we will work in atomic units only ($\hbar = 1$) ; the equations (2.2) read

$$i\frac{\partial}{\partial t}C_\epsilon = AC_\epsilon - \epsilon(t)BC_\epsilon, \ C_\epsilon(t=0) = C_0 \tag{3.1}$$

$$C_0 = (c_{0i})_{i=1}^N, \ c_{0i} =< \Psi_0, \Psi_i > \tag{3.2}$$

[4] This space is given by our model and the functions $\Psi_i(x)$ are usually the first eigenfunctions of H_0 constructed by a prior computation or by a modeling based on observations.

The controllability of (3.1) has been dealt with in the literature (cf. [12]) by deriving results from the controllability of a system posed on the space of the unitary matrices of dimension N. This approach has the benefit of granting access to the general tools on the controllability of bilinear systems on Lie groups. However, these results give only sufficient conditions for exact controllability (due to the setting which is more general). Finally there exists a class of simple quantum systems controllable in a sense to be defined further on that do not verify the criteria emerging from the Lie group analysis. We have therefore judged instructive to study this issue in a new framework; we were thus lead into identifying simple **necessary and sufficient** conditions for the finite dimensional controllability (see also [5] for an introduction to this topic).

In the case of our modeling the A matrix is diagonal and B is symmetrical with null diagonal elements (see [15] for the general case). Let us denote by $\lambda_i,\ i = 1,..,N$ the diagonal elements of A (the energies of the states Ψ_i). Before presenting the theoretical results we will introduce the controllability concept used.

Let O_1,...,O_p be positive quantum observables (positive autoadjoint operators). We say that the distribution of observables $\delta = (\delta_i)_{i=1}^p,\ \delta_i \geq 0,\ i = 1,...,p$ is reachable from the initial state C_0 if for any $\eta > 0$ there exists a final time $T_d > 0$ and an electric field $\epsilon(t) \in L^2([0,T_d])$ such that the solution of (3.1) satisfies:

$$| < \Psi(x,T_d)|O_i|\Psi(x,T_d) > -\delta_i^2| < \eta,\ i = 1,...,p$$

If this is also true for $\eta = 0$ we say that *the distribution of observables δ can be exactly reached from the initial state C_0*.

A special case of positive observables are the projections on the eigenstates P_{Ψ_i} defined by $P_{\Psi_i}\Psi =< \Psi, \Psi_i >_{L^2} \Psi_i,\ i = 1,...N$. The observable quantities $< \Psi|P_{\Psi_i}|\Psi >$ corresponding to this operators are called *populations* of the eigenstates. In our case these are $|c_{\epsilon k}(T_d)|^2$. A remarkable property of these observables is that when the system is evolving freely ((3.1) with $\epsilon(t) = 0$) the populations of the eigenstates do not change.

As it was previously seen the system evolves on the unit sphere of $L^2_x(\mathbf{R}^\gamma)$ which in finite dimensional representation reads $\sum_{i=1}^N |c_{\epsilon i}(t)|^2 = 1,\ \forall t \geq 0$. We call *population distribution* for the system (3.1) any N-tuple $d \in \mathbf{R}^N$ such that

$$\sum_{i=1}^N d_i^2 = 1,\ d_i \geq 0,\ i = 1,...,N \qquad (3.3)$$

A population distribution being a particular case of distribution of observables we extend the reachability concepts defined above to this case also.

4 Transfer graph and necessary conditions

We define as in [15] the non-oriented *transfer graph* of the system $G = (V,E)$ which corresponds to the intuitive image of population flow among different eigenstates of the system. The set V of vertices is the set of eigenstates Ψ_i and the set of edges E is the set of all pairs of eigenstates coupled by the matrix B:

$$G = (V,E),\ V = \{\Psi_1,...,\Psi_N\}\ E = \{(\Psi_i,\Psi_j); B_{ij} \neq 0\} \qquad (4.1)$$

This graph can be decomposed into connected components $G_\alpha = (V_\alpha, E_\alpha),\ a = 1,..,K$ that correspond to a bloc-diagonal structure of the matrix B (modulo permutations on the indices). It is worthwhile mentioning that this operation is the discrete version of the decomposition using projection operators that was undertaken for the infinite dimensional case; indeed, for each connected component $G_\alpha,\ \alpha = 1,...,K$, one can associate the linear space spanned by the eigenfunctions in V_α and prove that the (discrete) projection operator on this space P_α commutes with A and B.

Let $\tilde{D} = \{\tilde{\Psi}_1,...,\tilde{\Psi}_N\}$ be an orthonormal basis for the finite dimensional space F and $\tilde{P}_1,...,\tilde{P}_N$ projections operators on $\tilde{\Psi}_1$,...,$\tilde{\Psi}_N$ respectively. Suppose moreover that these observables are commuting with $P_1,...,P_K$, which is equivalent to the fact that $\tilde{D}$ is the union of orthonormal basis for each subsystem. Denote by U the unitary matrix that allow to change between the orthonormal basis D and $\tilde{D}$: $\tilde{\Psi}_i = \sum_j U_{ij}\Psi_j$. We will suppose in all that follows that all entries in U are **real**.

One can check by the definition of G and using equations (3.1) that for all $\alpha = 1,..,K$: $i\frac{d}{dt}\|P_\alpha\Psi(x,t)\|^2_{L^2} = 0$; each subsystem (connected component) comply therefore with the conservation laws

$$\sum_{\{i;\Psi_i \in V_\alpha\}} < \Psi(x,t)|\tilde{P}_i|\Psi(x,t) >= constant,$$
$$t > 0, \alpha = 1,..,K \qquad (4.2)$$

This allows us to give necessary conditions for controllability

Lemma 4.1 *If the distribution of observables δ is reachable from the initial configuration C_0 then*

$$\sum_{\{i;\Psi_i \in V_\alpha\}} < \Psi_0|\tilde{P}_i|\Psi_0 >= \sum_{\{i;\Psi_i \in V_\alpha\}} \delta_i^2,\ \alpha = 1,..,K. \qquad (4.3)$$

As a particular case one obtains the following

Corollary 4.1 *If the population distribution d is*

reachable from the initial configuration C_0 then

$$\sum_{\{i;\Psi_i\in V_\alpha\}} |c_{0i}|^2 = \sum_{\{i;\Psi_i\in V_\alpha\}} d_i^2, \quad \alpha = 1,..,K. \qquad (4.4)$$

5 Controllability results

Denote $\omega_{kl} = \lambda_k - \lambda_l,\ k,l = 1,...,N$. Let us introduce the following hypothesis:

$\mathbb{A}$ *The components $G_\alpha,\ \alpha = 1,..,K$ of G remain connected after elimination of all edge pairs $(\Psi_i,\Psi_j),(\Psi_a,\Psi_b)$ such that $\omega_{ij} = \omega_{ab}$ (degenerate transitions).*

Theorem 5.1 *(Local exact controllability) Let $T > 0$ be a given final time, $\epsilon_0(t) \in L^2([0,T])$ a given laser field such that:*

$\mathbb{B}$ $\lim_{t\to T}\epsilon_0(t) = 0,$

so in particular the limit $\lim_{t\to T}\epsilon_0(t)$ is supposed to exist (see also Remark 5.1); let Ψ_T be the state at time T of the system propagated with the laser field ϵ_0 and δ_T (d_T) the distribution of observables (populations) associated to the state Ψ_T:

$$\delta_T = (\sqrt{<\Psi_T|\tilde{P}_i|\Psi_T>})_{i=1}^N,$$
$$d_T = (|<\Psi_T,\Psi_i>_{L^2}|)_{i=1}^N = (|c_{0i}|)_{i=1}^N.$$

Suppose $(d_T)_i \neq 0, (\delta_T)_i \neq 0,\ i = 1,...,N$ and that the hypothesis $\mathbb{A}$ is verified. Suppose also that:

$\mathbb{C}$ *For each connected component $G_\alpha,\ \alpha = 1,...,K$ of G it does* **not** *exists a partition $V_\alpha = V_\alpha^1 \cup V_\alpha^2, V_\alpha^1 \cap V_\alpha^2 = \emptyset$ such that*

$$|\sum_{a\in V_\alpha^1} U_{jq} <\Psi_T,\Psi_a>| = |\sum_{b\in V_\alpha^2} U_{jb} <\Psi_T,\Psi_b>|,\ \forall j \in V_\alpha \qquad (5.1)$$

or if such a partition exists then

$$\frac{\sum_{a\in V_\alpha^1} U_{ja} <\Psi_T,\Psi_a>}{\sum_{b\in V_\alpha^2} U_{jb} <\Psi_T,\Psi_b>} = constant,\ \forall j \in V_\alpha.$$

Then there exists an open neighborhood D of δ_T on the surface of $\mathbf{R}^N$ given by the necessary conditions (4.3) endowed with the canonical topology such that one can exactly reach any distribution of observables δ in D from C_0.

Remark 5.1 *The hypothesis $\mathbb{B}$ is not really restrictive. In all practical cases $\epsilon_0(t)$ is continuous (at least at final/initial time) which assures the existence of the limit. The requirement that the limit of $\epsilon_0(t)$ in T be exactly 0 can be readily satisfied by replacing the triplet (ϵ_0, A, B) by $(\epsilon - \epsilon_0(T), A + \epsilon_0(T)B, B)$, where $\epsilon_0(T) = \lim_{t\to T}\epsilon_0(t)$. Note that in this situation the hypothesis $\mathbb{A}$ has to be verified for the eigenvalues of $A + \epsilon_0(T)B$ which are in general different from those of A. Finally, note that the set of final states Ψ_T that do* **not** *comply with the hypothesis $\mathbb{C}$ is of null canonical measure for any (real) unitary matrix U.*

Remark 5.2 *The result above may be somehow surprising due to the specific concept of locality used. In fact, suppose that the evolution of the system has ended in some final state p_T with the corresponding distribution of observables δ_T. Then, in order to obtain some other admissible distribution δ_c close to δ_T one has to go back in time and modify the electric field rather than to start from p_T and go for δ_c ! To understand this one has to remember that the observables* **do not necessarily commute** *with the hamiltonian so the free evolution (from p_T) drags the distribution of observables towards the direction given by the evolution equations 2.2; there is therefore no reason to hope that small perturbations (after the time T) can always counter-balance this bias and at the same time fill out a neighborhood of δ_T.*

Remark 5.3 *The technical conditions $(\delta_T)_i \neq 0,\ i = 1,...,N$ can also be intuitively justified. Indeed if some $(\delta_T)_i = 0$ one have to take care when choosing the good target set to expect exact controllability into, since there is no reason to hope in (exactly) reaching "distributions" having some* **strictly negative** *observables, as any projection-like observable is a* **positive** *operator.*

Proof: For the sake of simplicity we treat only the case $\omega_{ij} \neq \omega_{ab},\ \forall (i,j) \neq (a,b)$, the general case bearing no new concepts. Let us denote $\overline{A} = -iA$ and $\overline{B} = -iB$. Then (3.1) become:

$$\frac{\partial}{\partial t}C_\epsilon = (\overline{A} + \epsilon(t)\overline{B})C_\epsilon, \quad C_\epsilon(t=0) = C_0 \qquad (5.2)$$

Denote by $c(\epsilon, C_0, t) = (c_a(\epsilon, C_0, t))_{a=1}^N$ the solution at the time t of (3.1) for the initial ($t = 0$) data C_0 and electric field $\epsilon(t)$. Denote also $w(t) = c(\epsilon_0, C_0, t)$ and consider the canonical base $\{e_1,...,e_N\}$ of $\mathbf{R}^N$.

We define the application $M : L^2(\mathbf{R}) \to \mathbf{R}^N$ given by

$$M(\epsilon) = (<c(\epsilon, C_0, T)|\tilde{P}_a|c(\epsilon, C_0, T)>)_{a=1}^N \qquad (5.3)$$

Note that by the necessary conditions (4.3) the range of M is a subset of

$$\Big\{(x_i)_{i=1}^N \in \mathbf{R}^N; \sum_{\{i;\Psi_i\in V_\alpha\}} x_i = \sum_{\{i;\Psi_i\in V_\alpha\}} < \Psi_0|\tilde{P}_i|\Psi_0 > \quad \alpha = 1,..,K\Big\}$$

The local controllability is in fact a particular surjectivity property of M. We will prove that the differential DM of M has the surjectivity property we desire and by the implicit function theorem the conclusion will follow then for M itself. More precisely we prove that DM is onto the linear manifold (P) (product of hyper-planes of $\mathbf{R}^{cardinality(S_\alpha)}$, $\alpha = 1,..,K$):

$$\Big\{(x_i)_{i=1}^N \in \mathbf{R}^N; \sum_{\{i;\Psi_i\in V_\alpha\}} x_i = 0,\ \alpha = 1,..,K\Big\}$$

whose $M(\epsilon_0)$-translation is tangent to the range of M.

Denote by $f_a,\ a = 1,...,N$ the components of DM:

$$DM(\epsilon)|_{\epsilon=\epsilon_0}\cdot\tilde{\epsilon} = \big(< f_a, \tilde{\epsilon} >_{L^2}\big)_{a=1}^N \tag{5.4}$$

Due to the finite dimensionality of our setting we just have to show that the range of $DM(\epsilon)|_{\epsilon=\epsilon_0}$ has a null orthogonal with respect to (P), that is any vector $\mathbf{k} = (k_a)_{a=1}^N \in \mathbf{R}^N$ such that

$$\sum_{\{i;\Psi_i\in V_\alpha\}} k_i = 0, \quad \alpha = 1,..,K \tag{5.5}$$

$$\sum_{i=1}^N k_i \cdot < f_i, \tilde{\epsilon} >_{L^2} = 0,\ \forall\tilde{\epsilon} \in L^2([0,T]) \tag{5.6}$$

is necessary the null vector. Equation (5.6) can also be written

$$\sum_{i=1} k_i \cdot f_i(s) = 0,\ \forall 0 \le s \le T \tag{5.7}$$

The system (5.2) can be written in the integral form:

$$c(t) = e^{\int_0^t \overline{A}+\epsilon_0\overline{B}}c(0) + \int_0^t e^{\int_s^t \overline{A}+\epsilon_0\overline{B}}(\epsilon(s)-\epsilon_0(s))\overline{B}c(s)ds \tag{5.8}$$

which gives [2] the formula of the (Fréchet) derivative $D_\epsilon c(\epsilon, C_0, t)$ of $c(\epsilon, C_0, t)$ with respect to ϵ computed at $\epsilon(t) = \epsilon_0(t)$:

$$D_\epsilon c(\epsilon, C_0, t)|_{\epsilon=0}\cdot\tilde{\epsilon} = \int_0^t e^{\int_s^t \overline{A}+\epsilon_0\overline{B}}\tilde{\epsilon}(s)\overline{B}e^{\int_0^s \overline{A}+\epsilon_0\overline{B}}c(0)ds \tag{5.9}$$

Then it is easy to see that

$$DM(\epsilon)|_{\epsilon=0}\cdot\tilde{\epsilon} = \Big[2Re < D_\epsilon w(T)\cdot\tilde{\epsilon}|\tilde{P}_a|w(T) >\Big]_{a=1}^N \tag{5.10}$$

so we obtain after some manipulations

$$f_a(s) = 2Re < e^{\int_s^T \overline{A}+\epsilon_0\overline{B}}\overline{B}e^{\int_T^s \overline{A}+\epsilon_0\overline{B}}w(T)|\tilde{P}_a|w(T) > \tag{5.11}$$

From 5.7 we obtain that

$$\sum_{a=1}^N k_a \frac{d^k}{ds^k} f_a(s)|_{s=T} = 0 \tag{5.12}$$

To compute the derivatives $\frac{d^k}{ds^k}f_a(s)|_{s=T}$ we make use of a variant of the Campbell - Baker - Hausdorff formula:

$$e^{-Y\tau}Ze^{Y\tau} = Z + \tau[Y,Z] + \frac{\tau^2}{2}[Y,[Y,Z]] + ... \tag{5.13}$$

Define recursively $ad_Z^i Y = [Y, ad_Z^{i-1}Y]$ and $ad_Z^0 Y = Z$; we obtain after making use of the hypothesis $\epsilon_0(T) = 0$:

$$Re < ad_{\overline{B}}^q \overline{A}w(T)|\sum_{a=1}^N k_a\tilde{P}_a|w(T) >= 0,\ q \ge 0 \tag{5.14}$$

The matrix of the operator $\tilde{P}_a$ in the basis $\tilde{D}$ is simply $diag\{\delta_{ia}\}_{i=1}^N$ so the matrix of $\sum_{a=1}^N k_a\tilde{P}_a$ with respect to the basis D is $U^t\left(diag\{k_a\}_{a=1}^N\right)U$. By straightforward computations one obtains

$$ad_{\overline{B}}^q\overline{A} = \big((-i)^{q+1}\omega_{ab}^q B_{ab}\big)_{a,b=1}^N.$$

Note also the general property that when $\omega_{ab}(a < b)$ are all different then the only way to have

$$\sum_{a<b}\omega_{ab}^q r_{ab} = 0,\ q = 0,1,...$$

is when r_{ab} are **all** zero. Denote $\tilde{w}(T) = Uw(T)$ (the coefficients of $\sum_i w_i(T)\Psi_i$ in the base $\tilde{D}$). Using the ingredients above one proves that

$$If\ B_{kl} \ne 0: \sum_{i,j=1}^N (k_i - k_j)\tilde{w}_i(T)\overline{\tilde{w}_j(T)}U_{ik}U_{jl} = 0 \tag{5.15}$$

All that remains to be done is to show that the only way to have (5.5, 5.15) is when $\mathbf{k} = 0$. Note that because any connected component G_α has at least cardinality$(V_\alpha) - 1$ edges, in (5.5, 5.15) there are at least N relations so this is in fact a linear system to solve. We will suppose in all that follows that G has only one connected component; the general case can be

reduced to this one due to the commutation relations $[\tilde{P}_a, P_b] = 0,\ a = 1, ..., N,\ b = 1, ..., K$.

The remaining of the proof being rather technical so we will only sketch it. Denote $v = U^t(diag\{k_j\}_{j=1}^N)\overline{Uw(T)}$; then equation 5.15 can be written $w_k(T)\overline{v_l} = \overline{w_l(T)}v_k$ or, since $w_i(T) \neq 0,\ i = 1, ..., N$:

$$\text{If } B_{kl} \neq 0 \text{ then: } \frac{v_k}{w_k(T)} = \overline{\left(\frac{v_l}{w_l(T)}\right)} \qquad (5.16)$$

By the connectivity of G one obtains that there exists γ such that for each $i = 1, ..., N$ $v_i = \gamma w_i(T)$ or $v_i = \overline{\gamma} w_i(T)$. If γ is real one can infer $\mathbf{k} = 0$ by the definition of v. If γ is not real then divide indexes in two sets V_1 and V_2 such that for $i \in V_1$: $v_i = \gamma w_i(T)$ and for $j \in V_2$: $v_j = \overline{\gamma} w_j(T)$. One can obtain then a formula for k_i:

$$k_i = \frac{(U^t v)_i}{\tilde{w}_i(T)} = \frac{\gamma \sum_{j \in V_1} U_{im} w_m(T) + \overline{\gamma} \sum_{j \in V_1} U_{im} w_m(T)}{\tilde{w}_i(T)}. \qquad (5.17)$$

By this formula, for k_i to be real equation 5.1 from the hypothesis $\mathbb{C}$ has to be true; if this is not the case then γ is real and thus $\mathbf{k} = 0$. On the other side if the second assumption of $\mathbb{C}$ is true then it is easy to prove k_a is a constant that does not depend of $a,\ a = 1, ..., N$ so by 5.5 we obtain again $\mathbf{k} = 0$. ∎

A straightforward application of the theorem above is the following (Thm. 2 from [15]):

Corollary 5.1 *(Local exact controllability for populations) Let d_0 be the population distribution associated to the initial state C_0: $d_0 = (|c_{0i}|)_{i=1,...,N}$. Suppose $d_{0i} \neq 0,\ i = 1, ..., N$ and that the hypothesis $\mathbb{A}$ is verified. Then there exists an open neighborhood D of d_0 on the surface of $\mathbf{R}^N$ given by the necessary conditions (4.4) endowed with the canonical topology such that one can exactly reach any population distribution d in D from C_0.*

Proof: Apply the theorem 5.1 for arbitrary final time T and null electric field $\epsilon(t) \equiv 0$. Since the free evolution of the system preserves the populations of the eigenstates, we obtain for $\tilde{P}_i = P_{\Psi_i}$ that $\delta_T = d_T = (|c_{0i}|)_{i=1}^N$ so the only hypothesis left to verify is $\mathbb{C}$. This also is trivial since in this case $U = I$ and for j such that (for instance) $j \in V_a^1$ the first part of $\mathbb{C}$ can be written $| < \Psi_T, \Psi_j > | = 0$, impossible since $| < \Psi_T, \Psi_j > | = |c_{0j}| \neq 0$. ∎

Let us also mention for the sake of completeness the global exact controllability result that can be proved [15] using on the one hand the Corollary 5.1 and on the other hand approximate global controllability results (Thm. 3 [15]).

Theorem 5.2 *(Global exact controllability) Let d_0 be the population distribution associated to the initial state C_0: $d_0 = (|c_{0i}|)_{i=1,...,N}$. Under the hypothesis $\mathbb{A}$ any population distribution $d = (d_i)_{i=1}^N$ such that $d_i \neq 0,\ i = 1, ..., N$ which verifies the necessary conditions (4.4) can be* **exactly** *reached from C_0.*

6 Conclusions

Controllability of the bilinear quantum systems has been studied in the infinite and finite dimensional settings. The classical control concepts seem to be not very well adapted to the the infinite dimensional case and a negative result has been given as illustration. For the finite dimensional case, positive results have been obtained for exact local controllability of sets of projection-type observables and global controllability has been proven for the particular case when the observables are the populations of eigenstates. Easy to check and intuitively simple to understand necessary and sufficient conditions have been obtained to characterize the attainable set.

Acknowledgements. It is a pleasure to acknowledge helpful talks that we had on this topic with Prof. Yvon Maday (ASCI Laboratory).

References

[1] A. Assion et al. "Control of Chemical Reactions by Feedback-Optimized Phase-Shaping Femtosecond Laser Pulses" Science vol. 282 (1998) pp. 919-922

[2] J.M.Ball, J.E.Madersen and M.Slemrod, "Controllability for distributed bilinear systems", SIAM J.Control and Optimization, vol 20, No.4, 1982, pp.575-597

[3] C. Le Bris, "Control theory applied to Quantum Chemistry: Some tracks", ESAIM : Proceedings, vol. 8, 2000, pp 77-94.

[4] P.Brumer and M.Shapiro, Acc.Chem Res. Vol. 22, p.407 , 1989

[5] A.G. Butkovskiy, Yu.I.Samoilenko, "Control of quantum-mechanical processes and systems", Kluwer,1990

[6] Reinhard Diestel "Graph Theory", 2nd ed. Springer-Verlag, New York, Graduate Texts in Mathematics, Vol. 173, Feb. 2000

[7] C.M. Dion et al., Chem. Phys.Lett 302(1999) 215-223

[8] C.M. Dion, A.Keller, O.Atabek & A.D. Bandrauk, Phys. Rew. A 59(2) 1999, p.1382

[9] Kime K., Appl. Math. Lett. 6 (3) (1993) 11–15.

[10] G.M.Huang, T.J.Tarn & J.W.Clark, "On the controllability of quantum-mechanical systems", J.Math.Phys 24(1983), p.2608

[11] A.P.Pierce, M.A. Dahleh and H.Rabitz, Phys Rev.A **37**, 4950 (1988)

[12] V. Ramakrishna & al. "Controllability of molecular systems" , Phys. Rew. A, Vol 51, No.2, 1995 pp. 960-966

[13] S.Shi, A.Woody, and H.Rabitz, J.Chem Phys. 88(1988), p.6870

[14] D.J.Tannor and S.A.Rice, J.Chem Phys 83(1985), p.5013

[15] G. Turinici "On the controllability of bilinear quantum systems" in M.Defranceschi, C.LeBris (Eds.), "Mathematical models and methods for ab initio Quantum Chemistry", Lecture Notes in Chemistry, volume 74, Springer, 2000 ISBN: 3-540-67631-7

[16] Herschel Rabitz, Regina de Vivie-Riedle, Marcus Motzkus, and Karl Kompa "Whither the Future of Controlling Quantum Phenomena?" Science 2000 May 5; 288: 824-828.

[17] W.S.Warren, H.Rabitz and M.Dahleh, "Coherent control of quantum dynamics", Science 259 (1993)

1.2 Résultats théoriques

Wavefunction controllability in quantum systems

Gabriel Turinici
ASCI-CNRS Laboratory, Bat. 506, Université Paris Sud, 91405 Orsay Cedex
Herschel Rabitz
Department of Chemistry, Princeton University, Princeton, New Jersey 08544-1009

Abstract

We present controllability results for quantum systems interacting with lasers. Exact controllability for the wavefunction in these bilinear systems is proved in the finite dimensional case under very natural hypotheses. The controllability conditions are necessary and sufficient.

PACS number(s): 32.80.Qk

1 Introduction

Controlling chemical reactions at the quantum level is a long-lasting goal (cf. [2, 3, 5, 9, 11, 12, 13, 15, 16, 17, 22]) going back the very beginning of laser technology. Due to the subtle nature of the interactions involved, manipulation of quantum dynamics is expected to allow for finer control than classical tools (e.g. temperature and pressure) and possibly for new reactions and/or products. Controlling quantum phenomena also goes beyond chemical reactions to encompass many other applications [13].

The earliest experiments showed that designing a laser pulse capable of steering the system to the desired target state is a rather difficult task that physical intuition alone generally cannot accomplish. It is only recently that tools from control theory were introduced and began to give satisfactory results in some particular cases; finding the optimal laser electric field as a design objective is treated by numerical methods and a need exists for new

methods that are reliable and computationally inexpensive. A legitimate question arises in this context: what quantum states can be attained using such an external field ? Some answers are given below for finite dimensional quantum systems.

2 Dynamical Equations

This section introduces the general infinite dimensional equations for controllability analysis; their discretization is discussed in the next section. Consider a quantum system (treated first as isolated) without control interaction with internal Hamiltonian H_0 and prepared in the initial state $\Psi_0(x)$ where x denotes the relevant coordinate variables; the state $\Psi(x,t)$ at time t satisfies the time-dependent Schrödinger equation

$$\begin{aligned} i\hbar\tfrac{\partial}{\partial t}\Psi(x,t) = &\quad H_0\Psi(x,t) \\ \Psi(x,t=0) = &\quad \Psi_0(x),\ \|\Psi_0\|_{L^2(\mathbf{R}^\gamma)} = 1 \end{aligned} \tag{1}$$

In the presence of an external interaction, taken here as an electric field modeled by a laser amplitude $\epsilon(t) \in \mathbf{R}$ coupled to the system through a time independent dipole moment operator $\mathcal{B}$ (see also [23]) the (controlled) dynamical equations become:

$$\begin{aligned} i\hbar\tfrac{\partial}{\partial t}\Psi_\epsilon(x,t) = &\quad H_0\Psi_\epsilon(x,t) + \epsilon(t)\cdot\mathcal{B}\Psi_\epsilon(x,t) = H\Psi_\epsilon(x,t) \\ \Psi_\epsilon(x,t=0) = &\quad \Psi_0(x) \end{aligned} \tag{2}$$

In order to avoid trivial control problems we suppose $[H_0,\mathcal{B}] \neq 0$, where the Lie bracket $[\cdot,\cdot]$ is defined as $[U,V] = UV - VU$.

The goal is to find if any final time $T > 0$ and finite energy laser pulse $\epsilon(t) \in L^2([0,T])$ exist such that $\epsilon(t)$ is able to steer the system from $\Psi_0(x)$ to some predefined target $\Psi_\epsilon(x,T) = \Psi_{target}(x)$. If the answer to this question is affirmative, then the system is controllable. Given that H is Hermitian one can easily prove that the L^2 norm of Ψ_ϵ is conserved throughout the evolution:

$$\|\Psi_\epsilon(x,t)\|_{L^2_x(\mathbf{R}^\gamma)} = \|\Psi_0\|_{L^2(\mathbf{R}^\gamma)},\ \forall t > 0. \tag{3}$$

Note that $\Psi_\epsilon(x,t)$ evolves on the unit sphere $S(0,1)$ of $L^2(\mathbf{R}^\gamma)$:

$$S(0,1) = \{f \in L^2(\mathbf{R}^\gamma); \|f\|_{L^2(\mathbf{R}^\gamma)} = 1\}$$

3 Finite dimensional system

Let $D = \{\Psi_i(x); i = 1, .., N\}$ be the set of the first $N, N \geq 3$ eigenstates of the infinite dimensional Hamiltonian H_0, let M be the linear space they generate, and let A and B be the matrices of the operators H_0 and $\mathcal{B}$ respectively, with respect to this base; as in the infinite dimensional setting it is supposed that $[A, B] \neq 0$. Negative generic results concerning the infinite dimensional controllability (cf. [1, 10, 19, 21]) are available that show the need for tailored controllability concepts and for a good understanding of the finite dimensional case; moreover the existence of intrinsically finite dimensional quantum chemical situations ("N-level" systems, spin systems, etc.) motivates a finite dimensional analysis.

We denote $C = (c_i)_{i=1}^N$ to be the coefficients of $\Psi_i(x)$ in an expansion of the evolving state $\Psi(t, x) = \sum_{i=1}^N c_i(t)\Psi_i(x)$; Eq (2) now becomes

$$\begin{cases} i\hbar\frac{\partial}{\partial t}C = AC + \epsilon(t)BC \\ C(t=0) = C_0 \end{cases} \tag{4}$$

$$C_0 = (c_{0i})_{i=1}^N,\ c_{0i} =< \Psi_0, \Psi_i >,\ \sum_{i=1}^N |c_{0i}|^2 = 1 \tag{5}$$

The controllability of Eq. (4) has been dealt with in the literature (cf. [14]) by considering the problem of the controllability of a system posed on the space of the unitary matrices of dimension N. This approach has the benefit of drawing on the general tools and results from bilinear controllability on Lie groups. However, verifying those criteria may be computationally difficult when N is large; moreover the results obtained this way give only sufficient conditions for exact controllability (due to a setting that is more general than often required). Thus, we consider identifying the **necessary and sufficient** conditions for the finite dimensional controllability (see also [4] for an overview of the topic).

We make the common assumptions that the A matrix is diagonal and that the B matrix is real symmetric (Hermitian). We denote $\lambda_i \in \mathbf{R},\ i = 1, .., N$, to be the diagonal elements of A (the energies of the states Ψ_i). With the notation $S_M(0, 1) = S(0, 1) \cap M$, it was previously stated that the system evolves on $S_M(0, 1)$, which in a finite dimensional representation reads:

$$\sum_{i=1}^N |c_i(t)|^2 = 1,\ \forall t \geq 0 \tag{6}$$

4 Connectivity graph and necessary conditions

The B matrix plays the critical role of specifying the kinematic coupling amongst the eigenstates of the system reference Hamiltonian matrix A. We associate to the system a graph $G = (V, E)$ called the *connectivity graph* (we refer the reader to [6] for graph theory concepts). We define the set V of vertices as the set of eigenstates Ψ_i and the set of edges E as the set of all pairs of eigenstates coupled by the matrix B. Since B is symmetric we can consider G as non-oriented:

$$G = (V, E): \quad V = \{\Psi_1, ..., \Psi_n\} \quad E = \{(\Psi_i, \Psi_j); i \neq j, \ B_{ij} \neq 0\} \tag{7}$$

We may decompose this graph into (connected) components $G_\alpha = (V_\alpha, E_\alpha)$, $a = 1, .., K$. Note that this decomposition corresponds to a bloc-diagonal structure of the matrix B (modulo some permutations on the indices). From the definition of G and using Eq. (4) (A is diagonal) it follows that

$$i\hbar \frac{\partial}{\partial t} \sum_{\{i; \Psi_i \in V_\alpha\}} |c_i(t)|^2 = 0 \tag{8}$$

Using Eq. (8) one can write new conservation laws for each component:

$$\sum_{\{i; \Psi_i \in V_\alpha\}} |c_i(t)|^2 = constant, \ \forall t > 0, \ \alpha = 1, .., K \tag{9}$$

Denote by $U(A, B, \epsilon, t_1 \to t_2)$ the propagator associated with Eq. (4); for any state $\chi(t_1)$, $U(A, B, \epsilon, t_1 \to t_2)\chi(t_1)$ is defined as the solution at time $t = t_2$ of Eq. (4) with the initial state at time $t = t_1$ being $\chi(t_1)$. This operator is also symbolically written $U(A, B, \epsilon, t_1 \to t_2) = e^{-i \int_{t_1}^{t_2} (A + \epsilon(t) B) dt}$.

Definition 1 *We say that Ψ_2 is reachable from Ψ_1 if there exists $0 < T < \infty$, $\epsilon(t) \in L^2([0, T]; \mathbf{R})$ such that $U(A, B, \epsilon(t), 0 \to T)\Psi_1 = \Psi_2$.*

This allows us to give necessary conditions for controllability:

Lemma 1 *If the state $\chi = \sum_{i=1}^N d_i \Psi_i(x)$ is reachable from the initial configuration C_0 then*

$$\sum_{\{i; \Psi_i \in V_\alpha\}} |c_{0i}|^2 = \sum_{\{i; \Psi_i \in V_\alpha\}} |d_i|^2, \quad \alpha = 1, .., K \tag{10}$$

In order to simplify the presentation of the results we will introduce the following hypothesis:

$\mathbb{HA}$ *The graph G is connected, i.e. $K = 1$.*

The assumption $\mathbb{HA}$ is not at all restrictive, it is just a matter of specifying the number of independent subsystems we want to simultaneously control (see [19] for the general case). Note also that this does **not** imply that any two states are directly connected one with another, but only that for any two states Ψ_i and Ψ_j there is a path in the graph G that connects Ψ_i and Ψ_j.

5 Controllability

We denote $\omega_{kl} = \lambda_k - \lambda_l,\ k, l = 1, ..., N$ as the eigenvalue differences for the matrix A, and atomic units ($\hbar = 1$) will be utilized. Consider the hypothesis:

$\mathbb{HB}$ *The connectivity graph G does not have "degenerate transitions", that is for all $(i, j) \neq (a, b)$, $i \neq j$, $a \neq b$ such that $B_{ij} \neq 0$, $B_{ab} \neq 0$: $\omega_{ij} \neq \omega_{ab}$.*

Remark 1 *In all that follows this hypothesis could be relaxed to*

$\mathbb{HC}$ *The connectivity graph G remains connected after elimination of all edge pairs $(\Psi_i, \Psi_j), (\Psi_a, \Psi_b)$ such that $\omega_{ij} = \omega_{ab}$ (degenerate transitions).*

However, to ease of presentation $\mathbb{HB}$ will be assumed to be true.

We also introduce one more hypothesis:

$\mathbb{HD}$ *For each $i, j, a, b = 1, ..., N$ such that $\omega_{ij} \neq 0$: $\frac{\omega_{ab}}{\omega_{ij}} \in Q$, where Q is the set of all rational numbers.*

Remark 2 *The assumption $\mathbb{HD}$ implies that there exists a $T > 0$ such that $U(A, B, 0, 0 \to T) = e^{-iTA} = I$ (i.e. the free evolution is periodic). Note that $\mathbb{HD}$ is in particular verified if $\lambda_i \in Q, i = 1, ..., N$, which is often the case in practice (e.g. [18]). Moreover alternative proofs of the controllability result without $\mathbb{HD}$ are possible and will be presented in a future paper.*

We will conclude with a simple but important remark: the reverse (i.e. the same dynamics but with time reversed) of the system (4) given by $(A, B, \epsilon(t))$ is equivalent to a system of the same kind $(-A, -B, \tilde{\epsilon}(t) = \epsilon(-t))$, such that:

$$(U(A, B, \epsilon(t), t_1 \to t_2))^{-1} = U(-A, -B, \epsilon(-t), -t_2 \to -t_1).$$

We call $(A, B, \epsilon(t))$ the "direct system" and $(-A, -B, \tilde{\epsilon}(t))$ the corresponding "reverse system".

The goal is to prove that under hypothesis $\mathbb{HA}$,$\mathbb{HB}$,$\mathbb{HD}$ the system is controllable, i.e. for any $\Psi_1 \in S(0,1) \cap M$ the set of reachable states from Ψ_1 is $S(0,1) \cap M$. The proof has two parts: local controllability and global controllability.

5.1 Local controllability

We begin by introducing two particular subsets of M; if the graph G admits a bipartite decomposition $V = P_1 \cup P_2, P_1 \cap P_2 = \emptyset, P_1 \neq \emptyset, P_2 \neq \emptyset, E \subset P_1 \times P_2$ denote

$$X = \{\chi = \sum_{i=1}^{N} w_i \Psi_i; \sum_{i \in P_1} \lambda_i |w_i|^2 = \sum_{j \in P_2} \lambda_j |w_j|^2\} \tag{11}$$

If G does not have a bipartite decomposition (thus it has at least one odd-length cycle, see [6] p.24) set $X = \emptyset$. We also introduce the set Z:

$$Z = \{\Psi = \sum_{i=1}^{N} c_i \Psi_i; \exists i : c_i = 0\} \tag{12}$$

Theorem 1 *Let $\Psi \in S_M(0,1) \setminus X \setminus Z$. Under the assumptions $\mathbb{HA}$, $\mathbb{HB}$, $\mathbb{HD}$ the set of reachable states from Ψ is a neighborhood of Ψ (in the canonic topology of $S_M(0,1)$). The same result is true for the reverse system, that is, the set of states from which Ψ can be reached is a neighborhood of Ψ.*

Proof. We will use on M its **real** Hilbert space structure (and not the canonical **complex** Hilbert space structure) given by the scalar product:

$$< \chi_1, \chi_2 >_{\mathbf{R}} = Re(< \chi_1, \chi_2 >) = \frac{1}{2}(< \chi_1, \chi_2 > + < \chi_2, \chi_1 >). \tag{13}$$

Consider the mapping $S : L^2(\mathbf{R}) \times \mathbf{R} \to M$ given by $S(\epsilon, t) = U(A, B, \epsilon, 0 \to t)\Psi$. We want to prove that $S(L^2(\mathbf{R}) \times (0, \infty))$ is a neighborhood of Ψ. Note that $S(0,T) = \Psi$ and that S is differentiable in $(0,T)$(see [1]). Therefore it suffices to prove that the differential DS of S in $(0,T)$ is onto the tangent plane $\mathcal{P}$ in Ψ at $S_M(0,1)$ given by the equation:

$$\mathcal{P} = \{\chi \in M :< \chi, \Psi > + < \Psi, \chi >= 0\}. \tag{14}$$

Since the image of the differential is a linear space, it is enough to prove that the only $\chi \in M$ such that:

$$\begin{cases} < DS_{(\epsilon,t)=(0,T)}(\epsilon,t), \chi >_{\mathbf{R}} = 0, \forall (\epsilon,t) \in L^2(\mathbf{R}) \times \mathbf{R} \\ < \chi, \Psi >_{\mathbf{R}} = 0 \end{cases} \tag{15}$$

is $\chi \equiv 0$. Let χ satisfy (15). Denote by DS_ϵ the differential of S with respect to ϵ in $(0,T)$ and by DS_t the differential of S with respect to t in $(0,T)$. Then (see also [1]):

$$\begin{cases} DS_\epsilon(\tilde{\epsilon}) = -i \int_0^T e^{-iA(T-s)} \tilde{\epsilon}(s) B e^{iA(T-s)} \Psi ds \\ DS_t = -iA\Psi \end{cases} \tag{16}$$

So (15) is equivalent to:

$$\begin{cases} Im(< e^{-iA(T-s)} B e^{iA(T-s)} \Psi, \chi >) = 0, \forall 0 < s < T \\ Im(< A\Psi, \chi >) = 0 \\ Re(< \Psi, \chi >) = 0 \end{cases} \tag{17}$$

Denote $\Psi = \sum_i c_i \Psi_i$, $\chi = \sum_i w_i \Psi_i$. Making use of the hypothesis $\mathbb{HB}$ as in [20] we obtain:

$$B_{ab}(c_a \overline{w_b} - \overline{c_b} w_a) = 0, \forall 1 \leq a < b \leq N \tag{18}$$

$$Re(\sum_{a=1}^N c_a \overline{w_a}) = 0 \tag{19}$$

$$Im(\sum_{a=1}^N \lambda_a c_a \overline{w_a}) = 0. \tag{20}$$

Equation (18) implies that for each a, b such that $B_{a,b} \neq 0$

$$\frac{w_a}{c_a} = \overline{\left(\frac{w_b}{c_b}\right)}. \tag{21}$$

Since the connectivity graph is fully connected we obtain easily that there exists a complex constant α such that for each $1 \leq a \leq N$ $w_a = \alpha c_a$ or $w_a = \overline{\alpha} c_a$. If G is not bipartite then it has an odd-length cycle. Using (21) along this cycle one obtains $\alpha \in \mathbf{R}$ so $\chi = \alpha\Psi$ and by (19) it follows that $\alpha = 0$ so $\chi \equiv 0$.

If G is bipartite with decomposition $V = P_1 \cup P_2$ we conclude from (21) that

$$\begin{cases} w_a = \alpha c_a, \forall a \in P_1 \\ w_a = \overline{\alpha} c_a, \forall a \in P_2 \end{cases} \tag{22}$$

From (19) and (20) one concludes that either $\alpha = 0$ (so $\chi \equiv 0$) either $\Psi \in X$. Replacing $(A, B, \epsilon(t))$ by $(-A, -B, \epsilon(-t))$ one obtains the second part of the theorem.

5.2 Global controllability

Theorem 2 *Under the assumptions* $\mathbb{HA}$, $\mathbb{HB}$, $\mathbb{HD}$ *the system (4) is controllable, that is for any* $\Psi \in S_M(0,1)$ *the set of reachable states from* Ψ *is* $S_M(0,1)$*; the same result is true for the reverse system.*

Proof. The proof is based on the following lemmas:

Lemma 2 *("exit lemma") For any* $\Psi \in S_M(0,1)$ *there exists at least one state in* $S_M(0,1) \setminus X \setminus Z$ *that can be reached from* Ψ*; the same is true for the reverse system.*

Lemma 3 *("pass lemma") If* $X \neq \emptyset$ *then, in any given open (for the canonical topology of* $X \cap S_M(0,1)$*) subset* V *of* $X \cap S_M(0,1)$ *there exists a "pass state"* $\gamma \in V \setminus Z$ *such that from* γ *one can reach at least one point in any (of the two)* **local in** γ *connected components of* $S_M(0,1) \setminus X$ *separated by* X*; moreover these points can be chosen not to be in* Z*; the same is true for the reverse system.*

Suppose lemmas 2 and 3 are both true; suppose also $X \neq \emptyset$ (the simpler alternative $X = \emptyset$ follows along the same lines). By the "exit-lemma" it is enough to prove (for the direct and inverse system) that for any $\Psi \in S_M(0,1) \setminus X \setminus Z$ the set of reachable states from Ψ is $S_M(0,1) \setminus X \setminus Z$. That is, use the lemma for the direct system to reach a state in $S_M(0,1) \setminus X \setminus Z$, and use it once more for the reverse system to obtain a state in $S_M(0,1) \setminus X \setminus Z$ from which the target can be reached and in the "middle" use the controllability from $S_M(0,1) \setminus X \setminus Z$ to $S_M(0,1) \setminus X \setminus Z$. The proof proceeds in two steps:

1. Suppose the initial state ϕ and target δ are in the same connected component of $S_M(0,1) \setminus X$. Then there exists a continuous curve $C(t) : [0,1] \to S_M(0,1) \setminus Z \setminus X$ with $C(0) = \phi$, $C(1) = \delta$. We will prove that each $C(t)$, $t \in [0,1]$ is reachable from ϕ. Indeed, let us denote by η the minimal value t such that $C(t)$ is not reachable from ϕ. By the local controllability result for the state ϕ we obtain $\eta > 0$. Since $C(\eta) \in S_M(0,1) \setminus Z \setminus X$ one can apply the local result for the reverse system in $C(\eta)$ and deduce that there exists $\eta' < \eta$ such $C(\eta)$ is reachable from $C(\eta')$. But, by the minimal property of η, $C(\eta')$ is reachable from ϕ so by transitivity $C(\eta)$ is also reachable from ϕ.

2. Let the initial state ϕ and target δ be in different connected components of $S_M(0,1) \setminus X$. For the sake of simplicity suppose that the connected components are adjacent (two components are called adjacent if the

intersection of their frontiers has a non void interior in the canonic topology of $S_M(0,1) \cap X$), the general case being a mere reiteration of the arguments below It can be proved, see also the discussion on the geometry of the set X below, that any two components of $S_M(0,1) \setminus X$ can be linked by a chain of adjacent components. Then there exists a "pass-state" $\gamma \in X \setminus Z$ given by lemma 3 on the boundary of the two connected components. By the properties of a "pass state" there exists two states ϕ' (in the same component as ϕ) and χ' (in the same component as χ) and an electric field such that the corresponding evolution starting from ϕ' passes by γ and arrives in χ'. Since by the previous case ϕ' is reachable from ϕ and χ from χ' an electric field realizing an evolution $\phi \to \phi' \to \gamma \to \chi' \to \chi$ can be found, and therefore χ is reachable from ϕ, which concludes our proof.

Before giving proofs for the lemmas above let us denote by D_i the L^2 projector to Ψ_i, $i = 1, ..., N$ and by O the operator $\sum_{i \in P_1} \lambda_i D_i - \sum_{j \in P_2} \lambda_j D_j$. We make use of the classical "bra-ket" notation for self-adjoint operators V (such as O, D_i): $< \chi_1 | V | \chi_2 > := < \chi_1, V\chi_2 > = < V\chi_1, \chi_2 >$). We obtain the following characterizations

$$X = \{\chi; < \chi | O | \chi > = 0\}, \ Z = \bigcup_{i=1}^{N} \{\chi; < \chi | D_i | \chi > = 0\} \tag{23}$$

Note also: $[H_0, O] = [H_0, D_i] = 0, \ i = 1, ..., N$, but $[\mathcal{B}, O] \neq 0$, $[\mathcal{B}, D_i] \neq 0, \ i = 1, ..., N$. We will use the same notation for the matrices of these operators with respect to the base D.

Proof of lemma 2. a) We begin by proving that for any $k = 1, ..., N$, $\chi \in S_M(0,1)$, $\eta > 0$, and $\tau > 0$ there exists at least an $\epsilon(t) \in L^2(0,\tau)$, $\|\epsilon\|_{L^2} < \eta$: such that:

$$\{U(A, B, \epsilon(t), 0 \to s)\chi; 0 \leq s \leq \tau\} \setminus D_k^{-1}\{0\} \neq \emptyset \tag{24}$$

Denote $U(A, B, \epsilon(t), 0 \to s)\chi = \chi(s) = \sum_{l=1}^{N} c_l(s)\Psi_l$ as the solution of (4). Suppose (24) is not true, then $c_k(s)$ vanishes on $[0, \tau]$ as well as all its derivatives, for any smooth electric field $\epsilon(t) \in C^\infty \cap L^2(0,\tau)$, $\|\epsilon\|_{L^2} < \eta$. We obtain to first order:

$$i\frac{d}{dt}c_k(s) = \epsilon(s)\sum_{j=1}^{N} B_{kj}c_j(s) = 0, \ \forall \ s \in [0,\tau], \ \epsilon(t) \in C^\infty \cap L^2(0,\tau), \ \|\epsilon\|_{L^2} < \eta \tag{25}$$

Take $\epsilon_n(t) = \frac{\eta}{n\sqrt{\tau}}$ and denote by $\chi_n(s) = \sum_{l=1}^N c_{nl}(s)\Psi_l$ the corresponding evolution. Since $\epsilon_n(s) \neq 0$ on $[0,\tau]$ it follows that

$$\sum_{j=1}^N B_{kj}c_{nj}(s) = 0,\ 0 \leq s \leq \tau,\ n = 1, ... \tag{26}$$

For $n \to \infty$ the limiting trajectory is the free evolution $c_j(s) = e^{-i\lambda_j s}c_j(0)$, therefore

$$\sum_{j=1}^N B_{kj}c_j(s) = \sum_{j=1}^N B_{kj}e^{-i\lambda_j s}c_j(0) = 0,\ 0 \leq s \leq \tau \tag{27}$$

By the hypothesis $\mathbb{HB}$ this can be true only if $c_j(0) = 0$ for all j connected to k in G ($B_{kj} \neq 0,\ j \neq k$). Selecting the initial time arbitrarily in $[0,\tau]$ one obtains that for any $\epsilon(t) \in L^2([0,\tau]),\ \|\epsilon\|_{L^2} < \eta$ and corresponding evolution $U(A,B,\epsilon(t),0 \to s)\chi = \chi(s) = \sum_{l=1}^N c_l(s)\Psi_l$ the coefficient $c_j(s)$ is zero for all $s \in [0,\tau]$ and all j connected to k in G. Repeating this reasoning as many times as necessary (starting each time from the newly obtained zero coefficients) and using the **connected** graph structure of B it follows that $c_j(s) = 0,\ 0 \leq s \leq \tau, j = 1, ..., N$, which is in obvious contradiction with $\chi \in S_M(0,1)$.
b) An immediate consequence of the assertion (24) is that for each state $\chi \in S_M(0,1)$ and each neighborhood V of χ there exists a reachable state from χ that is not in Z.
c) Since Z is a closed set, all that remains to prove is that for any state $\chi \in S_M(0,1) \setminus Z$ and neighborhood V of χ there exists at least one reachable state from χ in $V \cap S_M(0,1) \setminus X$.

Suppose that this is not true; then there exists $\chi \in S_M(0,1) \setminus Z)$, $\eta > 0$, and $\tau > 0$ such that for any $\epsilon(t) \in L^2(0,\tau)$, $\|\epsilon\|_{L^2} < \eta$:

$$< U(A,B,\epsilon(t),0 \to s)\chi|O|U(A,B,\epsilon(t),0 \to s)\chi >= 0,\ \forall s \in [0,\tau] \tag{28}$$

Denote $U(A,B,\epsilon(t),0 \to s)\chi = \chi(s) = \sum_{l=1}^N c_l(s)\Psi_l$ as the solution of (4) and $O(t) =< \chi(t)|O|\chi(t) >$. Then for any $\epsilon(t) \in L^2(0,\tau)$, $\|\epsilon\|_{L^2} < \eta$ $O(t)$ and all its derivatives vanish in $[0,\tau]$. To compute the first derivative use the formula for the evolution of an observable represented by a matrix V: $V(t) =< \chi(t)|V|\chi(t) >$:

$$\frac{d}{dt}V(t) =< \chi(t)|i[A,V]|\chi(t) > +\epsilon(t) < \chi(t)|i[B,V]|\chi(t) > . \tag{29}$$

Denote $J = [B, O]$; there exists $a \neq b$ such that $J_{ab} \neq 0, B_{ab} \neq 0$ and for $V = O$ in (29):

$$\frac{d}{ds}O(s) = \epsilon(s) < \chi(s)|iJ|\chi(s) >= 0,\ \forall s \in [0,\tau],\ \epsilon(t) \in C^\infty,\ \|\epsilon\|_{L^2} < \eta \tag{30}$$

Using the same technique as above one concludes that

$$J_{ij}Re(c_i(s)\overline{c_j(s)}) = 0,\ i \neq j,\ i,j = 1,...,N,\ \forall s \in [0,\tau] \tag{31}$$

so finally:

$$Re(c_a(s)\overline{c_b(s)}) = 0,\ \forall s \in [0,\tau],\ \forall \epsilon(t) \in L^2(0,\tau),\ \|\epsilon\|_{L^2} < \eta \tag{32}$$

If suffices to note that $2Re(c_a(s)\overline{c_b(s)}) =< c_a(s), c_b(s) > + < c_b(s), c_a(s) >$ is (a particular observable) **not** conserved by the free evolution ($\epsilon(t) \equiv 0$) so (32) cannot be true.

Proof of lemma 3. We begin with some geometry considerations concerning the set X. Following the definition (23) denote by $f : S_M(0,1) \to \mathbf{R}$ the function $f(\sum_{i=1}^N z_i\Psi_i) = \sum_{i\in P_1}\lambda_i|z_i|^2 - \sum_{j\in P_2}\lambda_j|z_j|^2$. Then $X = f^{-1}\{0\}$. The differential Df of f never vanishes in general and vanishes only on $KD = \{e^{i\phi}\Psi_k; 0 \leq \phi \leq 2\pi\}$ if some $\lambda_k = 0$, so for any open set $V_1 \in S_M(0,1) \cap X$ there exists a subset $V_2 \subset V_1$ such that Df never vanishes on V_2; locally on V_2 only two connected components $f^{-1}(]0,\infty[) \cap V_2$ and $f^{-1}(]-\infty, 0[) \cap V_2$ are present and globally KD does not introduce new connected components. For any two points $\phi, \delta \in S_M(0,1) \setminus X$ there exists a continuous curve from ϕ to δ that does not intersect KD, the real codimension of KD in X being at least 2. We can therefore suppose $V \cap KD = \emptyset$.

Let $\chi(s)$ be the solution of (4) for initial data $\chi(0)$ and electric field $\epsilon(t)$. By the definition of X (cf. (23)) the local connected components separated by X in $S_M(0,1)$ correspond to regions where the observable O has constant sign. In order to prove this lemma it is therefore enough to find a $\gamma \in V \setminus Z$ such that $< \gamma|[\mathcal{B}, O]\gamma > \neq 0$, with at least one state in each connected component being then reached from γ by choosing the good sign for $\epsilon(0)$. Since $V \setminus Z \neq \emptyset$ there exists $\gamma' \in V \setminus Z$. Note as above by J the matrix representation of $[\mathcal{B}, O]$ in the basis D and find $a \neq b$ such that $J_{ab} \neq 0$. Choose τ such that the free evolution $\gamma'(s) = \sum_{i=1}^N g_i(s)\Psi_i$ of a system starting from $\gamma'(0) = \gamma' \in X$ do not exit $V \setminus Z$ before time $s = \tau$ (when the laser is off the system is **guaranteed** to remain in X). We have seen before that the equality $Re(g_a(s)\overline{g_b(s)}) = 0$ is not conserved during the free evolution so we may also suppose $Re(g_a(0)\overline{g_b(0)}) \neq 0$. If at least one $s \in [0,\tau]$ is found such

$< \gamma'(s)|J|\gamma'(s) > \neq 0$ the lemma is proved; if this is not true, notice that $J_{ij} \neq 0$ only when $B_{ij} \neq 0$ and use the formula for the free evolution and the hypothesis $\mathbb{HB}$ to obtain that $J_{ab}Re(g_a(0)\overline{g_b(0)}) = 0$, which is a contradition.

Remark 3 *Even when the hamiltonian matrix A does not comply with $\mathbb{HB}$, thm 2 may still be used; indeed, it suffices to find a $\mu \in \mathbf{R}$ such that the eigenvalues of $A + \mu B$ satisfy $\mathbb{HB}$, apply the theorem for the system $(A + \mu B, B)$ and obtain an field $\tilde{\epsilon}(t)$; the answer is then the field $\tilde{\epsilon}(t) + \mu$ as the system $(A + \mu B, B, \tilde{\epsilon})$ is obviously equivalent to $(A, B, \tilde{\epsilon}(t) + \mu)$.*

6 Discussion and conclusions

Wavefunction controllability of finite dimensional bilinear quantum systems was analyzed and necessary and sufficient conditions were found under reasonable physical hypothesis on the system under consideration. Under hypothesis $\mathbb{HB}$ the only restrictions on the attainable set appear from conservations laws (Eq.(10)) in effect. Various other hypothesis ($\mathbb{HA}$, $\mathbb{HD}$) are only necessary for the present proof and will be eliminated in a future paper. The status of the hypothesis $\mathbb{HB}$ is more subtle; in certain cases its removal brings about new conservation laws (that will necessarily contract the attainable set) very different from those in Eq.(10). On the other hand, an analysis of the case $N = 3$ leads us to state the following

Conjecture *As long as no new conservation laws –besides L^2 norm conservation – appear, the system is controllable, i.e. any state on the unit sphere can be reached (in finite time and with finite laser energy) from any other.*

The merit of the formulation above is intrinsically related to the properties of the systems and **not** on their mathematical transcription. The existence of conservation laws possibly may *prevent* controllability or correspondingly just restrict the set of attainable states (i.e., if the necessary conditions thus introduced are also sufficient). On the other hand we remark that in some cases, in the absence of $\mathbb{HB}$, conservation laws may involve quantities that are **not** necessarily observables.

7 Acknowledgements

H.R. acknowledges support from the National Science Foundation and DOD. G.T. thanks Mathieu Pilot from CERMICS (École nationale des ponts et chaussées, Marne-la-Vallée, France) , for helpful discutions on this topic.

References

[1] J.M.Ball, J.E.Marsden and M.Slemrod, "Controllability for distributed bilinear systems", SIAM J.Control and Optimization, vol 20 (4) (1982), 575–597

[2] C. Le Bris, "Control theory applied to Quantum Chemistry: Some tracks" , International Conference on systems governed by PDEs, Nancy, March 1999, ESAIM : Proceedings, vol. 8, 2000, pp 77-94.

[3] P.Brumer and M.Shapiro, "Coherence Chemistry: Controlling Chemical Reactions with Lasers", Acc.Chem Res. 22, 12 (1989) 407–413.

[4] A.G. Butkovskiy, Yu.I.Samoilenko, "Control of quantum-mechanical processes and systems", Kluwer,1990

[5] M.Demilrap and H.Rabitz, Phys. Rew A., **47** 2 1983, p.831

[6] Reinhard Diestel "Graph Theory", 2nd ed. Springer-Verlag, New York, Graduate Texts in Mathematics, Vol. 173, Feb. 2000

[7] C.M. Dion et al., Chem. Phys.Lett 302(1999), 215-223

[8] C.M. Dion, A.Keller, O.Atabek & A.D. Bandrauk, Phys. Rew. A 59(2) 1999, p.1382

[9] Kime K., "Control of transition probabilities of the quantum-mechanical harmonic oscillator", Appl. Math. Lett. 6 (3) (1993) 11–15.

[10] Huang G.M., Tarn T.J., Clark J.W., "On the controllability of quantum-mechanical systems", J. Math. Phys. 24, 11 (1983) 2608–2618.

[11] Mei Kobayashi,"Mathematics make molecules dance" , SIAM News 24 (1998)

[12] A.P.Pierce, M.A. Dahleh and H.Rabitz, Phys Rev.A **37** (1988), p.4950

[13] Herschel Rabitz, Regina de Vivie-Riedle, Marcus Motzkus, and Karl Kompa "Whither the Future of Controlling Quantum Phenomena?" Science 2000 May 5; 288: 824-828.

[14] V. Ramakrishna, et al. "Controlability of molecular systems" Phys. Rev. A 51 (2) (1995) 960–966.

1.3 Applications

Quantum Wavefunction Controllability

Gabriel Turinici
ASCI-CNRS Laboratory, Bat. 506, Université Paris Sud, 91405 Orsay Cedex
Herschel Rabitz
Department of Chemistry, Princeton University, Princeton, New Jersey 08544-1009

Abstract

Theoretical results are presented on the ability to arbitrarily steer about a wavefunction for a quantum system under time-dependent external field control. Criteria on the field free Hamiltonian and the field coupling term in the Hamiltonian are presented that assure full wavefunction controllability. Numerical simulations are given to illustrate the criteria. A discussion on the theoretical and practical relationship between dynamical conservation laws and controllability is also included.

PACS number(s): 32.80.Qk

1 Introduction

There is much interest on controlling quantum systems through their interaction with external fields [1] - [11]. This activity is motivated by a potential wide range of applications [7] that this framework can accommodate. Encouraging positive results have already been obtained in closed loop experiments [12, 13], but both theoretical and experimental research is still needed to understand the subtle nature of the control processes.

Early efforts at achieving quantum control based on intuitive physical understanding generally gave poor results. Significant advances have come through the introduction of rigorous control theory tools together with enhanced laser pulse shaping capabilities. An important preliminary step to any experiment are indications of its feasibility through theoretical studies

and computer simulations. Such analyses can indicate the set of objectives that can reasonably be met and present the nature of a laser pulse to most likely meet the objectives. The study of the set of quantum states that can be attained is an aspect of control theory aimed at deciding whether the system is *controllable*, i.e. if any admissible quantum state can be attained with some (admissible) laser field. Until recently the answer to this question was given using results available in [14] or [15] ; although useful in many cases, these results may prove more general than often required, as in [14] where general results are derived for the evolution of unitary operators, or too pessimistic as in [15] where negative results are presented for infinite dimensional controllability. A theoretical study was then undertaken [16] to shed some light on the phenomena involved when controllability for the wavefunction is investigated in finite dimensional bilinear quantum systems. The purpose of this paper is to explore and discuss the practical utility of these latter formal theoretical results along with simple illustrations through computer simulations. The outline of the paper is as follows: the theoretical results are presented in section 2 ; supporting numerical simulations and some practical extensions of the theory are presented in section 3. A discussion on the connections between dynamical conservation laws and controllability of quantum systems is given in section 4 ; concluding remarks are presented in section 5.

2 Theoretical Controllability Criteria

Consider a quantum system with internal Hamiltonian H_0 prepared in the initial state $\Psi_0(x)$, where x denotes the relevant coordinate variables. The external interaction will be taken here as a control field amplitude $\epsilon(t) \in \mathbb{R}$ coupled to the system through a time independent (e.g, dipole moment) operator $\mathcal{B}$ (see also [17]) ; then the time-dependent control Schrödinger equation that gives the evolution of the state $\Psi(x,t)$ at time t is :

$$\begin{aligned} i\hbar\tfrac{\partial}{\partial t}\Psi(x,t) = &\quad H_0\Psi(x,t) + \epsilon(t)\cdot\mathcal{B}\Psi(x,t) = H\Psi(x,t) \\ \Psi(x,t=0) = &\quad \Psi_0(x) \end{aligned} \tag{1}$$

In order to avoid trivial control problems we suppose $[H_0, \mathcal{B}] \neq 0$, where the Lie bracket $[\cdot,\cdot]$ is defined as $[U,V] = UV - VU$.

The goal is to find if any final time $T > 0$ and finite energy control pulse $\epsilon(t) \in L^2([0,T])$ exist such that $\epsilon(t)$ is able to steer the system from $\Psi_0(x)$ to some predefined target $\Psi(x,T) = \Psi_{target}(x)$. If the answer to this question is affirmative, then the system is controllable. Given that H is Hermitian, the

L^2 norm $\|\Psi(x,t)\|_{L^2_x(\mathbb{R}^\gamma)}$ of Ψ is conserved throughout the evolution so that $\Psi(x,t)$ evolves on the unit sphere $S(0,1)$ of $L^2(\mathbb{R}^\gamma)$:

$$S(0,1) = \{f \in L^2(\mathbb{R}^\gamma); \|f\|_{L^2(\mathbb{R}^\gamma)} = 1\}$$

Numerical simulations on the system (1) require the introduction of a finite dimensional setting. A common choice is to consider the set $D = \{\Psi_i(x); i = 1, .., N\}$ of the first N eigenstates of the infinite dimensional Hamiltonian H_0 and restrict the operators involved to the linear space that D generates. Let A and B be the matrices of the operators H_0 and $\mathcal{B}$ respectively, in terms of this base, and as above, it is supposed that $[A, B] \neq 0$.

Before leaving the infinite dimensional setting, we remark that the controllability of (bilinear) quantum systems on infinite dimensional spaces is a difficult problem and the resolution of this matter is only partially solved. Moreover, the generic results obtained so far in this setting are **negative** [20, 15, 21, 22, 23] showing the need for tailored controllability concepts and a thorough understanding of the finite dimensional case in order to appropriately extend the **positive** controllability results available [14, 16] to the infinite dimensional setting. The present study is also motivated by the existence of intrinsically finite dimensional quantum mechanical situations (e.g. N-level spin systems, etc.).

We denote $C = (C_i)_{i=1}^N$ to be the coefficients of $\Psi_i(x)$ in an expansion of the evolving state $\Psi(t,x) = \sum_{i=1}^N C_i(t)\Psi_i(x)$, $N \geq 3$; Eq (1) now becomes

$$\begin{cases} i\hbar\frac{\partial}{\partial t}C = AC + \epsilon(t)BC \\ C(t=0) = C_0 \end{cases} \tag{2}$$

$$C_0 = (C_{0i})_{i=1}^N,\ C_{0i} =< \Psi_0, \Psi_i >,\ \sum_{i=1}^N |C_{0i}|^2 = 1 \tag{3}$$

The controllability of Eq. (2) has been already dealt with in the literature [14] by considering the problem of the controllability of a system posed on the space of the unitary matrices of dimension N. This elegant approach has the benefit of drawing on the general tools and results from bilinear controllability on Lie groups. However, verifying those criteria may be computationally difficult when N is large; moreover the results obtained this way give only sufficient conditions for exact controllability (due to a setting that is more general than often required). Thus, we consider identifying computationally convenient and intuitive conditions for finite dimensional wavefunctions to be reachable from an arbitrary initial state (see also [24] for an overview of the topic).

We make the common assumptions that the A matrix is diagonal and that the B matrix is real symmetric (Hermitian). Denote $\lambda_i \in \mathbb{R},\ i = 1, .., N$, as

the real diagonal elements of A (i.e. the energies associated with the states Ψ_i). Denote $S_M(0,1) = S(0,1) \cap M$. The conservation of the L^2 norm of the wavefunction can be written in the finite dimensional representation:

$$\sum_{i=1}^{N} |C_i(t)|^2 = 1, \ \forall t \geq 0 \tag{4}$$

2.1 Connectivity Graph

The B matrix plays the critical role of specifying the kinematic coupling amongst the eigenstates of the system reference Hamiltonian matrix A. The structure of the set of all direct and indirect couplings between eigenstates is very relevant to assessing controllability. In order to formalize the concepts, we associate to the system a non-oriented graph $G = (V, E)$ called the *connectivity graph* (the reader is referred to [25] for graph theory concepts). We define the set V of vertices as consisting of the eigenstates Ψ_i and the set of edges E as consisting of all pairs of eigenstates **directly** coupled by the matrix B.

$$G = (V, E): \quad V = \{\Psi_1, ..., \Psi_n\}, \quad E = \{(\Psi_i, \Psi_j); i < j, \ B_{ij} \neq 0\} \tag{5}$$

This graph can be decomposed into (connected) components $G_\alpha = (V_\alpha, E_\alpha)$, $a = 1, .., K$. In more intuitive terms, two eigenstates Ψ_α and $\Psi_{\alpha'}$ are in the same connected component (we will say that they are *indirectly coupled*) if there exist a path $\Psi_{j_1} = \Psi_\alpha, \Psi_{j_2}, ..., \Psi_{j_l} = \Psi_{\alpha'}$ from Ψ_α to $\Psi_{\alpha'}$ such that any **consecutive** vertices Ψ_{j_a} to $\Psi_{j_{a+1}}$ of this chain are *directly coupled*, i.e. the dipole moment $B_{j_a j_{a+1}}$ is non-zero (which is the same as $(\Psi_{j_a}, \Psi_{j_{a+1}}) \in E$) ; note that there is no need for non-consecutive vertices Ψ_{j_a} to Ψ_{j_b} to be directly connected, i.e. if $b \neq a+1$ and $a \neq b+1$ the entry $B_{j_a j_b}$ may be zero. This decomposition corresponds to a bloc-diagonal structure of the matrix B (modulo some permutations on the indices), so it is just a matter of specifying the number of independent subsystems we want to **simultaneously** control (see [21] for the general case). We will consider the following hypothesis as true

$\mathbb{TA}$ *The graph G is connected, i.e. $K = 1$.*

Remark 1 *In agreement with the definition above, note that $\mathbb{TA}$ does* **not** *imply that any two states are necessarily* **directly** *connected, one with the other, but only that for any two states Ψ_α and $\Psi_{\alpha'}$ there is a path in the graph G that connects Ψ_α and $\Psi_{\alpha'}$.*

Denote by $U(A, B, \epsilon, t_1 \to t_2)$ the propagator associated with Eq. (2); for any state $\chi(t_1)$, $U(A, B, \epsilon, t_1 \to t_2)\chi(t_1)$ is defined as the solution at time $t = t_2$ of Eq. (2) with the initial state at time $t = t_1$ being $\chi(t_1)$.

Definition 1 *We say that $\tilde{\Psi}$ is reachable from $\overline{\Psi}$ if there exists $0 < T < \infty$, $\epsilon(t) \in L^2([0, T]; \mathbb{R})$ such that $U(A, B, \epsilon(t), 0 \to T)\tilde{\Psi} = \overline{\Psi}$.*

2.2 Controllability

Denote $\omega_{kl} = \lambda_k - \lambda_l,\ k, l = 1, ..., N$ as the eigenvalue differences for the matrix A, and atomic units ($\hbar = 1$) will be utilized. Consider the hypothesis:

$\mathbb{TB}$ *The graph G does not have "degenerate transitions", that is for all $(i, j) \neq (a, b)$, $i \neq j$, $a \neq b$ such that $B_{ij} \neq 0$, $B_{ab} \neq 0$: $\omega_{ij} \neq \omega_{ab}$.*

Remark 2 *This hypothesis could be relaxed to requiring only that the graph G remains connected after elimination of all edge pairs $(\Psi_i, \Psi_j), (\Psi_a, \Psi_b)$ such that $\omega_{ij} = \omega_{ab}$ (degenerate transitions). However, to ease of presentation, $\mathbb{TB}$ will be assumed to be true.*

We also introduce one more hypothesis:

$\mathbb{TC}$ *For each $i, j, a, b = 1, ..., N$ such that $\omega_{ij} \neq 0$: $\frac{\omega_{ab}}{\omega_{ij}} \in Q$, where Q is the set of all rational numbers.*

Remark 3 *Alternative controllability results completely excluding the need of the assumption $\mathbb{TC}$ are also possible and will be presented in a future paper.*

The main controllability result in [16] can be summarized as follows:

Theorem 1 *Under the assumptions $\mathbb{TA}$, $\mathbb{TB}$, $\mathbb{TC}$ the system (2) is controllable, that is for any $\Psi \in S_M(0, 1)$ the set of reachable states from Ψ is $S_M(0, 1)$.*

Remark 4 *Under the assumption $\mathbb{TB}$, the controllability criteria above has very strong uniform properties with respect to the coupling matrix B. Indeed, the only information needed to know is whether B_{ij} is null or not for each $i,j = 1, ..., N$; the* **exact value** *of B_{ij} is not important. Thus, the controllability analysis is generally independent of small errors in the entries of B. Note also that when adding, for example, one more eigenstate to the basis D, the controllability criterion is easy to check for the new system: it is necessary to ensure that the new state is connected through B to at least one eigenstate*

in the old basis and then check that the transition energies thus introduced do not equal other transition energies in the system - non-degeneracy - (see also remark 2).

When $\mathbb{TB}$ is not satisfied, changing the exact values of the entries of the coupling matrix B may transform a system that is not controllable into a system that is controllable; other techniques that allow for assessing the controllability (see the situation presented in Eq. (7) later in section 3) may also be sensitive to changes in coupling matrix entries.

Remark 5 *Theorem 1 is a result complementary to the work in [14] as the settings are different. Thorem 4.2 in [14] is appropriate when controllability on spaces of unitary matrices is under study (e.g., in quantum computing and in general where the Lie group transformation structure is relevant to the system), while theorem 1 above is suitable for assessing wavefunction controllability. Extensions of theorem 1 are available in [21] for the case of multiple independent subsystems (non connected graphs) along the same paradigm.*

A detailed proof of theorem 1 may be found in [16]. Below we go beyond the latter work and demonstrate the physical meaning and applicability of the theorem.

3 Numerical Simulations

Numerical experiments have been undertaken to illustrate the theoretical result above. All of the examples correspond to model systems with an external laser electric field coupled in through a dipole matrix B. The controllability Theorem 1 is not constructive in that its satisfaction does not produce a particular control field. Thus the controlling fields in the examples were computed using a genetic algorithm search procedure. Consider the following model [26] five-level system having internal Hamiltonian and coupling matrices,

$$A = \begin{pmatrix} 1.0 & 0 & 0 & 0 & 0 \\ 0 & 1.2 & 0 & 0 & 0 \\ 0 & 0 & 1.3 & 0 & 0 \\ 0 & 0 & 0 & 2.0 & 0 \\ 0 & 0 & 0 & 0 & 2.15 \end{pmatrix}, \; B = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix}. \tag{6}$$

Prior numerical studies with optimal control calculations hinted that this system might be controllable, but such computations cannot assure a full assessment (c.f., discussion later in this section). The coupling graph of the

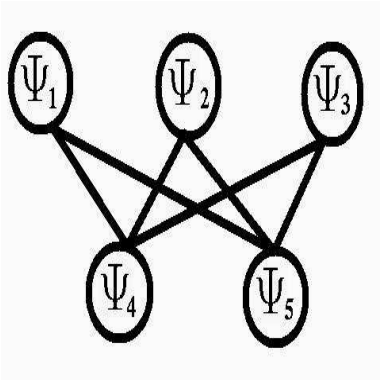


Figure 1: The graph associated with the B matrix of the system in (6). Note that the graph remains connected even after removal of some edges, e.g., (Ψ_3, Ψ_4) and (Ψ_1, Ψ_5).

system plotted in Figure 1 is obviously connected. In addition, it can be easily checked that the system has non-degenerate transitions. It follows by the controllability theorem that this system is completely controllable, implying that any superposition of states is reachable from any other in finite time and with finite laser energy.

An example of control is given in Figure 2; we plot the overlap of the wavefunction with the initial state and the distance to the target state. This situation was choosen to demonstrate control to a superposition of states. The initial state was taken to be Ψ_4 and the target was set to $\frac{\sqrt{3}}{3}\Psi_1 + \frac{\sqrt{6}}{3}\Psi_2$. The target goal is achieved to high accuracy at $T_{final} = 550$.

Although theorem 1 is true only with satisfaction of the hypothesis $\mathbb{TB}$, various situations where $\mathbb{TB}$ at first glance appears to be violated may arise in practice. In this case a simple technique is available to assess if $\mathbb{TB}$ is valid and then return to the setting that accommodates theorem 1. One such example is given below.

Consider the system given by the following Hamiltonian and dipole moment matrix [27, 28]:

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & .004556 & 0 & 0 \\ 0 & 0 & 0.095683 & 0 \\ 0 & 0 & 0 & 0.095683 \end{pmatrix}, \; B = \begin{pmatrix} 0 & 1 & 1 & -1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 \end{pmatrix}. \quad (7)$$

As presented, the system does not comply with $\mathbb{TB}$, being degenerate $\lambda_3 = \lambda_4 = 0.095683 E_h$ and therefore with degenerate transitions e.g. $\lambda_3 - \lambda_1 = \lambda_4 - \lambda_1$. However the states 3 and 4 can be distinguished by having different

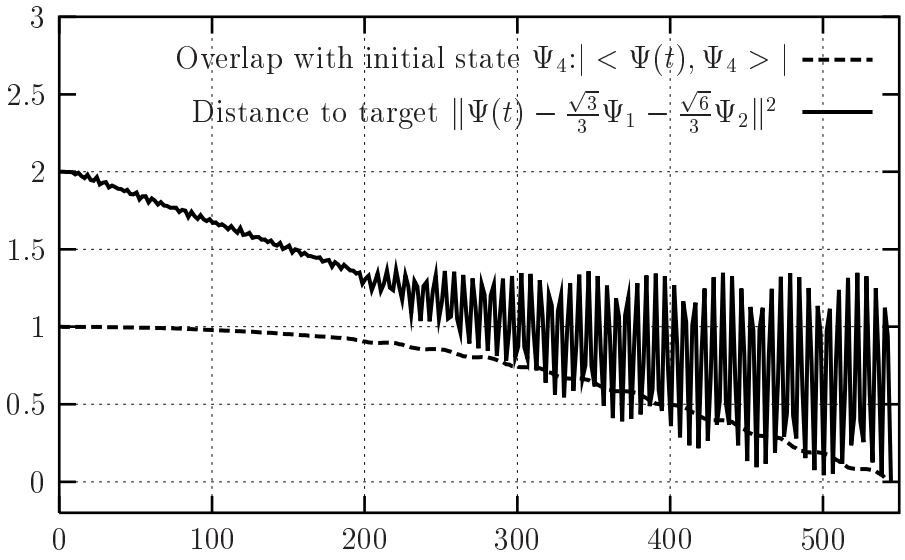


Figure 2: The evolution of the system in (6) under a control field realizing the target: $\Psi(T_{final}) = \frac{\sqrt{3}}{3}\Psi_1 + \frac{\sqrt{6}}{3}\Psi_2$.

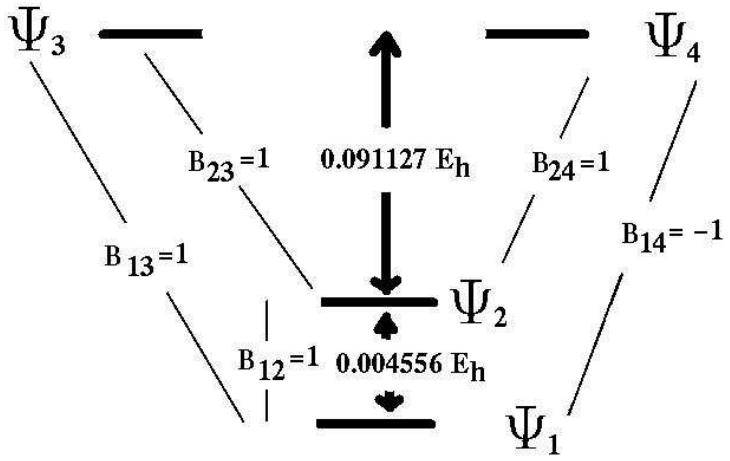


Figure 3: Graphical representation of the system in Eq. (7). It is seen that energy level degeneracy is present.

dipole moments with state 1, and therefore the system is expected to be controllable, as suggested by numerical optimal control calculations [27, 28].

Note that by writing $\epsilon(t) = \mu + \tilde{\epsilon}(t)$ the triplet $(A, B, \epsilon(t))$ that characterizes the control system is transformed to $(A + \mu B, B, \tilde{\epsilon}(t))$. Here $A + \mu B$ is the matrix of the new Hamiltonian $H_\mu = H_0 + \mu\mathcal{B}$. A unitary matrix U_μ may be found such that $\tilde{A} = U_\mu(A + \mu B)U_\mu^t$ is diagonal, and the dipole matrix B changes accordingly $\tilde{B} = U_\mu B U_\mu^t$. The dynamical equations to control are now

$$\begin{cases} i\hbar\frac{\partial}{\partial t}\tilde{C} = \tilde{A}\tilde{C} + \tilde{\epsilon}(t)\tilde{B}\tilde{C} \\ \tilde{C}(t=0) = U_\mu C_0 \end{cases} \tag{8}$$

It can be proven (and it is also trivial to check as soon as the precise value of μ is known) that the number of connected components of the connectivity graph $\tilde{G}$ associated to $\tilde{B}$ is the same as the connected components of G and so, according to hypothesis $\mathbb{TA}$, $\tilde{G}$ is connected. Therefore if $\tilde{A}$ complies with $\mathbb{TB}$ it follows (see also the remark 3) that the system under study is controllable. The controllability of the initial system (2) reduces then to finding μ such that $A + \mu B$ has no degenerate transitions. Many values for μ are often acceptable. The constant μ may be viewed as a Stark field which acts to suitably shift the energy levels so as to remove the degenerate transitions. However, this procedure is just a mathematical construction to reveal if the criteria underlying theorem 1 are valid. The identification of $\mu \neq 0$ does not imply that a laboratory implementation of the control requires a DC bias field to be successful. Satisfaction of $\mathbb{TA}$, $\mathbb{TB}$, $\mathbb{TC}$ just assure that at least some control exists to steer about the system in any arbitrary manner. As an illustration of the procedure above consider the example in Eq. (7) with $\mu = 0.1$ and then the eigenvalues of $A + \mu \cdot B$ are -0.172362, -0.042466, 0.170297 and 0.240453 (non-degenerate). It is easy to check that the eigenvalues also comply with $\mathbb{TB}$. The system (7) is therefore completely controllable. So, despite the degeneracy in the Hamiltonian, it is possible for instance to steer the system from from the state $\Psi(0) = \Psi_1$ to $\Psi(T) = \Psi_4$; such a laser pulse is presented in Figure 4 together with the evolution of the populations of the eigenstates in Figure 5.

In practice, the design of a control is implemented by the computation of a laser pulse that best meets the prescribed goals ; a general approach to executing this search is through the formalism of optimal control theory (OCT) where a cost functional for optimization is constructed that contains penalties for missing the target and various other costs (e.g. the fluence of

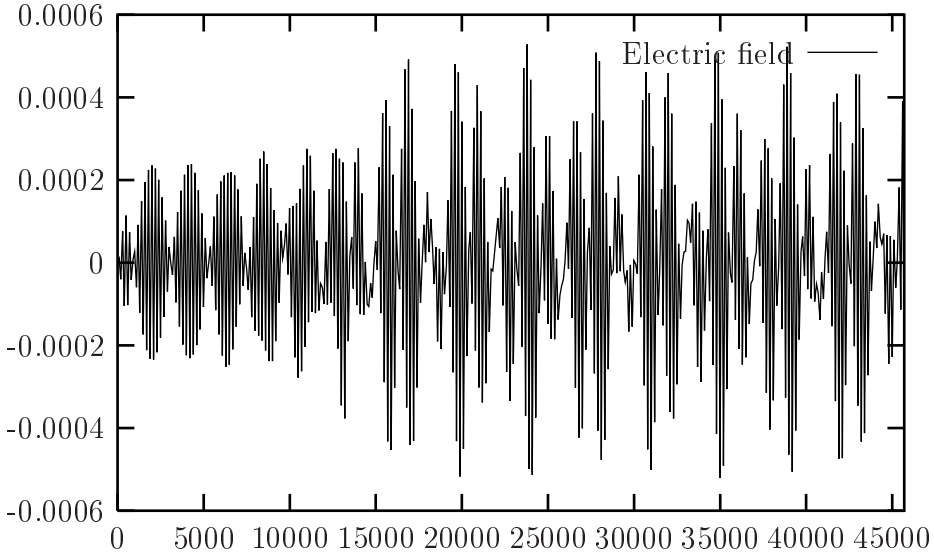


Figure 4: The field realizing the target $\Psi(T_{final}) = \Psi_4$ for the system in (7).

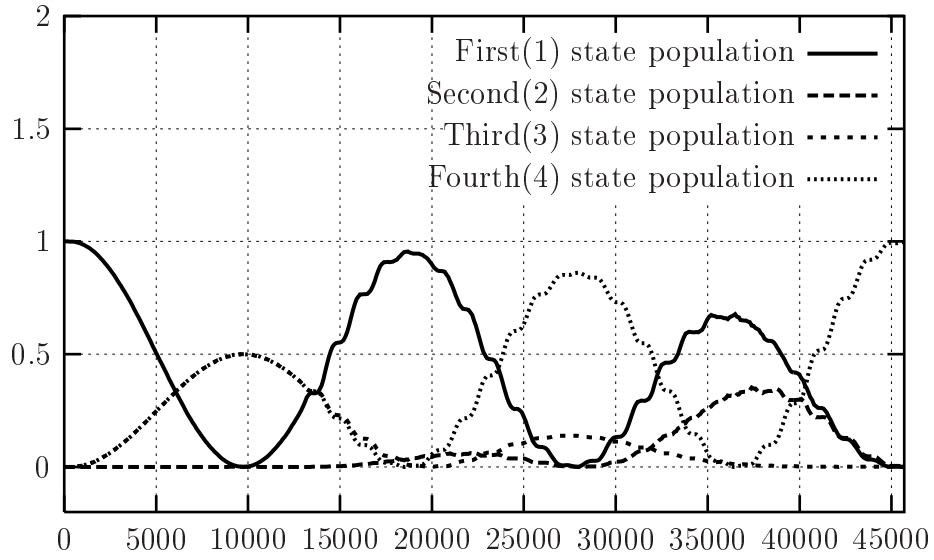


Figure 5: Evolution for the system (7) and field in Figure 4.

the laser). A simple choice for such a cost functional is :

$$J(\epsilon, T) = \|\Psi_\epsilon(T) - \Psi_{target}\|^2 + \alpha \int_0^T \epsilon^2(t)dt \tag{9}$$

It is important to stress that theorem 1 (as any other **exact** controllability theorem) does not guarantee quantitative results for the minimization of $J(\epsilon, T)$, but only insures that in the absence of costs beyond reaching the target state (e.g., the fluence term in Eq. (9)) the minimum value of $J = 0$ **can** be reached for some $T > 0$ and $\epsilon_0(t) \in L^2([0, T])$. An analysis of the existence of at least one minimiser to a class of quantum mechanical OCT cost functionals is given in [6]. The trade-off between the two extremes of fully reaching the target state versus fully meeting the additional cost criteria is the task of the OCT optimization. In this framework, any field that gives exact controllability is a minimizer of $J(\epsilon, T)$ for $\alpha = 0$. When other values for α are chosen, the fluence generally will be smaller but the overlap with the the target will also be reduced. In the example of Figures 4 and 5 where the target was required to be exactly reached, the laser fluence was 0.0302. In another example (not shown here) the fluence term was retained in the OCT cost functional and an overlap with the target of 80% was achived. The optimal field was found to reduce the fluence to 0.021 at the expense of dropping the yield in the target state.

4 Dynamical Conservation Laws and Controllability Restrictions

In light of the manipulations on the system in Eq. (7) an Figure 3 it may seem that the hypothesis $\mathbb{TB}$ has merely a technical role. Therefore a legitimate question to ask is whether theorem 1 remains true in the absence of this assumption. The answer to this question is negative and the presentation of some very particular phenomena that arise when $\mathbb{TB}$ is invalid is the purpose of this section.

We begin with some simple observations. For any Hermitian operator O such that the commutators $[H_0, O]$ and $[\mathcal{B}, O]$ are both zero it is easy to prove that :

$$< \Psi(t)|O|\Psi(t) >=< \Psi_0|O|\Psi_0 >, \ \forall t > 0. \tag{10}$$

The quantity $< \Psi(t)|O|\Psi(t) >$ is therefore **conserved** during the evolution of the system, irrespective of the field $\epsilon(t)$. The presence of any conservation relation on $\Psi(t)$, other than Eq. (4), implies some controllability restrictions.

One special class of Hermitian operators are L^2-projections to closed subspaces. Let P be such a projection to a closed subspace X of $L^2(\mathbb{R}^\gamma)$. The equalities $[H_0, P] = [\mathcal{B}, P] = 0$ mean in particular that X and its orthogonal complement $X^\perp$ are involutive for H_0 and $\mathcal{B}$, i.e.

$$\begin{cases} \forall \Psi \in X: \ H_0\Psi \in X, \ \mathcal{B}\Psi \in X \\ \forall \Psi \in X^\perp: \ H_0\Psi \in X^\perp, \ \mathcal{B}\Psi \in X^\perp \end{cases} \tag{11}$$

The system can then be viewed as decomposed into two **independent** subsystems with wavefunctions $P\Psi$, $(I-P)\Psi$ (the projections of the total wavefunction Ψ to X and $X^\perp$). This decomposition can be further refined for any additional projection operator that commutes with H_0 and $\mathcal{B}$ to obtain a finite number of independent subsystems, each being associated with its L^2-projector P_1,...,P_K such that:

$$\begin{gathered} [H_0, P_i] = [\mathcal{B}, P_i] = 0, \ \forall i = 1, ..., K \\ \begin{cases} P_i P_j = 0, \ \forall i \neq j, \ i, j = 1, ...K \\ \sum_{i=1}^K P_i = I \end{cases} \end{gathered} \tag{12}$$

By using (10) for the projectors P_1,...,P_k one can prove that the system evolves on the product of hyper-spheres S_{Ψ_0}

$$S_{\Psi_0} = \{ f \in L^2(\mathbb{R}^\gamma); \|P_i f\|_{L^2(\mathbb{R}^\gamma)} = \|P_i \Psi_0\|_{L^2(\mathbb{R}^\gamma)}, \ i = 1, ..., K \} \tag{13}$$

Thus, we obtain conditions for controllability : if Ψ is reachable from Ψ_0 then Ψ is necessary in S_{Ψ_0}.

This example shows how the existence of conservation laws for the system introduce restrictions for controllability. For projectors to closed subspaces, the situation lends itself to an easy intuitive understanding. More complicated situations are possible when the conservation law in effect does **not** correspond to a projection and not even to a Hermitian operator. We may see this point through a simple example. Consider the 3-level system:

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}, \ B = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \tag{14}$$

and the corresponding evolution equations

$$\begin{aligned} i\frac{d}{dt}C_1(t) &= C_1(t) + \epsilon(t)C_2(t) \\ i\frac{d}{dt}C_2(t) &= 2C_2(t) + \epsilon(t)C_1(t) + \epsilon(t)C_3(t) \\ i\frac{d}{dt}C_3(t) &= 3C_3(t) + \epsilon(t)C_2(t) \end{aligned}$$

This system has degenerate transitions e.g. $\lambda_2 - \lambda_1 = \lambda_3 - \lambda_2$ and no $\mu \in \mathbb{R}$ can be found such that $A + \mu B$ comply with $\mathbb{TB}$ moreover no (non-trivial) observable O exists that commutes with both A and B. Upon closer examination, a "hidden symmetry" is however found for this system. More precisely it is easy to prove that for any $t > 0$ and $\epsilon(t) \in L^2([0,t])$:

$$|C_1(t)C_3(t) - \frac{C_2(t)^2}{2}| = |C_{01}C_{03} - \frac{C_{02}^2}{2}|. \tag{15}$$

Therefore, if any controllability result is to be true for this setting, it must take into account the conservation law (15) ; any $\Psi(t) = \sum_{i=1}^{3} C_i(t)\Psi_i(x)$ that is reachable from $\Psi(0) = \sum_{i=1}^{3} C_{0i}\Psi_i(x)$ must satisfy the constraint (15). As an illustration of this point consider a simple numerical example. Suppose that the initial state is the ground state (Ψ_1) and the target is the first excited state (Ψ_2). A simple computation gives for Ψ_1 : $|C_{01}C_{03} - \frac{C_{02}^2}{2}| = |1 \cdot 0 - \frac{0^2}{2}| = 0$ and for Ψ_2 : $|C_1(t)C_3(t) - \frac{C_2(t)^2}{2}| = |0 \cdot 0 - \frac{1^2}{2}| = \frac{1}{2}$. Since the two quantities are different, one infers that Ψ_2 **is not reachable from** Ψ_1 and therefore the system is not controllable, despite the fact that the connectivity assumption $\mathbb{TA}$ is satisfied.

A detailed analysis of the case $N = 3$ shows that in each circumstance where the theorem 1 cannot be used, conservation laws are in effect. This leads us to state the following

Conjecture *As long as no new conservation laws appear –besides L^2 norm conservation – the system is controllable, i.e. any state on the unit sphere may be reached (in finite time and with finite energy) from any other.*

The statement above, if true, would have the merit of giving a controllability result independent of the mathematical transcription of the precise control situation (no mathematical properties of the matrices A and B are involved but only properties of the system they describe). When the Lie group corresponding to the Lie algebra generated by the internal Hamiltonian and the coupling matrix is a compact Lie group, a proof that appears to support the conjecture was communicated to us by V. Ramakrishna [29]. In general, it is not known whether the presence of conservation laws *prevents* controllability or only *restricts* the reachable set accordingly.

Remark 6 *Finite dimensional controllability results are only a part of the effort necessary for the theoretical understanding of quantum control problems. One still has to make compatible the* **positive** *results above or elsewhere ([14]) with the generic* **negative** *results for the infinite dimensional systems [20, 15, 21, 23]. The introduction of proper controllability concepts*

seems necessary to make further advances. Furthermore, it is interesting to note that usually when a quantum system is to be controlled the aim is not to precisely obtain a prescribed wavefunction, but rather to ensure that that some useful projections or expectation values have the desired behaviour.

Remark 7 *In the absence of positive infinite dimensional results, controllability conclusions based on some finite discretization should be treated with care. The number of eigenstates considered relevant to the control problem is important, as can be seen from the example in Eq. (14) : when discretized with only two eigenstates, the system is trivially controllable but the introduction of a third eigenstate generates the "hidden symmetry" with its associated loss of controllability. When the system is intrinsically infinite dimensional, the controllability of a low dimensional discretization does not imply the controllability of a larger (and more truthful) discretization involving all states that have important coupling matrix elements with the low dimensional space or domain of interest. As with numerical wave packet modelling calculations, it is suggestive that convergence of controllability conclusions may also occur within the domain of interest as the overall space is expanded in dimension.*

5 Conclusions

Wavefunction controllability of finite dimensional quantum systems interacting with external fields was explored from a practical perspective suggested by recent theoretical results [16]. The criteria presented was seen to be useful for a wide range of problems and very easy to check. Systems with unusual conservation laws that prevent controllability were also presented and the relationship with the theoretical criteria was investigated. Open questions with positive answers in some particular cases were stated as a conjecture. Numerical experiments were undertaken to illustrate the theoretical results and the connection with optimal control theory was discussed. The assessment of controllability is fundamental to the manipulation of quantum systems. Some tools are now available to make this assessment, but a full comprehensive analysis still needs to be developped.

6 Acknowledgements

H.R. acknowledges support from the National Science Foundation and DOD. G.T. thanks Mathieu Pilot from CERMICS (École nationale des ponts et chaussées, Marne-la-Vallée, France), for helpful discussions on this topic. Authors thank V. Ramakrishna for his comments on this paper.

References

[1] C. Le Bris, International Conference on systems governed by PDEs, Nancy, March 1999, ESAIM : Proceedings, vol. 8, 2000, pp 77-94.

[2] P.Brumer and M.Shapiro, Acc. Chem. Res. 22, 12 (1989) 407–413.

[3] M.Demiralp and H.Rabitz, Phys. Rew A., **47** 2 1983, p.831

[4] Kime K., Appl. Math. Lett. 6 (3) (1993) 11–15.

[5] Mei Kobayashi,"Mathematics make molecules dance" , SIAM News 24 (1998)

[6] A.P.Pierce, M.A. Dahleh and H.Rabitz, Phys Rev.A **37** (1988), p.4950

[7] H. Rabitz, R. de Vivie-Riedle, M. Motzkus, and K. Kompa, Science 2000 May 5; 288: 824-828.

[8] Shi S., Rabitz H., Chem. Phys. 97 (1992) 276–287.

[9] S.Shi, A.Woody, and H.Rabitz, J.Chem Phys. 88(1988), p.6870

[10] Tannor D.J., Rice S.A., J. Chem. Phys. 83 (1985) 5013–5018.

[11] W.S.Warren, H.Rabitz and M.Dahleh, Science 259 (1993) 1581–1589.

[12] A. Assion et al., Science vol. 282 (1998) pp. 919-922

[13] R. S. Judson and H. Rabitz, Phys. Rev. Lett. **68**, 1500 (1992)

[14] V. Ramakrishna, et al., Phys. Rev. A 51 (2) (1995) 960–966.

[15] Huang G.M., Tarn T.J., Clark J.W., J. Math. Phys. 24, 11 (1983) 2608–2618.

[16] G. Turinici and H. Rabitz, "Wavefunction controllability in quantum systems", J. Math. Phys., submitted

[17] Depending on the problem, one may need to go beyond this first-order, bilinear term when describing the interaction between the control field and the system, cf. [18, 19].

[18] C.M. Dion et al., Chem. Phys.Lett 302(1999), 215-223

[19] C.M. Dion, A.Keller, O.Atabek & A.D. Bandrauk, Phys. Rew. A 59(2) 1999, p.1382

[20] J.M.Ball, J.E.Marsden and M.Slemrod, SIAM J.Control and Optimization, vol 20 (4) (1982), 575–597

[21] G. Turinici, "Analysis of Numerical Simulation Methods in Quantum Chemistry", Ph.D. Thesis, work in progress

[22] G. Turinici "On the controllability of bilinear quantum systems" in M.Defranceschi, C.LeBris (Eds.), "Mathematical models and methods for ab initio Quantum Chemistry", Lecture Notes in Chemistry, volume 74, Springer, 2000 ISBN: 3-540-67631-7

[23] G. Turinici, "Controllable quantities for bilinear quantum systems" IEEE CDC 2000, Sydney, dec 2000.

[24] A.G. Butkovskiy, Yu.I.Samoilenko, "Control of quantum-mechanical processes and systems", Kluwer,1990

[25] Reinhard Diestel "Graph Theory", 2nd ed. Springer-Verlag, New York, Graduate Texts in Mathematics, Vol. 173, Feb. 2000

[26] S.H. Tersigni, P.Gaspard and S.A. Rice, J. Chem. Phys. 93, 3(1990) 1670–1680.

[27] P.Gross, D. Neuhauser, H. Rabitz, J.Chem.Phys **98** (6) (1993), 4557

[28] M.Q. Phan, H. Rabitz, Chem. Phys. 217 (1997) 389-400.

[29] V. Ramakrishna, private communication, Nov. 2000.

Chapitre 2

Algorithmes génétiques pour le contrôle des systemes quantiques(Genetic algorithms for the control of quantum systems)

We will present in the following a work in progress, in collaboration with Prof. Herschel Rabitz from Princeton University, concerning the practical computation of fields that allow for controlling quantum systems. The purpose of this section is twofold: firstly we will describe practical methods that complement the theoretical results of the previous sections; secondly we will present a "filtering" procedure that allows to avoid "noisy" solutions and hopefully identify general mechanisms of the controlling process having predictive properties.

A quantum control problem is typically transposed in mathematical terms as an optimization procedure defined by a cost functional that contains penalty terms for missing the target and other costs (e.g., the energy of the controlling field). This cost functional is then minimized and the result is the desired control. One of the most efficients algorithms often used for the resolution of this optimization problem is based on a so-called **G**enetic **A**lgorithm (GA). Although the present study is entirely based on this method, general novel paradigms closely related to GA are presently developped under the name of Evolutionary Computation (of which GA is only a branch) that include for instance Evolutionary Programming, Evolutionary Strategies, Genetic Programming methods. We refer the reader to [75] for an overview on

these topics and to Evonet website [71] for pointers and a presentation of the state of the art of the evolutionary methods in european research.

A genetic algorithm works by iterativelly improving the initial "population" of guessed control fields untill certain stopping criteria are met. These techniques require only the evaluation of the cost functional associated to any given field, all the rest of the population dynamics being egulated by a user-selected implementation of the Darwinian natural evolution of a population. It has been observed [53] that quantum systems have a very particular position among the problems that may be solved by the GA paradigm, namely the possibility of realizing practical laboratory implementations of GA. Thus, instead of computing the cost functional by means of computer simulations, the quantum system can be let "solve" its own evolution equation and laboratory measurements on the evolved system are then used to compute the cost [41]. The crucial argument that supports this approach is that such an experiment lasts only a small fraction of a second and therefore a huge number of control fields may be tested in a reasonable amount of time; also important is the fast response of the available optical aparatus that allow for a high duty cycle control field generation.

Although very efficient from many points of view, little is currently understood from the resulting control field concerning the structure and the mechanisms that enter in the control process and thus this practical solution has a limited predictive power. The study of the fields obtained by the GA strategies and the analysis of some control mechanisms are the objectives of this ongoing research.

2.1 Why GA ?

The purpose of this section is to motivate our study from the perspective of the applied mathematicians used to handle classical optimization tools available in the scientific computing.

Although very robust, the GA are generally less efficient than classical optimization methods, e.g. gradient. Accordingly, the main motivation of our choice is not the efficiency but lies in the diversity of the results that GA has access to (e.g. multiple solutions ...) . Let us recall that today, theoretical quantum controllability is an **offline** process with the immediate goal being the **understanding** of how control works [58].

In laboratory, more than 10^3 **different** laser pulses may produced in a second and thus the search for the optimal control field may even be carried out by *brute force* methods (and this remains very cheap since only one molecule is consumed at a time). In order to give this search a structure, GA

are used. The experimental setting is the following:

a) choose a set (population) of initial guesses (laser fields);

b) fire up selected lasers (each on a different molecule) and mesure results;

c) have a computer read these measurements and generate a new population using GA operators;

d) ask the laser machine to create last population individual by individual and mesure results (as in step b);

e) go to step c (or exit if satisfied);

On the contrary, numerical experiments are useful to understand what the reaction mechanisms are all about, rather than to solve a given problem. From this point of view, GA proves useful because it gives **multiple** solutions (so there are more chances to find one that is understandable theoretically), and it allows to introduce complicated concepts in the cost criteria ("avoid this target", nonlinearities, filters ...), whereas the gradient methods are not always such flexible. Moreover gradient like methods output only one solution, the best for the chosen cost functional. The difficulty is therefore transfered to the choice of the cost functional, which is a non trivial task as it is difficult to translate concepts like "most easily understandable theoretically". Indeed, gradient search often gives a result where in order to gain some extra percents in the cost functional, the resulting fielf is so involved that it is hard to decide what is fundamental mechanism and what is only cost functional extra percents gain (if the cost functional is not fully optimized noise is obtained).

On the other hand, understanding how a laser field acts is maybe more important that having a heavily optimized solution because in real world robustness is needed with respect to many external -some of them unknown- factors. GA is indeed more expensive, but here the purpose is neither speed (offline!) nor precision, but robust theoretical understanding of the control mechanisms ready to be generalized to a real world many atoms molecule. Note that the space of the exact wavefunction have such a high dimensionality that neither gradient nor GA may solve or propagate on such a space. From this perspective, the gradient is not anymore the only reasonably choice (as it cannot solve much more problems that GA since dimensions are dramatically increasing). Of course, when speed is the issue, classical optimization methods are the most efficient.

2.2 Introduction to Genetic Algorithms

The GA is a model of learning which derives its behavior from a metaphor of some of the mechanisms of Darwinian evolution in nature. This is done by

Chromosome 1	11011001
Chromosome 2	01011110

TAB. 2.2.1: Chromosomes can be used to code integres or rational numbers. For instance here we may interpret each chromosome as an integer in base 2. Then the number encoded by chromosome 1 is $217 = 1 \cdot 2^0 + 1 \cdot 2^3 + 1 \cdot 2^4 + 1 \cdot 2^6 + 1 \cdot 2^7$.

Chromosome 1	$11011 - 001$
Chromosome 2	$01011 - 110$
Offspring 1	$11011 - 110$
Offspring 2	$01011 - 001$

TAB. 2.2.2: Crossover selects genes from parent chromosomes and creates a new offspring. The simplest way how to do this is to choose randomly some crossover point and everything before this point point copy from a first parent and then everything after a crossover point copy from the second parent. Here "$-$" is the crossover point

the creation within a machine of a *population* of individuals represented by chromosomes, in essence elements of a fixed set of symbols that are analogous to the chromosomes in the DNA. The individuals in the population then undergo a process of simulated "evolution".

In the case of multidimensional optimization problems, the symbols in the chromosome can be used to encode discrete (or discretized) values for the different parameters being optimized.

In practice, the genetic model of computation is implemented by having arrays of bits or characters to represent the chromosomes. The chromosome should in some way contain information about solution which it represents. The most used way of encoding is a binary string. The chromosome then could look like in Table 2.2.1. Any bit in the binary string is also called a gene.

Each chromosome has one binary string and each bit in this string can represent some characteristic of the solution. Usually the whole string represents a number. There are many other ways of encoding and the choice depends mainly on the problem to be solved. For example, one can encode directly integers, real numbers or permutations.

Simple bit manipulation operations allow the implementation of crossover (see Table 2.2.2), mutation (see Table 2.2.3) or other operators on chromosomes.

A central notion for a GA is the fitness concept; it is a value assigned to an individual which reflects how well the individual solves the task at hand,

Original offspring 1	110**1**1110
Mutated offspring 1	110**0**1110
Original offspring 2	01**0**110**0**1
Mutated offspring 2	01**1**110**1**1

TAB. 2.2.3: After a crossover is performed, mutation may take place to avoid that all individuals in in population share too many common paterns. Mutation changes randomly the new offspring. For binary encoding we may switch a few randomly chosen bits from 1 to 0 or from 0 to 1. In this example bit 4 of offspring 1 and bits 3 and 7 of offspring 2 were mutated.

i.e. in our case the optimization problem. The "fitness function" is used to map a chromosome to a real value. Fitness may however contain population dependend terms; for example, when diversity is required from a certain population, two chromosomes with identical cost functional evaluation may have different fitness values depending on how similar they are to the rest of the population.

When the genetic algorithm is implemented it is usually done in a manner that involves the following cycle: evaluate the fitness of (all of the individuals in) the population. Create a new population by performing operations such as fitness-proportionate crossover (or other reproduction methods), and mutation on the individuals whose fitness has just been measured. Discard parts of the old and new population (usually worst fitness individuals) and iterate using the new population. One iteration of this loop is referred to as a generation.

The first generation of this is a population of randomly generated individuals. From there on, the genetic operations, in concert with the fitness measure, operate to improve the population.

2.3 Implementation of a GA search for quantum control problems

Consider a quantum system (isolated from the outer world for the moment) whose internal Hamiltonian is H_0 and let Ψ_0 be its initial state; the dynamics obeys the **T**ime **D**ependent **S**chrödinger **E**quation. Denoting by $\Psi(t)$ the wavefunction of the system at time t, the evolution equations read:

$$\begin{cases} i\hbar\frac{\partial}{\partial t}\Psi(t) = H_0\Psi(t) \\ \Psi(t=0) = \Psi_0,\ \|\Psi_0\|_{L^2(\mathbb{R}^\gamma)} = 1 \end{cases} \tag{2.3.1}$$

The interaction expected to allow for control is some external field of intensity $\epsilon(t) \in \mathbb{R}$ acting on the system through a certain time independent dipole moment operator $\mathcal{B}$. The (controlled) dynamical equations read:

$$\begin{cases} i\hbar\frac{\partial}{\partial t}\Psi(t) = H_0\Psi(t) - \epsilon(t)\mathcal{B}\Psi(t) = H\Psi(t) \\ \Psi_\epsilon(t=0) = \Psi_0 \end{cases} \tag{2.3.2}$$

In a first approximation the goal may be formalized as to find (if any) a final time T and a finite energy field $\epsilon(t)$, $\epsilon(t) \in L^2([0,T])$ able to steer the system from Ψ_0 to some predefined target $\Psi(T) = \Psi_{target}$.

Note that the L^2 norm of Ψ is conserved throughout the evolution:

$$\|\Psi(t)\|_{L^2_x(\mathbb{R}^\gamma)} = \|\Psi_0\|_{L^2(\mathbb{R}^\gamma)},\ \forall t > 0. \tag{2.3.3}$$

Theoretical results concerning the controllability of (2.3.2) have been given in the previous sections or are available elsewhere in the literature [59]. However none of them are constructive, and for this reason finding the appropriate control field for a given system is approached by numerical simulations.

It is rather seldom that an explicit value of T be precisely required or inferred from previous knowledge about the system; T is rather asked to vary into some acceptable interval with no further restrictions. The control problem is therefore formulated as a minimization of some cost functional depending on time and on the external field $\epsilon(t)$. A simple example of such a cost functional is

$$J(\epsilon,T) = \|\Psi(T) - \Psi_{target}\|_{L^2} + \alpha\int_0^T \epsilon(t)^2 dt. \tag{2.3.4}$$

Computer resolution of this optimization problem operates on a discretization of the wavefunction space. Let $D = \{\Psi_i; i = 1,..,N\}$ be an orthonormal basis for a finite dimensional sub-space F of $L^2(\mathbb{R}^\gamma)$ that we are interested in[1], and A and B be the matrices of the operators H_0 and $\mathcal{B}$ respectively, with respect to this base.

In the case of our modeling[1] the A matrix is diagonal and B is symmetrical. Denote by λ_i, $i = 1,..,N$ the diagonal elements of A (the energies of the states Ψ_i). and $\omega_{kl} = \lambda_k - \lambda_l$, $k,l = 1,...,N$ (transition frequencies).

Denote by $C = (c_i)_{i=1}^N$ the coefficients of Ψ_i in the formula of the evolving state $\Psi(t) = \sum_{i=1}^N c_i(t)\Psi_i$. As from now we will work in atomic units only,

[1] This space is given by our model and the functions Ψ_i are usually the first eigenfunctions of H_0 constructed by a prior computation or by a modeling based on observations.

that is $\hbar = 1$; then the equations (2.3.2) read

$$\begin{cases} i\frac{\partial}{\partial t}C = AC - \epsilon(t)BC \\ C(t=0) = C_0 \end{cases} \tag{2.3.5}$$

$$C_0 = (C_{0i})_{i=1}^N,\ C_{0i} =< \Psi_0, \Psi_i > . \tag{2.3.6}$$

For any given final time T and any field $\epsilon(t)$, $\epsilon(t) \in L^2([0,T])$ $J(\epsilon,T)$ can be computed by evolving the state C_0 according to Eq. (2.3.5). The genetic algorithm is used therefore to minimize $J(\epsilon,T)$.

The implementation of the GA search works with the field $\epsilon(t)$ in a Fourier form. First, the user has to define a list of prefered frequencies f_1,...,f_a. Due to some basic physical intuition, this list is usually a selection of elements ω_{ij}, $i,j = 1,...,N$, but this is not mandatory. Any ω_{ij} identified is given the corresponding label (see table 2.3.4). User is then asked to specify for each frequency f_i if sin form $\sin(f_i t)$ ("1") or cosin form $\cos(f_i t)$ ("-1") is prefered (frequencies may repeat, so both may be selected). In order for the transition from a null field to an oscillating field as above to be smooth, an integer P_i is required for each such frequency f_i to stand for the number of periods (in terms of the frequency f_i) that this transition is supposed to last. In practice the factors $\sin(f_i t)$ or $\cos(f_i t)$ are multiplied by

$$t_i(t) = \begin{cases} 0,\ if\ t \le 0 \\ \sin^2(\frac{f_i}{4P_i}t)\ if\ 0 \le t \le \frac{2\pi P_i}{f_i} \\ 1,\ if\ t \ge \frac{2\pi P_i}{f_i} \end{cases} \tag{2.3.7}$$

For each frequency an additional list of maximal amplitudes is also required. These amplitudes will be multiplicative factors $x_i(t)$ in front of $\sin(f_i t)t_i(t)$ or $\cos(f_i t)t_i(t)$. Each $x_i(t)$ is a piecewise constant function and the number of pieces is also a user selected variable. (called *number of switches* in table 2.3.4). The field is the sum over all chosen frequencies f_1,...,f_a of terms $x_i(t) \cdot \sin(f_i t)t_i(|t - S_i^t|)$ or $x_i(t) \cdot \cos(f_i t)t_i(|t - S_i^t|)$, where for any time t, S_i^t is the time of last switch between two constant pieces of $x_i(t)$ or the time of the next switch (from t) if this is closer than P_i periods).

A model for an input (as it is understood and recognized by the program) is given in the table 2.3.4 and Fig 2.1.

The GA search was implemented using primitives available in [72].

The contruction of the function to minimize has to take into account the presence of multiple solutions to the control problem [46]. Some of these solutions may contain "noisy" terms that are not relevant to control and will prevent intuitive understanding of the control mechanisms. Therefore, some

```
System dimension = 4
Field encoding dimension = 2
Internal Hamiltonian = 0       0       0              0
                       0       .15     0              0
                       0       0       0.8            0
                       0       0       0              0.8
Dipole matrix =        0       1       1       -1
                       1       0       1       1
                       1       1       0       0
                       -1      1       0       0
Frequencies considered= 0.15 0.65
Maximum amptitudes=     0.0025    0.0050
Sinus or cosinus =  1  -1
Number of switches= 2   2
Number of periods=  1   1
Initial time = 0.0
Final time   = 700.0
Best Guess =
sin(1->2;)      Time= +50.0   Ampt= +0.0025
                Time= +425.0  Ampt=  0.0000
cos(2->3;2->4;) Time= +200.0  Ampt= -0.0010
                Time= +575.0  Ampt= +0.0030
```

TAB. 2.3.4: Part of the input for the GA search. A "best guess" for the field is required. The labels in the first column of "Best Guess" are recognized and set by the program; in this case there are 2 frequencies $\omega_{12} = 0.15$ ("sin" form) and $\omega_{23} = 0.65 = \omega_{24}$ ("cos" form). The piecewise function corresponding to the first frequency (labeled $\sin(1->2;)$) is equal to $+0.0025$ in the interval $(+50.0, +425.0)$ and is zero for all other times. The piecewise function corresponding to the 2^{nd} frequency (labeled $\cos(2->3;2->4;)$) is zero till $t = +200.0$, is equal to -0.0010 from $t = +200.0$ untill $t = +575.0$ and is equal to $+0.0030$ from this value untill the final time 700.0. See Fig. 2.1 for a graphical representation of this field.

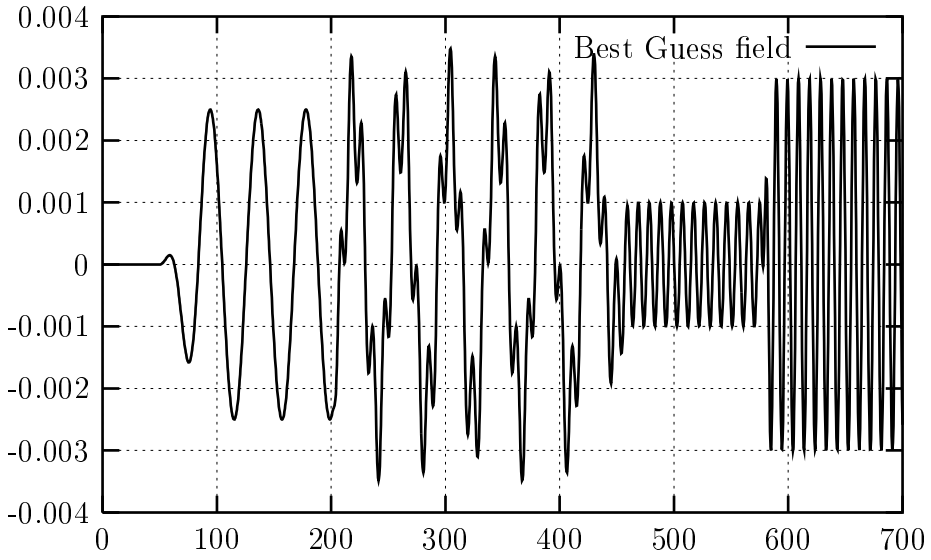


FIG. 2.1: The graphical representation of the "Best Guess" field described in Table 2.3.4.

filtering is to be enforced to eliminate the components that do not actively contribute to steering to target. Indeed, as we will see in the section 2.4, this procedure helps to make understandable solutions that are difficult to analyse in a rough format.

An important feature of the GA is that the search mechanism does **not** depend on the exact form of the cost functional, but rather on how well the individuals are performing. This is very different from the case of classical optimal control theory (OCT) that rely on iterative solving of the first order Euler-Lagrange equations associated to the minimization of $J(\epsilon,T)$ (the differential of $J(\epsilon,T)$ with respect to ϵ and T is required to be zero). As soon as the cost functional is different from the form in Eq. (2.3.4), the resolution of such a classical OCT problem becomes less efficient. By contrast, for the GA we can, at no additional cost (other than computing $J(\epsilon,T)$), use distance metrics adapted to the objectives.

We will use a cost functional such that when the final state $\Psi(T)$ is far from the target, improvement in the distance $\|\Psi(T)-\Psi_{target}\|$ be preffered; on the contrary, when this distance is small, the energy ($\int_0^T \epsilon^2(t)dt$) of the field should become important. In particular, when $\|\Psi(T)-\Psi_{target}\|$ drops below a certain threshold (e.g. 1%) target is considered reached and improvement is only seek in the energy of the field. So, rather that choosing a cost functional J linear in the square of the distance to target $d = \|\Psi(T)-\Psi_{target}\|$ and

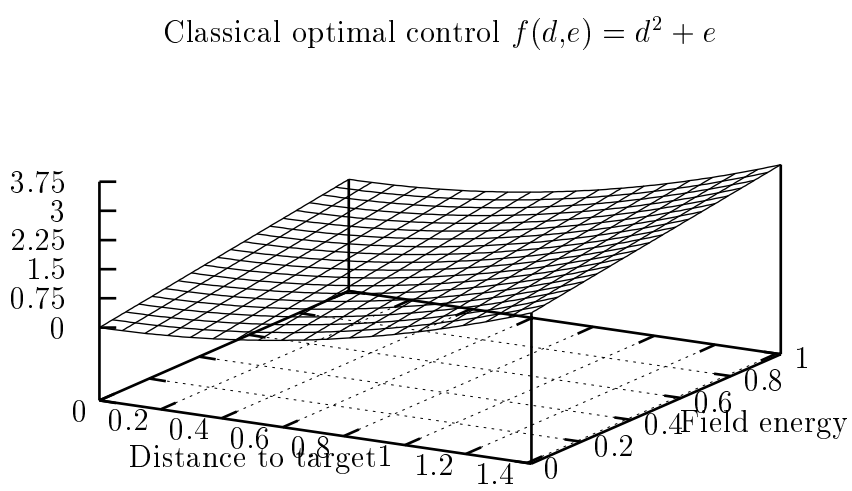


FIG. 2.2: Typical cost OCT functional (2.3.4) in terms of the distance to target and field energy for $\alpha = 1$. Note that the penalizations on the energy is uniform for all distances to target.

energy of the field $e = \int_0^T \epsilon^2(t)dt$ (Figure 2.2) we prefer a function (Figure 2.3) of the form:

$$f(d,e) = (\sqrt{2} - d)(c \cdot e - 1) + \sqrt{2} \tag{2.3.8}$$

which is decreasing when d,e are decreasing The parameter "c" a measure of how penalizing has to be large values of "e" when "d" is close to 0. In fact, in order to implement the mentionned threshold, variable d is multiplied with some cutoff function to set it to zero for (user-defined) small values.

In both cases (2.3.4) and (2.3.8) minimizing the energy is obtained by using a cost functional with a high value of α or c, see Fig. 2.4 and 2.5.

2.4 Numerical results

Numerical simulations have two objectives. Firstly, we aim at testing how well the (parallel) code is performing to produce fields that can exactly control a given initial state to an arbitrary target. In the second stage, where understanding of the control mechanism is sought after, we test the filtering procedure.

Consider the following model [63] five-level system having internal Ha-

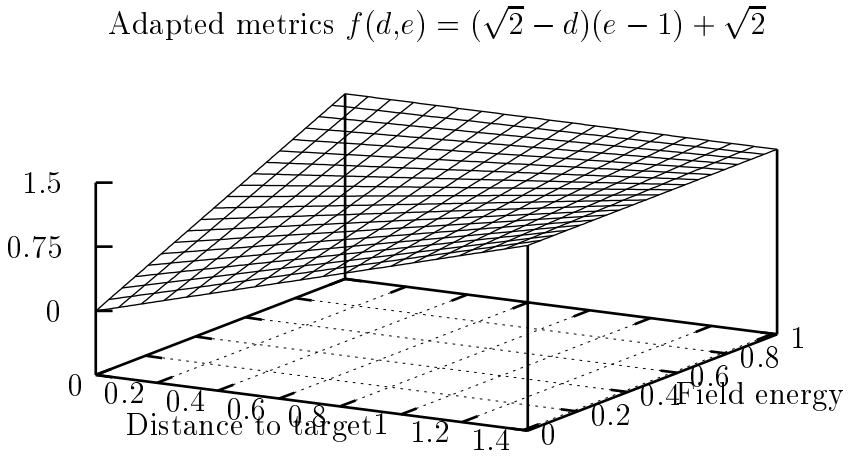


FIG. 2.3: Adapted cost functional (2.3.8) in terms of the distance to target and field energy for $c = 1$. Note that the relative weight of the penalization on the energy is larger as distance to target (d) decreases.

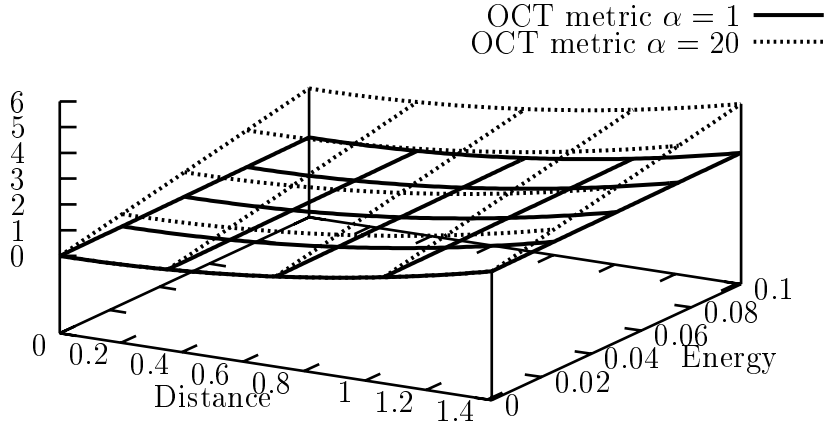


FIG. 2.4: Plot of two surfaces for the classical OCT metric in (2.3.4). Note that when increasing α, the optimization procedure will loose efficiency trying to optimize the energy for large values of the distance.

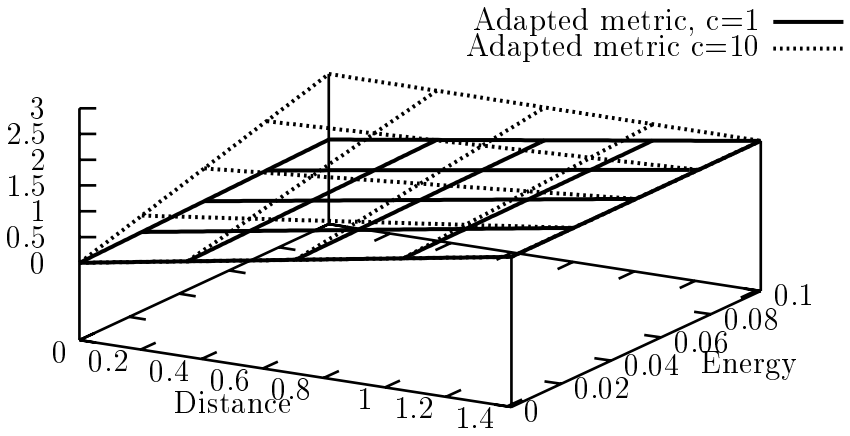


FIG. 2.5: Plot of two surfaces for the adapted metric in (2.3.8). Note that when increasing c, the optimization procedure will hopefully not spend time optimizing the energy for large values of the distance, but will rather try to arrive closer to target in the first place.

miltonian and dipole moment matrices,

$$A = \begin{pmatrix} 1.0 & 0 & 0 & 0 & 0 \\ 0 & 1.2 & 0 & 0 & 0 \\ 0 & 0 & 1.3 & 0 & 0 \\ 0 & 0 & 0 & 2.0 & 0 \\ 0 & 0 & 0 & 0 & 2.15 \end{pmatrix}, B = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix}. \tag{2.4.9}$$

It was seen in the previous chapters by application of the theoretical results that this system is controllable.

An example of control is given in the following.

The initial state is Ψ_1 and the target is Ψ_2. The target is reached by using state Ψ_4 as intermediary as there is no direct connection between Ψ_1 and Ψ_2. Other examples of control for the same system are given in the previous chapters. Since the purpose is to find a controlling field realizing the target here we set $c = 0$ in (2.3.8) (energy is not optimized).

Another interesting system to control is a situation where degenerate

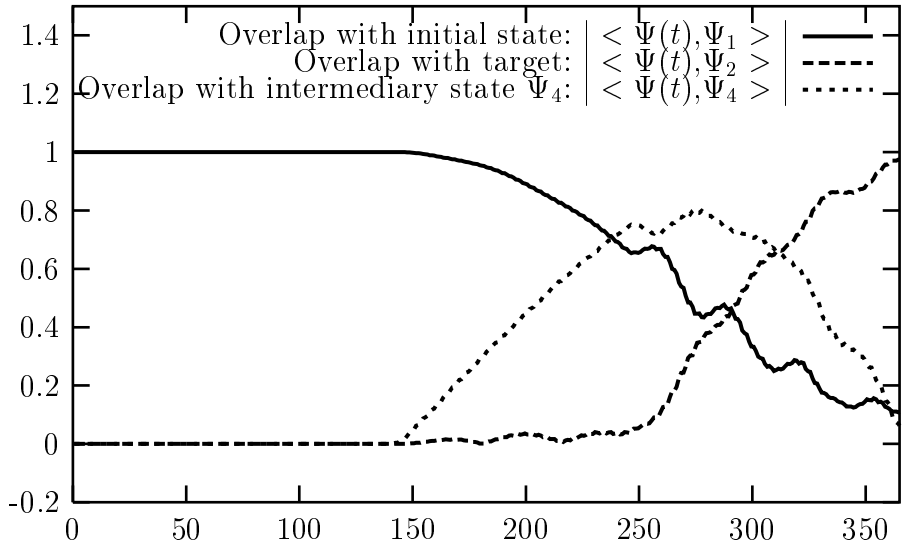


FIG. 2.6: The evolution of the system (2.4.9) for the optimal field found by the GA. Target is reached at time $T = 364.5$

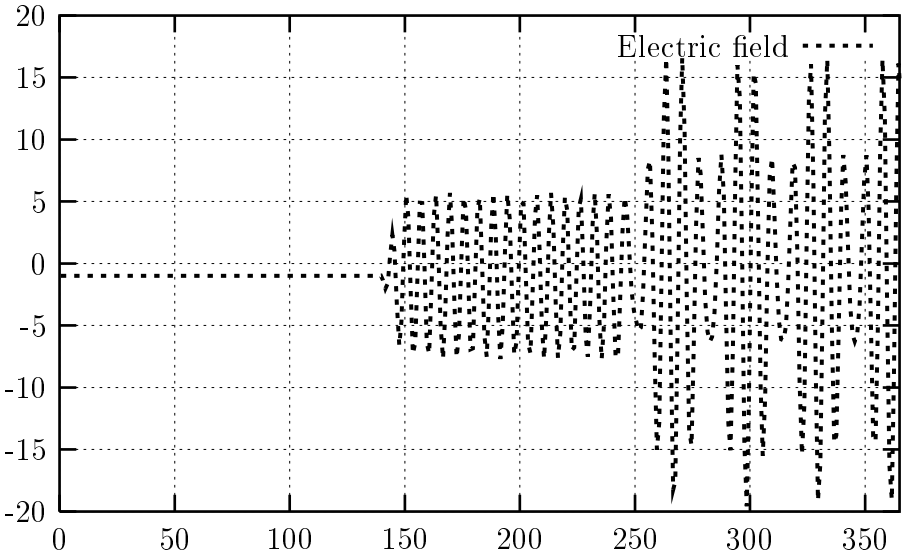


FIG. 2.7: The field used to reached the target in Figure 2.6

```
Best Guess=
cos(1->4;) Time= +138.33 Ampt= +0.01678
cos(2->4;) Time= +238.57 Ampt= -0.03000
```

FIG. 2.8: The exact formulas for the field plotted in Figure 2.7

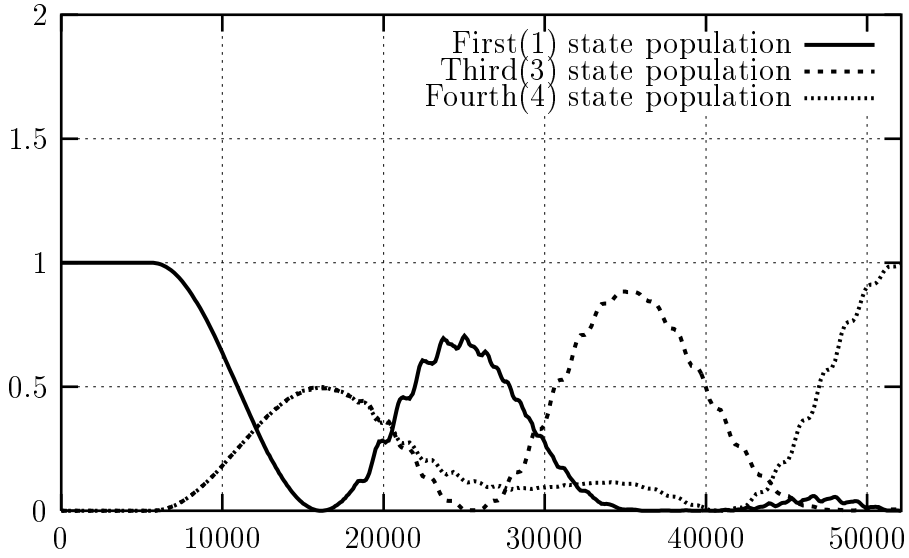


FIG. 2.9: The evolution of the system (2.4.10) for the optimal field found by the GA. Target is reached at time $T = 52113.3$.

eigenstates are present. Let us consider for instance the system:

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & .00455 & 0 & 0 & 0 \\ 0 & 0 & 0.09568 & 0 & 0 \\ 0 & 0 & 0 & 0.09568 & 0 \\ 0 & 0 & 0 & 0 & 0.14124 \end{pmatrix}, B = \begin{pmatrix} 0 & 1 & 1 & -1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}. \tag{2.4.10}$$

The states 3 and 4 have the same energy but are distinguished by a different coupling with state 1. The system is proven to be controllable by the same technique used in chapter 1.2 for a similar degenerate system. The initial state is taken as $\Psi(0) = \Psi_1$ and the goal is to populate the 4th state, i.e. to maximize $|C_4(T_{final}|^2 = | < \Psi(T_{final}), \Psi_4 > |^2$. The best individual given by the GA search is described in Figures 2.9 and 2.10 and Table 2.11.

Although field in Fig. 2.10 is reaching the target, there is not much that can be learned from the evolution in Fig. 2.9. A filtering procedure is thus

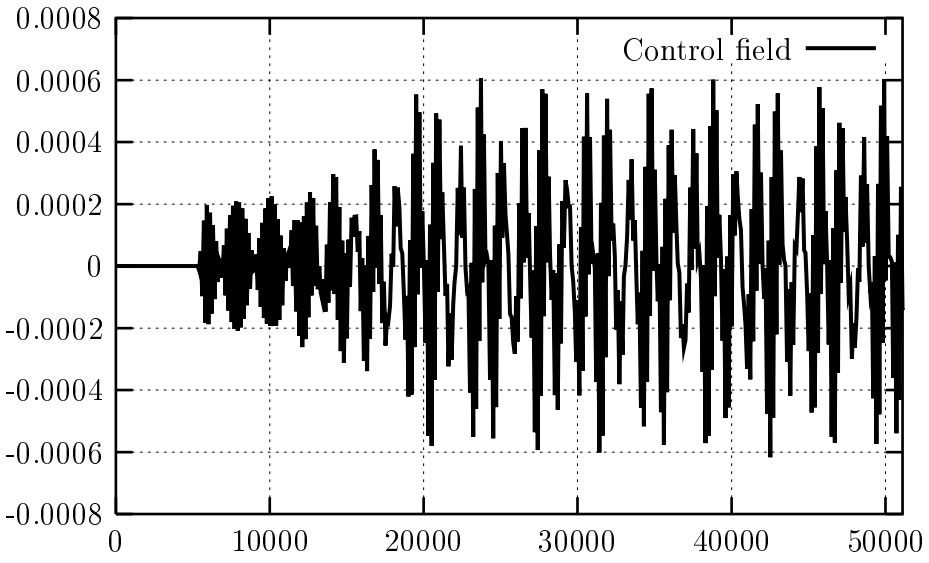


FIG. 2.10: The field used to reach the target in Figure 2.9.

```
sin(1->2;)        Time= +7695.250488281      Ampt= +0.000250000
                  Time= +56698.136718750     Ampt= +0.000250000
cos(1->2;)        Time= +49706.300781250     Ampt= -0.000250000
                  Time= +185184.656250000    Ampt= -0.000250000
sin(2->3;2->4;)   Time= +109382.320312500    Ampt= -0.000250000
                  Time=+159486.546875000     Ampt= +0.000250000
cos(2->3;2->4;)   Time= +17928.259765625     Ampt= +0.000246904
                  Time= +59867.460937500     Ampt= +0.000250000
sin(1->3;1->4;)   Time= +117658.828125000    Ampt= +0.000250000
                  Time= +145578.359375000    Ampt= +0.000250000
cos(1->3;1->4;)   Time= +5299.300781250      Ampt= +0.000209279
                  Time= +160703.984375000    Ampt= +0.000250000
sin(2->5;)        Time= +140905.359375000    Ampt= -0.000250000
                  Time= +198833.390625000    Ampt= +0.000250000
cos(2->5;)        Time= +59597.402343750     Ampt= -0.000250000
                  Time= +61023.167968750     Ampt= -0.000250000
sin(3->5;4->5;)   Time= +117505.625000000    Ampt= -0.000250000
                  Time= +182682.765625000    Ampt= -0.000250000
cos(3->5;4->5;)   Time= +102031.703125000    Ampt= +0.000250000
                  Time= +190480.453125000    Ampt= -0.000250000
```

FIG. 2.11: The exact formulas for the field plotted in Figure 2.10.

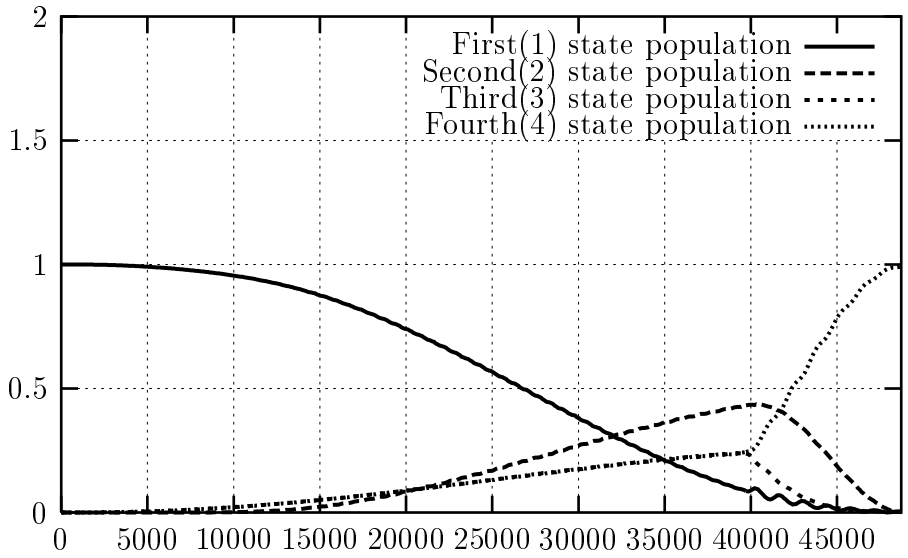


FIG. 2.12: The evolution of the system (2.4.10) for the optimal field found by the GA when $c = 10$. Target is reached at time $T = 48719.9$.

necessary if any general understanding is to be extracted from this example. We have run another GA search, this time with a large constant c in (2.3.8), that is $c = 10$. The best individual given by the GA search is described in Figures 2.12 and 2.13 and Table 2.14. This time the population flow is more simple, and we are thus able to identify a path-interference control [44] situation. As it can be seen from Fig. 2.12 this control mechanism has two phases: a "preparation" phase till $t = 39523.1$ (states 3 and 4 have same population) and an "active" path-interference phase that allow to selectively control population in state 4.

We will close this section by presenting on an example multiple solutions that can be obtained using GA for a given control proble

Consider the system in Eq. (2.4.10). The initial state is set to $\Psi(0) = \Psi_1$ and the goal is, as before, to populate the 4th state.

An automatic "classification" procedure was developed to impart all individuals generated by the GA evolution in classes that share common control mechanism characteristics. Two fields were said to be in the same class (i.e. correspond to the same control mechanism) if the two evolutions of the system for the two fields were close to a certain (user-defined) degree. The Figures 2.15 to 2.22 show, for some of the classes, the evolution of a representative of the class.

Although no interpretations are formally set for the classes presented,

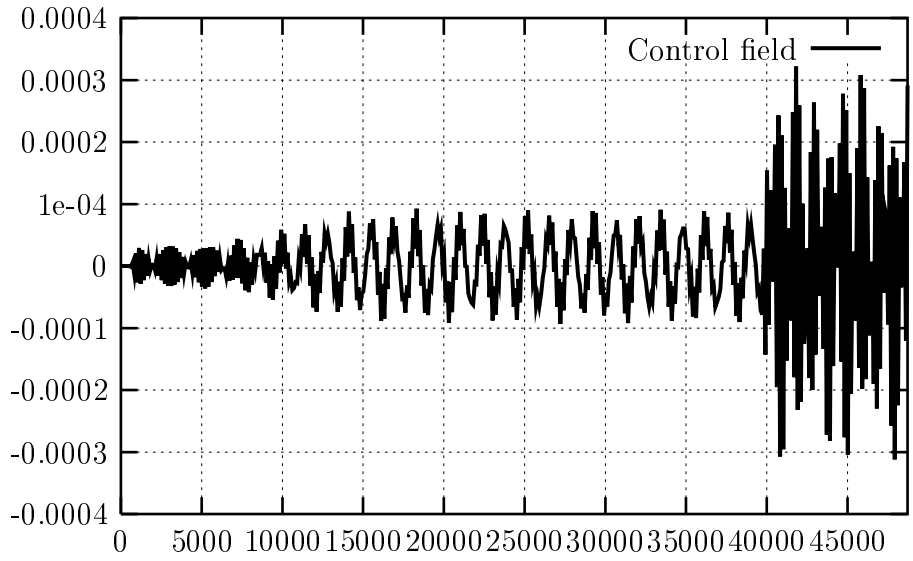


FIG. 2.13: The field used to reach the target in Figure 2.12.

```
sin(1->2;)       Time= +3018.028808594     Ampt= +0.000061445
                 Time= +161632.609375000   Ampt= -0.000250000
cos(1->2;)       Time= +49234.769531250    Ampt= -0.000250000
                 Time= +125710.226562500   Ampt= +0.000250000
sin(2->3;2->4;)  Time= +193640.187500000   Ampt= -0.000250000
                 Time= +193760.406250000   Ampt= +0.000250000
cos(2->3;2->4;)  Time= +39523.199218750    Ampt= -0.000250000
                 Time= +100943.093750000   Ampt= +0.000250000
sin(1->3;1->4;)  Time= +137410.156250000   Ampt= -0.000250000
                 Time= +186754.750000000   Ampt= +0.000250000
cos(1->3;1->4;)  Time= +539.105529785      Ampt= +0.000032386
                 Time= +172056.281250000   Ampt= +0.000250000
sin(2->5;)       Time= +59618.832031250    Ampt= +0.000250000
                 Time= +183306.765625000   Ampt= -0.000250000
cos(2->5;)       Time= +116067.187500000   Ampt= -0.000250000
                 Time= +159165.703125000   Ampt= +0.000250000
sin(3->5;4->5;)  Time= +62793.699218750    Ampt= +0.000250000
                 Time= +183631.109375000   Ampt= +0.000250000
cos(3->5;4->5;)  Time= +57645.250000000    Ampt= +0.000250000
                 Time= +75322.750000000    Ampt= +0.000250000
```

FIG. 2.14: The exact formulas for the field plotted in Figure 2.13.

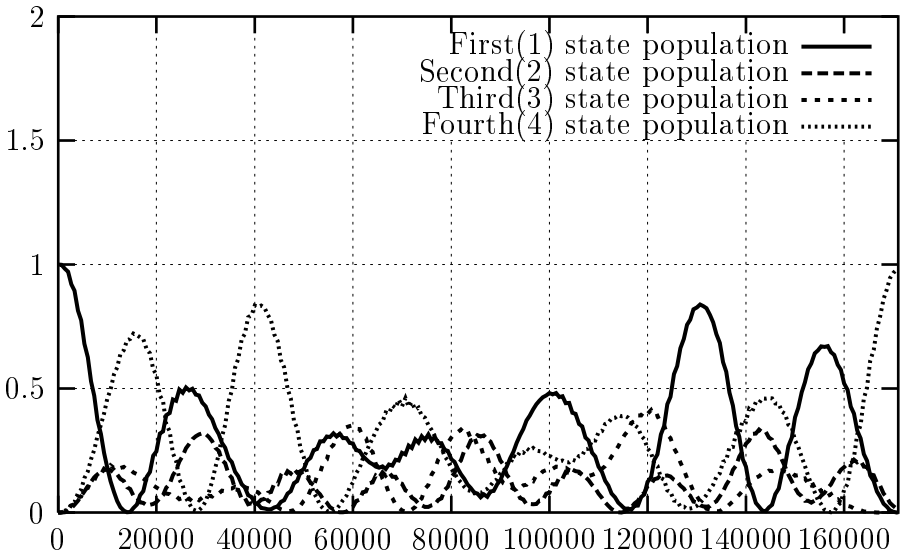


FIG. 2.15: The evolution of the system (2.4.10) for the field in Fig. 2.16. Target is reached at time $T = 171006.6$.

evolution in Fig. 2.15 and 2.17 seems to still contain some noise preventing a straightforward interpretation, Fig. 2.19 seems to correspond to a path-interference mechanism as the one in Fig. 2.12, while evolution in Fig. 2.21 is likely to derive from another, yet unknown, control mechanism. This issue is under study.

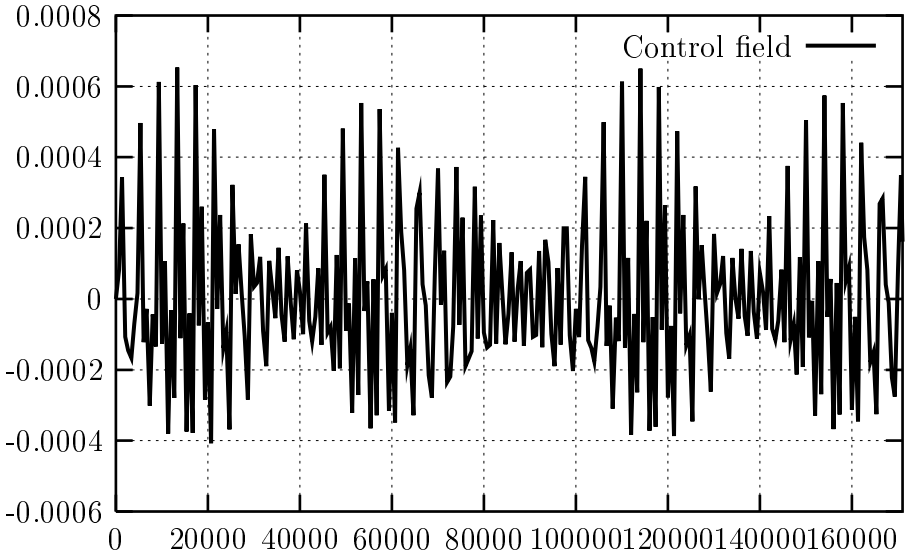


FIG. 2.16: The field used to reach the target in Figure 2.15.

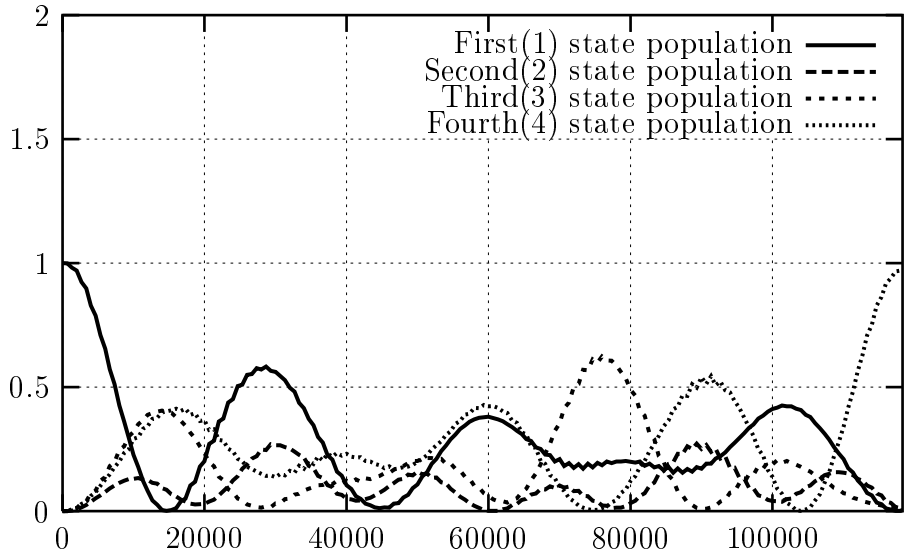


FIG. 2.17: The evolution of the system (2.4.10) for the field in Fig. 2.18. Target is reached at time $T = 118313.3$.

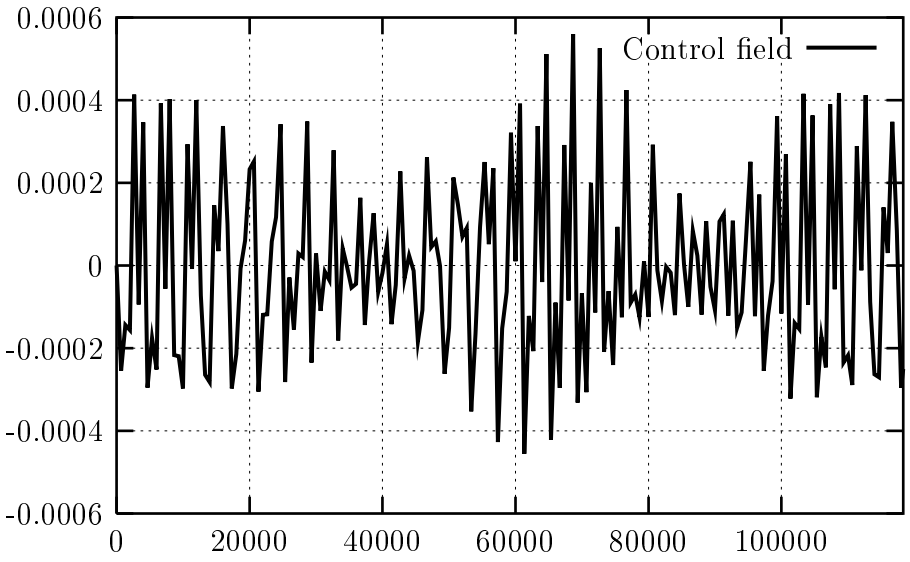


FIG. 2.18: The field used to reach the target in Figure 2.17.

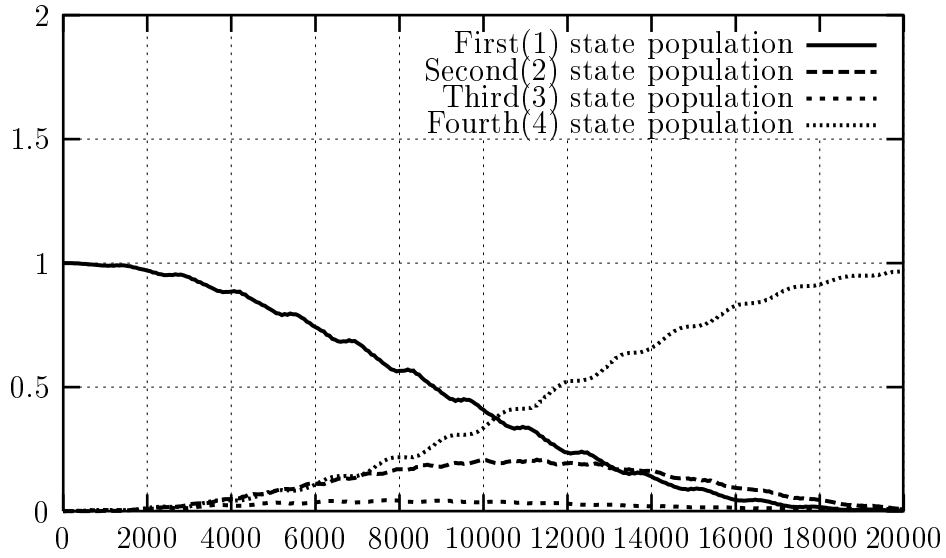


FIG. 2.19: The evolution of the system (2.4.10) for the field in Fig. 2.20. Target is reached at time $T = 19986.6$.

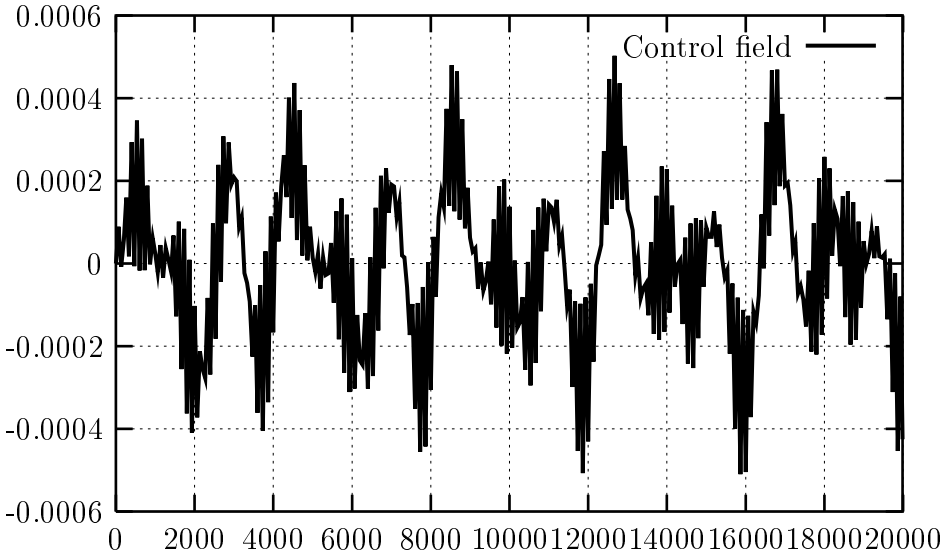


FIG. 2.20: The field used to reach the target in Figure 2.19.

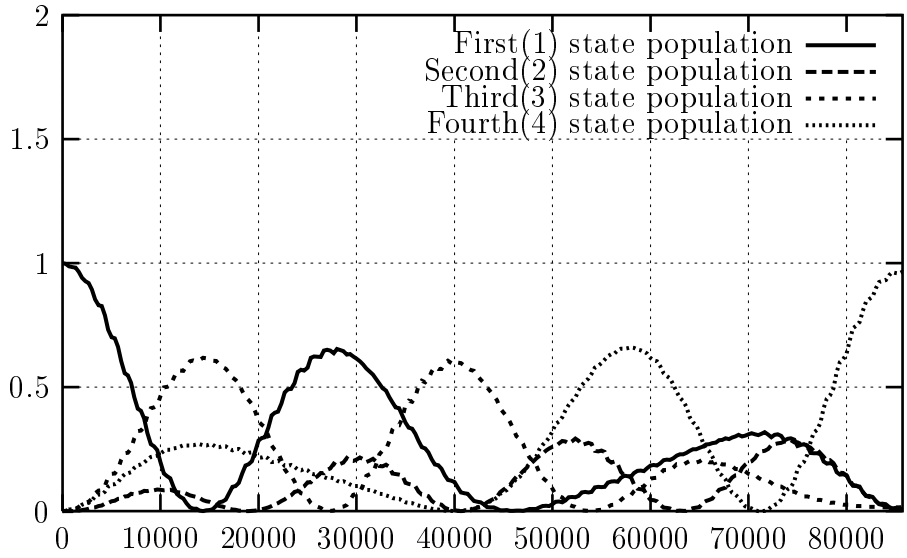


FIG. 2.21: The evolution of the system (2.4.10) for the field in Fig. 2.22. Target is reached at time $T = 85726.6$.

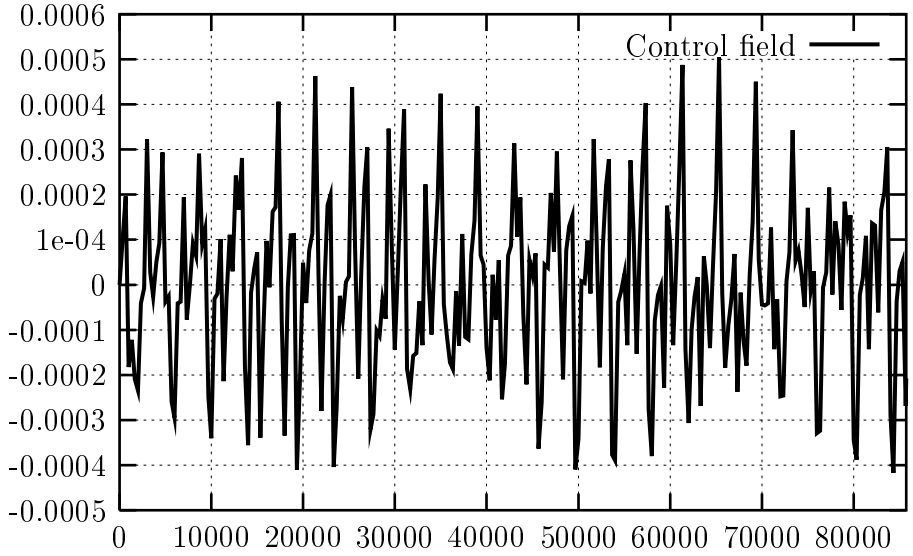


FIG. 2.22: The field used to reach the target in Figure 2.21.

Annexe A

Annexe (Appendix)

A.1 Improvements of controllability results

The aim of this section is to present extensions of the controllability results given in [67] that were obtained in collaboration with Mathieu Pilot from CERMICS (École nationale des ponts et chaussées, Marne-la-Vallée, France) and Herschel Rabitz from Princeton University.

The two extensions studied here concern the elimination of the periodicity hypothesis and the study of general connectivity graphs (i.e. when the number of connected components larger than 1).

The notations and definitions are those of section 1.2.

A.1.1 Elimination of the periodicity hypothesis

Let us remind the periodicity hypothesis:

$\mathbb{IA}$ *For each* $i,j,a,b = 1,...,N$ *such that* $\omega_{ij} \neq 0$*:* $\frac{\omega_{ab}}{\omega_{ij}} \in Q$*, where* Q *is the set of all rational numbers.*

We have seen that $\mathbb{IA}$ implies that the free evolution is periodic i.e. there exists a $T > 0$ such that $U(A,B,0,0 \to T) = e^{-iTA} = I$.

Suppose $\mathbb{IA}$ is not true. Let us remark that due to the finite dimensionality of the system the following quasi-periodicity property is true:

$\mathbb{IB}$ *For each* $\eta > 0$*,* $M > 0$*, there exists* $T_\eta > M$ *such that* $\|e^{-iT_\eta A} - I\| < \eta$.

Indeed, let $T > 0$ and consider the set $\{e^{-i(n \cdot T)A}; n \in \mathbb{N}\}$. Then one of the following alternatives in true:

1. there exist $p \neq q \in \mathbb{N}$ such that $e^{-i(p \cdot T)A} = e^{-i(q \cdot T)A}$;
2. for any $p \neq q \in \mathbb{N}$, $e^{-i(p \cdot T)A} \neq e^{-i(q \cdot T)A}$.

If the first case is true then, supposing $p > q$ we obtain the periodicity: $e^{-i((p-q)\cdot T)A} = I$ so in particular $\mathbb{IB}$ is true with T_η independent of η: $T_\eta = (p-q)\cdot T$. If $T_\eta < M$ choose a multiple of T_η large enough.

If the second case is true, note that all matrices in the set $\{e^{-i(n\cdot T)A}; n \in \mathbb{N}\}$ are unitary, so in particular their euclidian norm is bounded. Then, considering for any $\eta > 0$ the union of balls $B(e^{-i(n\cdot T)A},\eta)$ of radius η centered around each element of the **infinite** set $\{e^{-i(n\cdot T)A}; n \in \mathbb{N}\}$ it is clear there exists at least a pair of balls centered in $e^{-i(p_\eta\cdot T)A}$ and $e^{-i(q_\eta\cdot T)A}$ with $p_\eta - q_\eta > \frac{M}{T}$ having non-empty intersection (otherwise their union will have infinite Lebesque measure, in contradiction with the statement above).

We obtain thus p_η,$q_\eta \in \mathbb{N}$ such that $\|e^{-i(p_\eta\cdot T)A} - e^{-i(q_\eta\cdot T)A}\| \leq \eta$. But since $e^{-i(q_\eta\cdot T)A}$ is unitary it follows $\|e^{-i(p_\eta\cdot T)A} - e^{-i(q_\eta\cdot T)A}\| = \|(e^{-i((p_\eta-q_\eta)\cdot T)A} - I)\cdot e^{-i(q_\eta\cdot T)A}\| = \|(e^{-i((p_\eta-q_\eta)\cdot T)A} - I)\|\cdot\|e^{-i(q_\eta\cdot T)A}\| = \|(e^{-i((p_\eta-q_\eta)\cdot T)A} - I)\|$ which gives the conclusion for $T_\eta = (p_\eta - q_\eta)\cdot T$.

The controllability result Thm. 2 section 1.2 uses the periodicity hypothesis only by the intermediary of the local controllability theorem 1 section 1.2. Therefore, in order to prove that Thm. 2 section 1.2 remains valid in the absence of $\mathbb{IA}$ all that is to be proved is that Thm. 1 section 1.2 remains valid in the absence of $\mathbb{IA}$.

Let us remark that in the absence of $\mathbb{IA}$ the local result reads:

Lemma. A.1.1 *Let $\Psi \in S_M(0,1)\backslash X\backslash Z$, and suppose that the graph associated to the coupling matrix B is connected and has no degenerate transitions. Then, for any $T > 0$ the set of reachable states from Ψ contains a sphere of radius $\epsilon_{T,\Psi}$ (in the canonic metric of $S_M(0,1)$) centered around $e^{-iTA}\Psi$.*

Let then $\Psi \in S_M(0,1)$ be given and find T_0 such that

$$\|e^{-i(T+T_0)A} - I\| \leq \frac{\epsilon_{T,\Psi}}{2}. \tag{A.1.1}$$

Note that by Eq. (A.1.1) $B(\Psi,\frac{\epsilon_{T,\Psi}}{2}) \subset B(e^{-i(T+T_0)A}\Psi,\epsilon_{T,\Psi})$[1]. Consider a target state $y \in B(\Psi,\frac{\epsilon_{T,\Psi}}{2})$ and therefore $y \in B(e^{-i(T+T_0)A}\Psi,\epsilon_{T,\Psi})$, so that $e^{iT_0A}y \in e^{iT_0A}B(e^{-i(T+T_0)A}\Psi,\epsilon_{T,\Psi})$; since the internal Hamiltonian evolution is unitary we obtain

$$e^{iT_0A}B(e^{-i(T+T_0)A}\Psi,\epsilon_{T,\Psi}) = B(e^{-iTA}\Psi,\epsilon_{T,\Psi}).$$

By lemma $\mathbb{IA}$ it follows that $e^{iT_0A}y$ is reachable from Ψ ; but y is reachable from $e^{iT_0A}y$ by the free evolution (for final time equal to T_0) so we conclude that y is reachable from Ψ, which proves the following local result:

[1] We denote by $B(x,r)$ the ball of center x and radius r in the canonical metric of the finite dimensional state space.

Theorem. A.1.1 *Let $\Psi \in S_M(0,1) \setminus X \setminus Z$, and suppose that the graph associated to the coupling matrix B is connected and has no degenerate transitions. Then the set of reachable states from Ψ is a neighborhood of Ψ (in the canonic topology of $S_M(0,1)$). The same result is true for the reverse system.*

Let us mention for the sake of completedness the global result that can be proved from this local controllability theorem:

Theorem. A.1.2 *Suppose that the graph associated to the coupling matrix B is connected and has no degenerate transitions. Then the system*

$$\begin{cases} i\hbar\frac{\partial}{\partial t}C(t) = (A + \epsilon(t)B)C(t) \\ C(t=0) = C_0 \end{cases} \tag{A.1.2}$$

$$C_0 = (c_{0i})_{i=1}^N,\ c_{0i} =< \Psi(0),\Psi_i >\ ,\ \sum_{i=1}^N |c_{0i}|^2 = 1. \tag{A.1.3}$$

is controllable, that is for any $\Psi \in S_M(0,1)$ the set of reachable states from Ψ is $S_M(0,1)$; the same result is true for the reverse system.

A.1.2 Comments on the extension of controllability results for general connectivity graphs

We will give in this section some indications on how theoretical controllability results may be obtained for general control situations where the graph associated to the coupling matrix B is not connected.

Let us recall (Lemma 1 from [67]) the necessary conditions for general graphs with (connected) components $G_\alpha = (V_\alpha,E_\alpha)$, $a = 1,..,K$. Each component G_α of G is an independent subsystem of the initial system in the sense that for any $\alpha \neq \alpha'$ and $\Psi_1 \in V_\alpha$ the coupling of Ψ_1 with any state in $V_{\alpha'}$ is zero.

An important particular case is when some $V_{\overline{\alpha}}$ consists in only one element $V_{\overline{\alpha}} = \{\Psi_{\overline{\alpha}}\}$. In this case $\Psi_{\overline{\alpha}}$ is not coupled to any other eigenstate and therefore the projection $C_{\overline{\alpha}}(t) =< \Psi(t),\Psi_{\overline{\alpha}} >$ of the wavefunction on the linear space spanned by $\Psi_{\overline{\alpha}}$ evolves by the law $C_{\overline{\alpha}}(t) = e^{-i\lambda_{\overline{\alpha}}t}C_{\overline{\alpha}}(0)$ so its evolution **does not depend** on the controlling field $\epsilon(t)$; therefore $C_{\overline{\alpha}}(t)$ is **not controllable**. Such cases have to be excluded. We will suppose in all that follows:

$\mathbb{IC}$ *For all $\alpha = 1,...,K$: cardinality(V_α)≥ 2.*

Denote for $\Psi_0 = \sum_{i=1}^{N} c_i \Psi_i(x) \in S_M(0,1)$:

$$S_{\Psi_0} = \{\chi = \sum_{i=1}^{N} d_i \Psi_i(x) \in S_M(0,1); \sum_{\{i;\Psi_i \in V_\alpha\}} |c_i|^2 = \sum_{\{i;\Psi_i \in V_\alpha\}} |d_i|^2,\ i = 1,...,K\} \tag{A.1.4}$$

Then, if Ψ is reachable from Ψ_0 then $\Psi \in S_{\Psi_0}$. The question is whether under the non-degenerate transitions hypothesis for each $\Psi_0 \in S_M(0,1)$ the set of reachable states is S_{Ψ_0}. The answer is positive, and the proof is made up of two parts:

1. prove the local result: for each $\Phi \in S_M(0,1) \setminus X \setminus Z$ the set of reachable states is a neighbourhood of Φ in the canonical topology of S_Φ;
2. prove the "exit lemma" and the "pass lemma" (see [67]) for this situation.

Here the set X is defined as the set of all states that have the property that at least one projection on the space generated by eigenfunctions corresponding some connected component is in the "X" set of that component.

The proof of the step 1 is just a remake of the proof of the Theorem 1 from [67] combined with the techniques in the lemma A.1.1 above.

The proof of the "exit lemma" adapted to this particular control situation also follows the same lines as the corresponding result (Lemma 2) in [67]. Let us however state the version adapted to this general case:

Lemma. A.1.2 *("exit lemma") For any $\Psi \in S_M(0,1)$ there exists at least one state in $S_\Psi \setminus X \setminus Z$ that can be reached from Ψ; the same is true for the reverse system.*

The only result that remains to be proved is the analogue of the "pass lemma" (Lemma 3 in [67]). Let $\Phi \in S_M(0,1)$. The observation that is crucial to this setting is that in any open (for the canonical topology of $X \cap S_\Phi$) subset V of $X \cap S_\Psi$ there exists an open subset V' of V such that the points of V' belong to the frontier of components of S_Ψ separated by X for only one connected component (independent subsystem) G_α. We obtain then:

Lemma. A.1.3 *("pass lemma") If $X \neq \emptyset$ then, in any given open (for the canonical topology of $X \cap S_\Psi$) subset V of $X \cap S_\Psi$ there exists a "pass state" $\gamma \in V \setminus Z$ such that from γ one can reach at least one point in any (of the two)* **local in** *γ connected components of $S_\Psi \setminus X$ separated by X; moreover these points can be chosen not to be in Z; the same is true for the reverse system.*

A.2 Study of wavefunction controllability of 3-level systems

This section aims at studying the applicability of results in [67, 68] for the case of 3-level bilinear systems. We refer therefore to the papers above for all notations and definitions.

The situation that is of interest to us is $N = 3$. Let then $\Psi_i(x)$, $i = 1,2,3$ be the eigenstates of the internal Hamiltonian and denote by $\Psi(t,x) = \sum_{i=1}^{N} c_i(t)\Psi_i(x)$ the wavefunction of the system. The dynamical equations are (A.1.2- A.1.3).

Remark. A.2.1 *We make the common assumptions that the matrix A is diagonal and that the matrix B is real symmetric (Hermitian) and $[A,B] \neq 0$.*

We denote $\lambda_i \in \mathbb{R}$, $i = 1,..,N$, to be the diagonal elements of A (the energies of the states Ψ_i). Denote $S_M(0,1) = \{C; \sum_{i=1}^{3} |C_i|^2 = 1\}$ as the unit sphere. The conservation of the L^2 norm of the wavefunction can be written in the finite dimensional representation:

$$\sum_{i=1}^{3} |c_i(t)|^2 = 1, \, \forall t \geq 0 \tag{A.2.5}$$

We will start with the analysis of the controllability concept given by the following definition:

Definition. A.2.1 *The system (A.1.2) is called c-controllable if for any $\Psi_1,\Psi_2 \in S_M(0,1)$ at least one state in the class $\{e^{i\varphi}\Psi_2; \varphi \in \mathbb{R}\}$ is reachable from the initial state $\Psi \in S_M(0,1)$.*

Remark. A.2.2 *This concept of controllability is motivated by the fact that overall phase (i.e. complex multiplicative factors of unitary norm) of a quantum states is not an observable. Moreover, any observable related to a quantum state $\tilde{\Psi}$ is constant on the class $\{e^{i\varphi}\tilde{\Psi}; \varphi \in \mathbb{R}\}$.*

Remark. A.2.3 *It is important to remark that this definition of controllability allow to set the trace of A and B to zero. To understand this, consider the evolution equations : $i\hbar\frac{\partial}{\partial t}\Psi_1(t) = (A + \epsilon(t)B)\Psi_1(t)$, and $i\hbar\frac{\partial}{\partial t}\Psi_2(t) = (A + \epsilon(t)B)\Psi_2(t) + \alpha(t)\Psi(t)$ with $\Psi_1(0) = \Psi_2(0)$. Then it is easy to see that $\Psi_2(t) = e^{-i\int_0^t \alpha(\sigma)d\sigma}\Psi_1(t)$ for all $t \geq 0$, so $\Psi_1(t)$ and $\Psi_2(t)$ are in the same class for all times. It follows that substractiong from the matrices A and/or B a term of the form $\alpha(t)I$ (I is the identity matrix) does not change the class of the final state and therefore by the definition A.2.1 does not change the*

c-controllability of the system. In particular substracting from A the matrix $Tr(A)I$ and from $\epsilon(t)B$ the matrix $\epsilon(t)Tr(B)I$ one obtains a (new) system where the matrices involved have zero trace (here "Tr" is the trace operator). So we can suppose $Tr(A) = Tr(B) = 0$.

The aim of this section is to prove that

Lemma. A.2.4 *In all cases where (A.2.5) is the only conservation law* [2] *of the system, the system is controllable.*

Remark. A.2.4 *The presence of conservation laws other that (A.2.5)* **prevents** *controllability [68].*

Proof: The absence of other conservation laws than (A.2.5) implies that the graph associated to the coupling matrix B is connected. So, in agreement with the results in [67, 68], only the non-degenerate transitions hypothesis is to be assured; in fact, by the same argument as in [68] the system is controllable if at least an $\mu \in \mathbb{R}$ can be found such that the eigenvalues of $A + \mu B$ does not give rise to degenerate transitions.

In the case $N = 3$ the presence of degenerate transition mean either two eigenvalues are equal, either one eigenvalue is the arithmetic mean of the two other. Therefore, the only systems that may not controllable are the ones where an interval $[\mu_1,\mu_2]$, $\mu_1 < \mu_2$ exists such that at least one of the following alternatives is true:

1. for all $\mu \in [\mu_1,\mu_2]$ the matrix $A + \mu B$ has two equal eigenvalues
2. for all $\mu \in [\mu_1,\mu_2]$ one eigenvalue of the matrix $A+\mu B$ is the arithmetic mean of the two other.

Denote by λ_i^μ the $i-th$ eigenvalue of the matrix $A + \mu B$; for instance $\lambda_1^0,\lambda_2^0,\lambda_3^0$ are the eigenvalues of the matrix A.

The situation 2 is equivalent to the fact that the following function of μ is null over $[\mu_1,\mu_2]$:

$$P(\mu) = (\frac{\lambda_1^\mu + \lambda_2^\mu}{2} - \lambda_3^\mu) \cdot (\frac{\lambda_2^\mu + \lambda_3^\mu}{2} - \lambda_1^\mu) \cdot (\frac{\lambda_3^\mu + \lambda_1^\mu}{2} - \lambda_2^\mu). \qquad \text{(A.2.6)}$$

Denote $a(\mu) = \frac{\lambda_1^\mu+\lambda_2^\mu+\lambda_3^\mu}{3}$. Then $P(\mu) = \frac{27}{8}(a(\mu)-\lambda_1^\mu)\cdot(a(\mu)-\lambda_2^\mu)\cdot(a(\mu)-\lambda_3^\mu)$. We know that $a(\mu) = \frac{Tr(A+\mu B)}{3} = \frac{Tr(A)}{3} + \mu\frac{Tr(B)}{3}$ which by the remark A.2.3 can be set to zero. So in fact we obtain $P(\mu) = \lambda_1^\mu\lambda_2^\mu\lambda_3^\mu$. But we also have a simple form for this expression: $P(\mu) = \det(A + \mu B)$. It is easy to see that P^μ is a polynomial in μ; since it is zero on the non-trivial interval

[2] i.e. dynamical quantity that is conserved for any external field, see [68] for details

$[\mu_1,\mu_2]$ it will be zero for all μ. In particular for $P(0)=0$; to fix the notations we suppose $\lambda_2 = \frac{\lambda_1+\lambda_3}{2}$; together with $Tr(A)=0$ this gives $\lambda_2 = 0$ and $\lambda_1 = -\lambda_3$, $\lambda_1 \neq 0$, so the matrix A is:

$$A = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\lambda_1 \end{pmatrix}. \tag{A.2.7}$$

In this context the fact that $\det(A+\mu B)=0$ for all $\mu \in \mathbb{R}$ is equivalent to $Q(x) = \det(x \cdot A + B) = 0$ for all $x \in \mathbb{R}$ (use again the polynomial interpretation). A simple computation gives

$$Q(x) = -\lambda_1^2 B_{22} x^2 + \lambda_1(B_{22}(B_{33}-B_{11}) + B_{12}B_{21} - B_{23}B_{32})x + \det(B). \tag{A.2.8}$$

In order for $Q(x)$ to be zero for all x in $\mathbb{R}$ we obtain the following relations: $B_{22}=0$, $B_{12}B_{21} = B_{23}B_{32}$ and $\det(B)=0$, or, since B is symmetrical, $B_{22}=0$, $|B_{12}| = |B_{23}|$, $\det(B)=0$. Note first that if $B_{12}=0$ the matrix B has null elements in second row and second column so the associated graph cannot be connected, which is impossible by hypothesis.

Note that since $Tr(B)=0$ one obtains $B_{11}+B_{33}=0$. Therefore the relation $\det(B)=0$ imply $2B_{12}B_{23}B_{13}=0$ so $B_{13}=0$.

Two situations are possible: $B_{12}=B_{23}$ or $B_{12}=-B_{23}$. In each case conservation laws can be found: when $B_{12}=B_{23}$ one may prove as in [68] that for any $t>0$ and $\epsilon(t) \in L^2([0,t])$:

$$|C_1(t)C_3(t) - \frac{C_2(t)^2}{2}| = |C_{01}C_{03} - \frac{C_{02}^2}{2}|. \tag{A.2.9}$$

For the case $B_{12}=-B_{23}$ the conservation law reads:

$$|C_1(t)C_3(t) + \frac{C_2(t)^2}{2}| = |C_{01}C_{03} + \frac{C_{02}^2}{2}|. \tag{A.2.10}$$

A similar analysis may be carried out for the alternative 1 (two eigenvalues are equal). This time we consider the polynomial

$$R_\mu(X) = (X-\lambda_1^\mu)\cdot(X-\lambda_2^\mu)\cdot(X-\lambda_3^\mu) = X^3 + X(\lambda_1^\mu\lambda_2^\mu + \lambda_2^\mu\lambda_3^\mu + \lambda_3^\mu\lambda_1^\mu) - \lambda_1^\mu\lambda_2^\mu\lambda_3^\mu$$

where we have used the fact that $\lambda_1^\mu + \lambda_2^\mu + \lambda_3^\mu = 0$. Denote $\beta(\mu) = (\lambda_1^\mu\lambda_2^\mu + \lambda_2^\mu\lambda_3^\mu + \lambda_3^\mu\lambda_1^\mu)$ and $\gamma(\mu) = \lambda_1^\mu\lambda_2^\mu\lambda_3^\mu$ so that $R_\mu(X) = X^3 + \beta(\mu)X - \gamma(\mu)$. According to our alternative, for any $\mu \in [\mu_1,\mu_2]$ $R_\mu(X)$ has a root of multiplicity at least 2. It follows that $R_\mu(X)$ and its derivative $R'_\mu(X) = 3X^2 + \beta(\mu)$ have a common factor of first degree. But $R_\mu(X) - \frac{X}{3}R'_\mu(X) = \frac{2\beta(\mu)}{3}X - \gamma(\mu)$

which shows that $\frac{3\gamma(\mu)}{2\beta(\mu)}$ is a root[3] of $R'_\mu(X)$. Writing that $R'_\mu(\frac{3\gamma(\mu)}{2\beta(\mu)}) = 0$ we obtain

$$4\beta^3(\mu) = -27\gamma^2(\mu). \tag{A.2.11}$$

Note that $\gamma(\mu)$ is a third degree polynomial in μ while $\beta(\mu)$ is a second order polynomnial in μ; the fact that this quantities are equal for any $\mu \in [\mu_1,\mu_2]$ imply that this equality is true for any $\mu \in \mathbb{R}$. A closer look at (A.2.11) reveals first that $\gamma(\mu)$ and $\beta(\mu)$ must have the same roots and secondly that the set of roots cannot contain more than one element. This imply in particular that the two roots of $\beta(\mu)$ are equal. To fix the notations we suppose that $\lambda_2^0 = \lambda_3^0 = \lambda$ and therefore $\lambda_1^0 = -2\lambda$. Denote

$$B = \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix}, \; A = \begin{pmatrix} -2\lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}. \tag{A.2.12}$$

Compute $\beta(\mu)$:

$$\beta(\mu) = \det \begin{pmatrix} -2\lambda + \mu a & \mu b \\ \mu b & \lambda + \mu d \end{pmatrix} + \det \begin{pmatrix} -2\lambda + \mu a & \mu c \\ \mu c & \lambda + \mu f \end{pmatrix} +$$
$$\det \begin{pmatrix} \lambda + \mu d & \mu e \\ \mu e & \lambda + \mu f \end{pmatrix}.$$

Note that $Tr(B) = 0$ imply $a = -d - f$. We obtain by simple computations:

$$\beta(\mu) = \mu^2(-d^2 - f^2 - df - b^2 - c^2 - e^2) - 3\mu\lambda(d + f) - 3\lambda^2.$$

The polynomial $\beta(\mu)$ has degenerate roots only when $9\lambda^2(d+f)^2 = 4{\cdot}3\lambda^2(d^2+ f^2 + df + b^2 + c^2 + e^2)$ what is equivalent to $(d - f)^2 + 4(b^2 + c^2 + e^2) = 0$. This can happen only when $b = c = e = 0$ and $d = f$, leading to a diagonal matrix B:

$$B = \begin{pmatrix} -2d & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & d \end{pmatrix}, \tag{A.2.13}$$

which imply $[A,B] = 0$, in contradiction with the hypothesis.

In conclusion, if no conservation laws other than (A.2.5) are in effect, the system is c-controllable.

Let us consider now the problem of (classical) controllability with arbitrary matrices A and B, under the hypothesis formulated in remark A.2.1.

[3] the values of μ where $\beta(\mu) = 0$ are at most 2 and lead to trivial situations

Denote by $A' = A - Tr(A)I$ and $B' = B - Tr(B)I$. We have seen above that if no $\mu \in \mathbb{R}$ can be found such that the transitions corresponding to eigenvalues of $A' + \mu B'$ be non-degenerate, then either the graph associated to matrix B is non-connected, either $[A',B'] = [A,B] = 0$ either conservation laws (A.2.9) or (A.2.10) are in effect.

The eigenvalues of $A' + \mu B'$ are translations of eigenvalues of $A + \mu B$ so the transitions corresponding to eigenvalues of $A' + \mu B'$ are the same as the transitions corresponding to eigenvalues of $A + \mu B$.

By the remark A.2.3, the solution $\Psi'(t)$ of the evolution equation $i\hbar\frac{\partial}{\partial t}\Psi'(t) = (A' + \epsilon(t)B')\Psi'(t)$ differs from $\Psi(t)$ only by a multiplicative constant $e^{i\varphi_t}$, $\varphi_t \in \mathbb{R}$ (provided that $\Psi'(0) = \Psi(0)$). It suffices now to see by the particular form of conservation equations (A.2.9, A.2.10) that if $\Psi'(t)$ comply with one of these conservation laws then also does $\Psi(t)$ (here it is essential that the conservation laws are homogeneous).

In conclusion, if no $\mu \in \mathbb{R}$ can be found such that the transitions corresponding to eigenvalues of $A + \mu B$ be non-degenerate, then either the graph associated to matrix B is non-connected, either $[A,B] = 0$ or conservation laws (A.2.9, A.2.10) are in effect, which ends the proof of lemma A.2.4 $\square$.

Bibliographie générale

[1] **Références générales chimie quantique:**

[2] C. Cohen-Tannoudji, B. Diu et F. Laloë, "Mécanique quantique", en 2 tomes, Hermann 1977

[3] I.N.Levine, "Quantum Chemistry", 4th edition, Prentice Hall 1991

[4] J.L.Rivail "Éléments de chimie quantique", 2^eédition, InterÉditions/CNRS Éditions 1994

[5] A. Szabo, N.S. Ostlund "Modern Quantum Chemistry" Dover, 1996

[6] L.C. Biedenharn, J.D. Louck "Angular momentum in Quantum Physics, Enclyclopedia of Mathematics", vol. 8 (Addison-Wesley), Chap 7.10 et références, 1981

[7] E.B. Wilson Jr., J.C. Decius, P.C. Cross "Molecular vibrations" (Dover), Chap. 2

[8] **Références générales mathématiques**

[9] M. Azaiez, M. Dauge and Y. Maday "Méthodes spectrales et les éléments spectraux" Institut de Recherche Mathématique de Rennes

[10] C. Bernardi and Y. Maday "Spectral methods" in "Handbook of numerical analysis", Ph. G. Ciarlet and J.L. Lions (eds.), North-Holland, vol. V, Part 2, 1997

[11] C. Bernardi and Y. Maday "Approximations spectrales de problèmes aux limites elliptiques" Paris; Berlin ; NewYork NY : Springer, 1992.

[12] R.Dautray et J.L.Lions "Analyse mathématique et calcul numérique pour les sciences et les techniques" tome 5 , Masson , CEA, 1988

[13] J.L.Lions and E.Magenes "Problèmes aux limites non-homogènes et applications", DUNOD, Paris 1968

[14] C.Canuto, M.Y.Hussaini, A.Quarteroni and T.A.Zang "Spectral Methods in Fluid Dynamics" (Springer, Berlin, 1987)

[15] **Techniques a posteriori en mathématiques**

[16] I. Babuška "A posteriori error estimation for the finite element method", Internat. J. Numer. Methods Engrg. **12** (1978), 1597-1615

[17] I. Babuška and C. Schwab "A posteriori error estimation for hierarchic models of elliptic boundary value problems on thin domains" SIAM J. Numer. Anal. Vol **33** (1996), No.1, pp 241-246

[18] C. Bernardi, B. Metivet, "Indicateurs d'erreur pour l'equation de la chaleur. (Error indicators for the heat equation)" Rev. Eur. Elem. Finis 9, No.4, 425-438 (2000).

[19] P.Ladevèze and D.Leguillon "Error estimate procedure in the finite element method and applications" SIAM J.Numer Anal. Vol 20,1991, no 5, 485-504

[20] J.T.Oden and Y.Feng, "Local and pollution error estimation for finite element approximations of elliptic boundary value problems" J.Comput. Appl. Math, **74**, 245-293 (1996)

[21] Y. Maday, A.T. Patera "Numerical analysis of a posteriori finite element bounds for linear-functional outputs", Math. Models Methods Appl. Sci. 10 (2000), no. 5, 785–799.

[22] Y. Maday, A. T. Patera and J. Peraire, "A general formulation for a posteriori bounds for output functionals of partial differential equations; application to the eigenvalue problem", Comptes Rendus de l'Académie des Sciences - Serie I - Mathematiques (328) 9 (1999) pp. 823-828

[23] M. Paraschivoiu, A.T. Patera, "A hierarchical duality approach to bounds for the outputs of partial differential equations", Computer Methods in Applied Mechanics and Engineering 158 (3-4) (1998) pp. 389-407.

[24] M. Paraschivoiu, J. Peraire, A.T. Patera, "A posteriori finite element bounds for linear-functional outputs of elliptic partial differential equations", Computer Methods in Applied Mechanics and Engineering 150 (1-4) (1997) pp. 289-312.

[25] J. Peraire, A.T. Patera, "Asymptotic a posteriori finite element bounds for the outputs of noncoercive problems: the Helmholtz and Burgers equations", Computer Methods in Applied Mechanics and Engineering 171 (1-2) (1999) pp. 77-86.

[26] J. Pousin et J. Rappaz, "Consistency, stability, a priori and a posteriori errors for Petrov-Galerkin methods applied to nonlinear problems" (Report), EPFL, Lausanne, 1992

[27] G. Turinici and Y.Maday, "A posteriori numerical analysis for the Hartree-Fock equations and quadratically convergent methods", to be submitted to Numerische Mathematik

[28] R.Verfürth "A Review of A Posteriori Error Estimates and Adaptative Mesh-Refinement Techniques", Wiley-Teubner 1997

[29] R.Verfürth "A Posteriori Error Estimates For Non-Linear Problems. Finite Element Discretisations of Elliptic Equations" Math. of Comp. **62**, 206(1994), pp 445-475

[30] **Méthode de réduction adiabatique**

[31] G.Brocks and D. Van Koeven "The calculation of the van der Waals vibrational and rotational states of atom-large molecule complexes, with Ar-fluorene as an example" Mol. Phys. 63(6), p. 999-1019, 1988

[32] R.Friesner, J.Bentley, M.Menou and C.Leforestier, "Adiabatic pseudospectral methods for multimensional vibrational potentials " J.Chem.Phys., **99** , 324(1993).

[33] C.Leforestier "Grid representation of rotating triatomics", J.Chem.Phys., **94** , 6388(1991).

[34] K.Yamashita, K.Mokoruma and C.Leforestier "Theoretical study of the highly vibrationally excited states of FHF^-: Ab initio potential energy surface and hyperspherical formulation", J.Chem.Phys., **99** , 8848(1993).

[35] J.Antihainen, R.Friesner and C.Leforestier "Adiabatic pseudospectral calculation of the vibrational states of the four atom molecules: Application to hydrogen peroxide" J.Chem.Phys., **102** , 1270(1995).

[36] D. Kosloff and R.Kosloff "Fourier Method for the Time Dependent Schrödinger Equation as a Tool in Molecular Dynamics" J. Comp. Phys. **vol 52**, 35 (1983)

[37] R.Kosloff "Time-Dependent Quantum-Mecanical Methods for Molecular Dynamics" J. Chem. Phys., **92**, 2087(1988).

[38] Y.Maday et G.Turinici "Analyse numérique de la méthode des variables adiabatiques pour l'approximation de l'hamiltonien nucléaire", C.R.Acad.Sci. Paris, t. 326, série I, p.397-402, 1998

[39] Y.Maday et G.Turinici "Numerical Analysis of the Adiabatic Variable Method for the Approximation of the Nuclear Hamiltonian" , submitted to Mathematical Modelling and Numerical Analysis.

[40] **Références contrôle**

[41] A. Assion et al. "Control of Chemical Reactions by Feedback-Optimized Phase-Shaping Femtosecond Laser Pulses" Science vol. 282 (1998) pp. 919-922

[42] J.M.Ball, J.E.Marsden and M.Slemrod, "Controllability for distributed bilinear systems", SIAM J.Control and Optimization, vol 20 (4) (1982), 575–597

[43] C. Le Bris, "Control theory applied to Quantum Chemistry: Some tracks" , International Conference on systems governed by PDEs, Nancy, March 1999, ESAIM : Proceedings, vol. 8, 2000, pp 77-94.

[44] P.Brumer and M.Shapiro, "Coherence Chemistry: Controlling Chemical Reactions with Lasers", Acc.Chem Res. 22, 12 (1989) 407–413.

[45] A.G. Butkovskiy, Yu.I.Samoilenko, "Control of quantum-mechanical processes and systems", Kluwer,1990

[46] M.Demiralp and H.Rabitz, Phys. Rew A., **47** 2 1983, p.831

[47] Reinhard Diestel "Graph Theory", 2nd ed. Springer-Verlag, New York, Graduate Texts in Mathematics, Vol. 173, Feb. 2000

[48] C.M. Dion et al., Chem. Phys.Lett 302(1999), 215-223

[49] C.M. Dion, A.Keller, O.Atabek & A.D. Bandrauk, Phys. Rew. A 59(2) 1999, p.1382

[50] P.Gross, D. Neuhauser, H. Rabitz, "Teaching lasers to control molecules in the presence of laboratory field uncertainty and measurement imprecision" J.Chem.Phys **98** (6) (1993), 4557

[51] K. Kime, "Control of transition probabilities of the quantum-mechanical harmonic oscillator", Appl. Math. Lett. 6 (3) (1993) 11–15.

[52] G.M. Huang, T.J. Tarn, J.W. Clark, "On the controllability of quantum-mechanical systems", J. Math. Phys. 24, 11 (1983) 2608–2618.

[53] R. S. Judson and H. Rabitz, Phys. Rev. Lett. **68**, 1500 (1992)

[54] M. Kobayashi,"Mathematics make molecules dance" , SIAM News 24 (1998)

[55] M.Q. Phan, H. Rabitz, "Learning control of quantum-mechanical systems by laboratory indentification of effective input-output maps", Chem. Phys. 217 (1997) 389-400.

[56] A.P.Pierce, M.A. Dahleh and H.Rabitz, "Optimal control of quantum mechanical systems: existence, numerical approximation and applications" Phys Rev.A **37** (1988), p.4950

[57] R.Plass et al. "Cyclic Ozone Identified in Magnesium Oxide (111) Surface Reconstructions", Phys. Rev. Lett **81**, pp. 4891-4894, (1998).

[58] H. Rabitz, R. de Vivie-Riedle, M. Motzkus, and K. Kompa "Whither the Future of Controlling Quantum Phenomena?" Science 2000 May 5; 288: 824-828.

[59] V. Ramakrishna, et al. "Controlability of molecular systems" Phys. Rev. A 51 (2) (1995) 960–966.

[60] S. Shi , H. Rabitz, "Optimal control of selectivity of unimolecular reactions via an excited electronic state with designed lasers", Chem. Phys. 97 (1992) 276–287.

[61] S.Shi, A. Woody, and H.Rabitz, "Optimal control of selective vibrational excitation in harmonic linear chain molecules" , J.Chem Phys. 88(1988), p.6870

[62] D.J. Tannor, S.A. Rice, "Control of selectivity of chemical reaction via control of wave packet evolution", J. Chem. Phys. 83 (1985) 5013–5018.

[63] S.H. Tersigni, P.Gaspard and S.A. Rice, "On using shaped light pulses to control the selectivity of product formation in a chemical reaction: An application to a multiple level system", J. Chem. Phys. 93, 3(1990) 1670–1680.

[64] G. Turinici, "Contrôlabilité exacte de la population des états propres dans les systèmes quantiques bilinéaires" C. R. Acad. Sci. Paris, t.330, série I p. 327-332, 2000

[65] G. Turinici "On the controllability of bilinear quantum systems" in M.Defranceschi, C.LeBris (Eds.), "Mathematical models and methods for ab initio Quantum Chemistry", Lecture Notes in Chemistry, volume 74, Springer, 2000 ISBN: 3-540-67631-7

[66] G. Turinici, "Controllable quantities for bilinear quantum systems" 39th IEEE Conference on Decision and Control, Sydney Convention & Exhibition Centre, December 12-15, 2000

[67] G. Turinici and H. Rabitz, "Wavefunction controllability in quantum systems", submitted to J. Math. Phys.

[68] G. Turinici and H. Rabitz, "Quantum wavefunction controllability", in print in J. Chem. Phys.

[69] W.S.Warren, H.Rabitz and M.Dahleh, "Coherent control of quantum dynamics : the dream is alive", Science 259 (1993) 1581–1589.

[70] **Ressources WWW**

[71] Evonet Website:
`http://evonet.dcs.napier.ac.uk/Coordinator/evonet.html`

[72] GAlib: A C++ Library of Genetic Algorithm Components
`http://lancet.mit.edu/ga/`

[73] Action Concertée Incitative "Etude numérique et expérimentale du contrôle des réactions chimiques par laser"
`http://cermics.enpc.fr/equipes/chimie.html`

[74] **Références sur les algorithmes genetiques et evolutionnaires**

[75] Evolutionary Computations and Applications at CMAP
`http://www.eeaax.polytechnique.fr/marc/habilitation.ps.gz`

[76] T. Baeck, "Evolutionary Algorithms in Theory and Practice", Oxford, NY, 1996

[77] Lawrence Davis (editor), "Handbook of Genetic Algorithms", Van Nostrand Reinholt, NY, 1991

[78] D.B. Fogel, "Evolutionary Computation", IEEE Press, NY, 1995

[79] David Edward Goldberg, "Genetic Algorithms in Search and Optimization", Addison-Wesley Pub. Co., 1989 ISBN 0-201-15767-5

[80] J.H. Holland, "Adaptation in Natural and Artificial Systems", 2nd. Edition J.H. Holland, MIT Press, Cambridge, MA, 1992

[81] De Jong, K.A. (1975) "An analysis of the behavior of a class of genetic adaptive systems", Doctoral thesis, Dept. of Computer and Communication Sciences, University of Michigan, Ann Arbor.

[82] J. Koza, "Genetic Programming", MIT Press, Cambridge, MA, 1992

[83] Zbginew Michalewicz, "Genetic Algorithms + Data Structures = Evolution Programs", Springer, Berlin, 1994

[84] H.-P. Schwefel, "Evolution and Optimum Seeking", Wiley, NY, 1994

[85] **Thèses de mathématiques appliquées à la chimie quantique:**

[86] O. Bokanowski, "Formalisation mathématique de la théorie de la fonctionnelle de la densité", Thèse de l'Université Paul Sabatier, Toulouse, 1996

[87] I. Catto, "Analyse mathématique de modèles de la mécanique quantique", Thèse de l'Université Paris IX, 1991

[88] C. Le Bris, "Contribution à l'étude mathématique et numérique de problèmes non linéaires issus des Sciences Physiques : Chimie Quantique Moléculaire et Mécanique des Fluides", Habilitation de l'Université de Paris 9 Dauphine, 1997

[89] Eric Cancès, "Simulation moléculaire et effets d'environnement. Une perspective mathématique et numérique", Thèse de l'École Nationale des Ponts et Chaussées, 1998
URL: `http://cermics.enpc.fr/theses/98/Cances.ps`

[90] J.L. Fattebert, "Une méthode numérique pour la résolution des problèmes aux valeurs propres liés au calcul de structure électronique moléculaire", Thèse de l'École Polytechnique Féderale de Lausanne, 1997

[91] P. Fisher, "Ondelettes et analyse de Fourier dans l'étude d'un problème de chimie quantique", Thèse de l'Université Paris IX, 1994

[92] C. Le Bris, "Quelques problèmes mathématiques en chimie quantique moléculaire", Thèse de l'Ecole Polytechnique, 1993

[93] J-F. Léon, "Étude mathématique de quelques problèmes issus de la physique", Thèse de l'Université Paris IX, 1990

[94] Quelques références de mathématiques appliquées à la chimie quantique

[95] E. Cancès and C. Le Bris, "On the convergence of SCF algorithms for the Hartree-Fock equations", Mathematical Modelling and Numerical Analysis, vol. 34, 4, 2000. Pages: 749-774.

[96] E. Cancès and C. Le Bris, "Can we outperform the DIIS approach for electronic structure calculations?", International Journal of Quantum Chemistry, vol. 79, 2, 2000. Pages: 82-90. (see also typographical erratum, to be published)

[97] E. Cancès and C. Le Bris, "An efficient strategy to solve a nonlinear eigenvalue problem issued from electronic calculations in Quantum Chemistry", in preparation for Journal of Computational Physics.

[98] E. Cancès, C. Le Bris and M. Pilot "Optimal bilinear control for a Schrodinger equation" C.R. Acad. Sci. Paris , t. 330, Série 1, p 567-571, 2000.

[99] C. Le Bris "Some results on the Thomas-Fermi-Dirac-von Weizsacker model", Differential and Integral Equations, vol. 6, 2, p 337-353, 1993.

[100] C. Le Bris, M. Defranceschi, "Computing a molecule: A mathematical viewpoint", Journal of Mathematical Chemistry, vol. 21, 1, p 1-30, 1997.

[101] E.H.Lieb and B.Simon " The Hartree-Fock theory for Coulomb systems" Commun. Math. Phys. 53, 185-194 (1977)

[102] P.L.Lions "Solutions of Hartree-Fock Equations for Coulomb systems", Commun. Math. Phys. 109, 33-97(1987)

Index

(semi-)empiriques, 18
équation de Schrödinger
    dépendant du temps, 20
    indépendant du temps, 21
énergie cinétique, 21
énergie potentielle, 21
état excité, 21
état fondamental, 21
état stable, 21

ab initio, 18

a posteriori, 31
a priori, 31

Born-Openheimer, 22
bornes sur fonctionnelles de la solution, 33
bra-ket, 20
braket, 20

coordonnées de Jacobi, 38

déterminant de Slater, 24
densité électronique, 25

EDP, 31
espace des configurations, 19
estimateur d'erreur, 33

fonction d'onde, 19
fonction d'onde électronique, 22
fonction d'onde nucléaire, 23
formalisme d'Eckart, 38

Hamiltonian, 20
hamiltonien électronique, 22
hamiltonien localisé, 44
hamiltonien nucléaire, 22
hamiltonien réduit, 43
Hartree-Fock, 24

indicateur d'erreur, 33
intervalle de confiance, 35

Lanczos, 40
LCAO, 26

matrice de densité, 25

observable, 19
opérateur de Fock, 25

paire d'électrons de Lewis, 24
population, 146

Restricted Hartree-Fock, 24
réduction adiabatique (pseudo-)spectrale, 37

spin down, 19
spin up, 19

théorie quantique, 17