Development of fast methods for electronic structure calculations

Maxime Barrault

THÈSE

présentée pour l'obtention du titre de

Docteur de l'Ecole Nationale des Ponts et Chaussées

Spécialité : Mathématique, Informatique

par

Maxime BARRAULT

Sujet : *Développement de méthodes rapides pour le calcul de structures électroniques*

Rapporteurs :	François Alouges Frédéric Nataf
Examinateurs :	Eric Cancès Yvon Maday Jean-Louis Vaudescal
Directeur de thèse :	Claude Le Bris

A Amélie, ma famille, mes amis.

Remerciements

Ce travail a été réalisé dans le groupe I23 du département SINETICS de la branche Recherche et Développement d'Electricité de France[1] (EDF), en collaboration avec le Centre d'Enseignement et de Recherche en Mathématiques, Informatique et Calcul Scientifique (CERMICS). Jean-Louis Vaudescal, chef du département SINETICS, m'a accueilli au sein de son équipe. Je le remercie grandement de la confiance qu'il m'a accordée tout au long de ce travail.

Je tiens tout d'abord à témoigner toute ma gratitude à William Hager et Gilles Zerah de m'avoir fait l'honneur de participer avec enthousiasme au jury de soutenance, à François Alouges et Frédéric Nataf d'avoir accepté la lourde tâche d'évaluer avec le plus grand intérêt ce travail et à Yvon Maday d'avoir présidé le jury.

C'est avec une immense joie que j'exprime une profonde reconnaissance à Claude Le Bris et Eric Cancès qui m'ont initié dès mon entrée à l'Ecole Nationale des Ponts et Chaussées aux problématiques numériques et m'ont suivi et offert depuis ce jour les meilleures opportunités afin de m'épanouir dans ce domaine. C'est un réel bonheur de travailler avec ces personnes d'exception que j'admire par leurs qualités humaines, pédagogiques et scientifiques. Il me serait impossible de les remercier à la hauteur des enseignements basés sur la rigueur et l'humilité qu'ils m'ont prodigués, enseignements qui ont été d'indispensables atouts pour appréhender et surmonter les problèmes rencontrés dans ce travail.

Par ailleurs, c'est avec plaisir que j'exprime toute ma sympathie et ma gratitude à Jean-Pierre Daudey, Mireille Defranceschi, Cuong Nguyen, Tony Patera, Bernard Philippe, Gabriel Turinici et une nouvelle fois Yvon Maday au contact de qui je n'ai jamais cessé d'apprendre au cours des stimulantes et fructueuses discussions à chacune de nos rencontres et collaborations.

Une grande part de l'aboutissement de ce travail revient aux personnes qui m'ont cotoyé et soutenu tous les jours à EDF et au CERMICS. C'est pourquoi je remercie très sincèrement Laetitia Andrieu, Anne Auger, Adel Ben Haj Yedder, Xavier Blanc, Adrien Blanchet, Linda Elalaoui, Youstra Gati, Frédéric Legoll, Tony Lelièvre, François Lodier, Patricia Piat, les groupes I23, I26 et I27, Philippe Baranek et Muriel Robin. Je tiens à exprimer ma plus grande gratitude à Guy Bencteux qui m'a témoigné d'un soutien sans faille dans les moments les plus difficiles.

[1] Convention CIFRE N.20010079.

J'ai une pensée particulière pour Sylvie Berte, Jacques Daniel, Marine Danielle, Khadija Elouali, Marie-Claude Mansat, Alice Tran à l'Ecole Nationale des Ponts et Chaussées et Agnès De Cicco, Anna Derbyshire, Didier Lerondel, Liliane Prijac et Philippe Schoenberger à EDF ainsi que Cécile Petit[2] qui m'ont permis d'évoluer dans les meilleures conditions au quotidien.

Enfin, de simples mots ne suffiraient pas pour remercier la générosité de mes proches et de ma famille dans leurs encouragements face aux difficultés inhérentes à un travail de thèse.

[2]suivi de la thèse CIFRE au ministère.

Résumé :

Cette thèse présente quelques idées pour l'accélération des calculs *ab initio* de systèmes physico-chimiques.

Après une introduction générale aux modèles et aux méthodes, faite au chapitre 1, le chapitre 2 est consacré à une présentation mathématique de la construction des pseudo-potentiels qui mènent à une réduction considérable de la taille du problème électronique.

On s'intéresse ensuite au problème aux valeurs propres généralisé qui constitue l'étape limitante de la résolution du problème électronique. On propose dans le chapitre 3 une méthode de décomposition de domaine de complexité linéaire avec le nombre d'électrons du système en terme de temps CPU et d'encombrement mémoire. Cette méthode, adaptée au traitement des systèmes isolants, remédie à certaines insuffisances des méthodes existantes. Dans le même esprit, le chapitre 4 est dédié à une tentative d'adaptation des méthodes dites de projection pour le traitement des gros systèmes métalliques.

Un autre problème est abordé au chapitre 5. Il s'agit de l'application de la méthode des bases réduites au problème électronique. Dans un premier temps, des résultats montrant la faisabilité de l'approche ont été obtenus sur les systèmes H_2^+ et H_2 où la base de discrétisation pour la résolution du problème électronique dépend de la position des noyaux, paramètres du système. Dans un second temps, une adaptation de la méthode des bases réduites pour traiter un problème non linéaire est présentée.

Le chapitre 6 présente enfin des conclusions générales sur l'ensemble des approches abordées dans la thèse, ainsi que quelques pistes pour des développements futurs.

Abstract :

We investigate in this thesis some approaches for the acceleration of electronic structure calculations in *ab initio* simulation of molecular systems.

Chapter 1 is a general introduction to the models and methods. Chapter 2 is dedicated to a mathematical description of the construction of pseudo-potentials which allow a significant reduction of the dimension of the electronic problem.

We then turn to the resolution of the generalized eigenvalue problem which is the bottleneck of the resolution of the electronic problem. In chapter 3, a multilevel domain decomposition method, scaling linearly with respect to the size of the system, is proposed for the treatment of insulators. It overcomes some drawbacks of the existing methods. Chapter 4 presents an attempt to adapt the so called projection methods to the treatment of metallic systems.

Chapter 5 is devoted to the adaptation of the reduced basis method to the electronic problem. First, some preliminary results have been obtained on the systems H_2^+ and H_2, using a Galerkin basis for the resolution of the electronic problem which depends on the positions of the nuclei, parameters of the system. These results de-

monstrate the feasibility of the approach. Second, a strategy for the treatment of a general nonlinear problem is presented.

Finally, some general conclusions about all the approaches studied in the thesis are presented in Chapter 6, along with some tracks for further research.

Sommaire

Publications de l'auteur et communications orales

⋆ Publications de l'auteur

[A1] M. Barrault (2002), note HI-23/2002/020/A interne EDF, *Rapport de synthèse sur le domaine des Pseudo-Potentiels.*

[A2] M. Barrault, E. Cancès, W. W. Hager and C. Le Bris (2005), *Multilevel domain decomposition for electronic structure calculations*, Journal of Computational Physics, submitted in Journal of Computational Physics.

[A3] M. Barrault, Y. Maday, N.C. Cuong and T. Patera (2004), *An "empirical interpolation" method : application to efficient reduced-basis discretization of partial differential equations*, Comptes-Rendus Mathématiques, *339*, Issue 9, 667-672.

⋆ Communications orales à des congrès

[I1] *Toward a general purpose linear scaling method for eigenvalue problems in Quantum Chemistry*, Quatrième journée d'algorithmique numérique appliquée aux problèmes industriels, ULCO, Calais, 15-16 Mai 2003.

[I2] *Toward a general purpose linear scaling method for eigenvalue problems in Quantum Chemistry*, AMAM 2000 Conference, Nice, February 10-13, 2003.

[I3] *Toward a general purpose linear scaling method for eigenvalue problems in Quantum Chemistry*, CANUM 2003, Département de Mathématiques de l'Université de Montpellier II, 2-6 Juin 2003.

[I4] *Toward a general purpose linear scaling method for eigenvalue problems in Quantum Chemistry*, Journée GAMNI-PSMN, ENS Lyon, 05 Décembre 2003.

Chapitre 1

Introduction générale

La compréhension de nombreux phénomènes physico-chimiques conduit à simuler un système moléculaire isolé comportant N électrons et M noyaux de charge $(z_k)_{k=1,M}$ et de masse $(m_k)_{k=1,M}$ sur une période de temps T. Un modèle courant pour ces simulations est le modèle de dynamique moléculaire *ab initio* qui revient à résoudre le problème électronique statique pour un grand nombre de jeux de positions des noyaux $(x_k)_{k=1,M}$. Les méthodes actuelles pour la résolution de ce problème ne permettent pas d'atteindre des ordres de grandeur sur N et T satisfaisants. Cette thèse propose quelques voies pour le développement de méthodes rapides en vue d'aller vers de tels ordres de grandeur.

Dans cette section, on présente le contexte dans lequel s'insère le travail réalisé. Par souci de simplicité, on néglige les effets de spin et on suppose que le système est constitué de N *paires* d'électrons. On utilise dans la suite les unités atomiques

$$m_e = 1, \quad e = 1, \quad \hbar = 1, \quad \frac{1}{4\pi\epsilon_0} = 1.$$

L'ensemble des matrices réelles à k lignes et l colonnes est noté $\mathcal{M}(k,l)$, alors que $\mathcal{M}_S(k)$ est l'ensemble des matrices réelles symétriques de dimension k et I_k est la matrice identité de dimension k et δ_{ij} est le symbole de Kronecker.

Le lecteur pourra se référer à [1] pour une vision mathématique plus détaillée de la chimie computationnelle. Pour les aspects chimiques, il pourra consulter [3, 6, 7].

1.1 Le problème électronique

Les fondements de la chimie quantique associent au système moléculaire considéré une fonction d'onde $\Psi(t,.)$, vivant dans un sous-espace $\mathcal{H}$ de $L^2(\mathbb{R}^d)$ avec $d = 3(M+N)$ dont l'évolution est régie par l'équation de Schrödinger (excepté dans des cas très particuliers). La dimension de $\mathcal{H}$ rend presque toujours le traitement numérique de cette équation impossible. Afin de décrire le système, on fait dans un premier temps l'approximation adiabatique qui conduit au modèle de dynamique

moléculaire : les noyaux atomiques sont des particules classiques relevant d'une dynamique de Newton et si les électrons restent des objets quantiques ils sont supposés être à tout instant dans leur état fondamental (i.e. état de plus basse énergie). Les noyaux sont vus par les électrons comme un potentiel électrique "extérieur" et les électrons sont vus par les noyaux comme un champ de force additionnel qui requiert de résoudre le problème

$$\inf \Big\{ \langle \Psi_e | H_e | \psi_e \rangle,\ \Psi_e \in \mathcal{H}_e,\ \|\Psi_e\| = 1 \Big\} \tag{1.1}$$

où H_e est l'hamiltonien de la partie électronique. Comme $\mathcal{H}_e$ est un sous-espace de $L^2(\mathbb{R}^{3N})$ on fait, dans un second temps, l'approximation de (1.1) par des modèles de type Hartree-Fock [7] ou de type Kohn-Sham [6]. Formellement, on est alors ramené au calcul de N fonctions $\{\phi_i\}_{i=1,N}$ satisfaisant un problème de minimisation appelé problème électronique de la forme

$$\inf \Big\{ \mathcal{E}(\phi_1, \ldots, \phi_N), \phi_i \in H_1(\mathbb{R}^3),\ \int_{\mathbf{R}^3} \phi_i \phi_j = \delta_{ij},\ \forall\ 1 \leq i, j \leq N \Big\} \tag{1.2}$$

La fonctionnelle d'énergie $\mathcal{E}$ dépend paramétriquement de la position des noyaux $\bar{x}_k$ au pas de temps considéré. Lors de la résolution du problème électronique, les forces exercées par le nuage électronique sur les noyaux sont données par $\{\partial \mathcal{E} / \partial \bar{x}_k\}_{k=1,M}$ et utilisées pour calculer au pas de temps suivant la position des noyaux par l'intermédiaire d'un schéma adapté d'intégration des équations de Newton [1] :

$$\forall\ 1 \leq k \leq M, \quad m_k \frac{d^2 \bar{x}_k}{dt^2}(t) = -\frac{\partial \mathcal{E}}{\partial \bar{x}_k} - \frac{\partial \mathcal{W}_{nuc}}{\partial \bar{x}_k}, \quad \mathcal{W}_{nuc} = \sum_{1 \leq i < j \leq M} \frac{z_i z_j}{|\bar{x}_i - \bar{x}_j|}.$$

1.2 Méthodes de résolution du problème électronique

L'obtention des forces électroniques agissant sur les noyaux nécessite de résoudre le problème (1.2), et en fait, en pratique, le problème aux valeurs propres non linéaire associé, c'est-à-dire les équations d'Euler-Lagrange correspondantes.

Une première approche consiste naturellement à caractériser les ϕ_i par leurs valeurs sur une grille et d'évaluer leur dérivées par différences finies [8, 32]. Cette approche est peu utilisée en pratique pour les systèmes de petite taille, mais elle peut l'être pour les systèmes en phase solide de grande taille. Le plus souvent, on se place dans le contexte d'une discrétisation des ϕ_i sur une base de Galerkin.

1.2.1 Discrétisation du problème

L'utilisation d'une méthode de Galerkin pour résoudre le problème électronique consiste à considérer une base de fonctions $\{\chi_\mu\}_{\mu=1,N_b}$ et à chercher ϕ_i sous la forme

$$\forall\ 1 \leq i \leq N, \quad \phi_i = \sum_{\mu=1}^{N_b} C_{\mu i}\chi_\mu. \tag{1.3}$$

On suppose sans perte de généralité dans la suite que les fonctions χ_μ et la matrice C sont réelles.

1.2.1.1 Formulation du problème discrétisé

Le développement (1.3) conduit à la forme discrétisée du problème électronique (1.2) suivante

$$\inf\Big\{\mathcal{E}_d(CC^t), C \in \mathcal{M}(N_b,N),\ C^tSC = I_N\Big\} \tag{1.4}$$

où pour les modèles *Restricted Hartree-Fock* (RHF) et de *Kohn-Sham* (KS), on a respectivement

$$\left|\begin{array}{rcl} \mathcal{E}_d^{RHF}(D) & = & \mathrm{Tr}(hD) + \mathrm{Tr}\big(J(D)D\big) - \dfrac{1}{2}\mathrm{Tr}\big(K(D)D\big), \\ \\ \mathcal{E}_d^{KS}(D) & = & 2\mathrm{Tr}(hD) + 2\mathrm{Tr}\big(J(D)D\big) + \mathcal{E}_{xc}(D). \end{array}\right.$$

avec $\forall\ 1 \leq \mu,\nu \leq N_b$,

$$\begin{aligned} S_{\mu\nu} &= \int_{\mathbf{R}^3} \chi_\mu\chi_\nu, \\ h_{\mu\nu} = h^L_{\mu\nu} + h^V_{\mu\nu} &= \frac{1}{2}\int_{\mathbf{R}^3} \nabla\chi_\mu\nabla\chi_\nu + \int_{\mathbf{R}^3} V_{\bar{x}}\chi_\mu\chi_\nu, \quad V_{\bar{x}} = -\sum_{k=1}^{M}\frac{z_k}{|x-\bar{x}_k|}, \\ J(X)_{\mu\nu} &= \sum_{\kappa,\lambda=1}^{N_b} (\mu\nu|\kappa\lambda)X_{\kappa\lambda}, \\ K(X)_{\mu\nu} &= \sum_{\kappa,\lambda=1}^{N_b} (\mu\lambda|\nu\kappa)X_{\kappa\lambda}, \\ \forall\ 1 \leq \kappa,\lambda \leq N_b, \quad (\mu\nu|\kappa\lambda) &= \int_{\mathbf{R}^3}\int_{\mathbf{R}^3} \frac{\chi_\mu(x)\chi_\nu(x)\chi_\kappa(x')\chi_\lambda(x')}{|x-x'|}\,dx\,dx'. \end{aligned} \tag{1.5}$$

La matrice S est la *matrice de recouvrement*, h est la *matrice hamiltonienne* de cœur, $V_{\bar{x}}$ désigne le potentiel des noyaux agissant sur les électrons et $\mathcal{E}_{xc}(D)$ est l'*énergie d'échange-corrélation*. Cette fonctionnelle est généralement non convexe et non linéaire ; par exemple, pour une fonctionnelle dite LDA [4]

$$dite\mathcal{E}_{xc}(D) = \int_{\mathbf{R}^3} \rho(x)\varepsilon_{xc}^{LDA}\big(\rho(x)\big)\,dx \quad \text{avec } \rho(x) = 2\sum_{\mu,\nu=1}^{N_b} D_{\mu\nu}\chi_\mu(x)\chi_\nu(x). \tag{1.6}$$

Les coefficients $(\mu\nu|\kappa\lambda)$ sont appelés *intégrales biélectroniques*. Leur calcul est de complexité *a priori* $\mathcal{O}(N_b^4)$.

L'énergie $\mathcal{E}_d$ ne dépend que de la matrice $D = CC^t$, appelée *matrice densité*. Ceci est la conséquence de l'invariance de l'énergie de Hartree-Fock et de Kohn-Sham par rotation orthogonale des ϕ_i. Afin de s'affranchir de cette dégénérescence, on peut considérer la forme équivalente à (1.4) suivante

$$\inf\Big\{\mathcal{E}_d(D), D \in \mathcal{M}_S(N_b),\ DSD = D,\ \mathrm{Tr}(DS) = N\Big\} \tag{1.7}$$

1.2.1.2 Les équations d'Euler-Lagrange associées

Les équations d'Euler-Lagrange associées à (1.4) s'écrivent (en utilisant l'invariance par rotation orthogonale des colonnes de C pour diagonaliser immédiatement la matrice des multiplicateurs de Lagrange)

$$\left\{\begin{array}{rcll} F(D)C & = & SCE, & E = \mathrm{Diag}(\epsilon_1, \ldots, \epsilon_N), \\ C^tSC & = & I_N, & \\ D & = & CC^t. & \end{array}\right. \tag{1.8}$$

La matrice $F(D)$ est appelée *matrice de Fock*. Pour les modèles RHF et KS, elle s'écrit :

$$F^{RHF}(D) = h + 2J(D) - K(D), \quad F^{KS}(D) = h + 2J(D) + F_{xc}(D)$$

avec

$$\forall\ 1 \leq \mu, \nu \leq N_b, \quad \Big(F_{xc}(D)\Big)_{\mu\nu} = \frac{1}{2}\int_{\mathbf{R}^3} \mu_{xc}(\rho)\chi_\mu\chi_\nu \tag{1.9}$$

avec $\mu_{xc}(\rho) = \partial\mathcal{E}_{xc}/\partial\rho$ où ρ est définie en (1.6). En pratique, la résolution de (1.8) est presque toujours préférée à la minimisation directe de (1.4).

1.2.1.3 Introduction éventuelle d'un pseudo-potentiel

Pour une grande partie des systèmes physiques, la plupart des phénomènes chimiques sont déterminés par les électrons de valence (associés aux orbitales, dites de valence, de plus haute énergie) et non par les électrons de cœur (associés aux orbitales, dites de cœur, de plus basse énergie) [91]. Plus précisément, les grandeurs pertinentes d'un point de vue physico-chimique sont en fait les orbitales de valence *à l'extérieur* de régions entourant les noyaux du système. Ainsi, à chaque atome, on peut associer un cœur, qui coïncide avec la région de localisation des orbitales de cœur associées, dans lequel les orbitales sont peu influencées par le milieu externe. Il n'est donc d'aucun intérêt de calculer exactement les orbitales de cœur et les orbitales de valence dans les cœurs de chaque atome [102].

En raison de ces considérations physiques, on aboutit à une réduction notable de la dimension du problème électronique en remplaçant, surtout lorsque le système

considéré est un système de très grande taille, le potentiel nucléaire $V_{\bar{x}}$ par un pseudo-potentiel (ou potentiel effectif de cœur) V^{ps} défini en coordonnées sphériques [46] par

$$\begin{aligned} \forall\, \psi(r,\theta,\varphi) \quad V^{ps}.\psi(r,\theta,\varphi) &= V^{ps}_{loc}(r)\psi(r,\theta,\varphi) \\ &\quad + \sum_{l=0}^{l_{max}} \sum_{m=-l}^{l} \left(\int_{\mathcal{S}^2} \psi \mathcal{Y}_l^m d\theta d\varphi \right) V_l^{ps}(r) \mathcal{Y}_l^m(\theta,\varphi) \end{aligned}$$

où $(\mathcal{Y}_l^m)_{lm}$ désignent les harmoniques sphériques. Les fonctions radiales V^{ps}_{loc} et V^{ps}_l sont définies par un calcul préliminaire sur l'atome isolé. On définit r_c le rayon de la sphère centrée en l'atome dans laquelle les fonctions ϕ_i correspondant aux électrons de cœur (électrons de plus basse énergie) sont localisées. V^{ps}_{loc} est un potentiel régulier de longue portée. V^{ps}_l sont des potentiels de courte portée définis sur $[0, r_c]$. La matrice h^V est remplacée alors par la matrice h^{ps} définie par

$$\forall\, 1 \leq \mu, \nu \leq N_b, \quad h^{ps}_{\mu\nu} = \int_{\mathbf{R}^3} \chi_\mu \big(V^{ps}.\chi_\nu\big) = \big(h^{ps}_{loc}\big)_{\mu\nu} + \sum_{l=1}^{l_{max}} \big(h^{ps}_l\big)_{\mu\nu}. \tag{1.10}$$

Le principe et la construction des pseudo-potentiels, ainsi que les limites d'applicabilité de la méthode, sont détaillés dans le chapitre 2. En particulier, leur introduction est indispensable pour les méthodes de différences finies, lors de l'usage de la base d'ondes planes et de bases d'orbitales atomiques numériques ainsi que pour le traitement des atomes lourds.

1.2.1.4 Les bases de discrétisation

De par la forme particulière des intégrales à calculer, trois familles de bases se sont imposées dans les codes de calcul : les bases localisées, la base d'ondes planes et les bases mixtes qui combinent les deux premières. On renvoie le lecteur pour ces dernières aux références [11, 44].

Les bases localisées

Les bases localisées : on cite particulièrement les bases de gaussiennes [52] utilisées dans le code *GAUSSIAN* [105] et les bases d'orbitales atomiques numériques [12] utilisées dans les codes *SIESTA*, *DMol* [106, 107]. Leur construction se base sur l'approximation LCAO : chaque ϕ_i s'écrit comme une combinaison linéaire d'orbitales atomiques (c'est à dire de solutions calculées sur l'atome isolé). Ces bases dont l'efficacité nécessite un réel savoir-faire, sont caractérisées par leur petite taille ($N_b = \mathcal{O}(N)$) et le caractère creux des matrices introduites en (1.5). Toutefois, elles sont non orthogonales et peuvent présenter des erreurs de superposition [39] qui conduisent par suite à la singularité de la matrice de recouvrement S.

La base d'ondes planes

La base d'ondes planes utilisée dans les codes *VASP*, *Abinit* [108, 109] : cette base orthogonale ($S = I_{N_b}$) adaptée au traitement des systèmes périodiques est très populaire dans le domaine de la Physique du Solide. L'opérateur cinétique est diagonal dans cette base et on peut bénéficier avantageusement de la transformée de Fourier rapide. Le caractère oscillatoire des fonctions ϕ_i près des noyaux conduit à considérer une base de taille bien plus grande que la taille des bases localisées usuelles : $N_b \simeq 10^2 N$. Par ailleurs, les matrices introduites en (1.5) sont pleines.

1.2.1.5 Calcul d'intégrales

Le choix de la base de discrétisation dépend en premier lieu du système considéré. Le traitement des systèmes périodiques est réalisé généralement par une base d'ondes planes ou une base mixte. En revanche, les bases localisées sont préférées pour le traitement des systèmes moléculaires isolés. Dans un deuxième temps, ce choix est orienté par le calcul des intégrales nécessaires à la construction des matrices définies en (1.5), (1.9) et (1.10), et donc indirectement par le modèle *ab initio* considéré.

Les modèles Hartree-Fock

La complexité du calcul des intégrales biélectroniques conduit à opter le plus souvent pour une base de gaussiennes. Ces bases introduites par Boys [25] permettent *in fine* un calcul analytique des intégrales définissant les matrices définies en (1.5). En particulier, le calcul des intégrales biélectroniques $(\mu\nu|\kappa\lambda)$ se ramène au calcul d'une intégrale à une dimension. L'association de la localisation des gaussiennes et d'une méthode multipôle rapide [129] permet de construire la matrice $F(D)$ avec une complexité linéaire [98].

Les modèles de Kohn-Sham

Seul le calcul de J est nécessaire. Par définition de ρ en (1.6), on a

$$2J_{\mu\nu} = \int_{\mathbf{R}^3} \Big(\rho \star \frac{1}{|x|}\Big)\chi_\mu\chi_\nu. \tag{1.11}$$

L'exploitation de la propriété (1.11) permet le calcul de J avec une complexité $\mathcal{O}\big(N_b^\alpha\big)$ avec $\alpha \leq 3$ pour les bases autres que gaussiennes.
Pour la base d'onde plane, $S = I_{N_b}$ et le calcul de h^L est direct comme le laplacien est diagonal dans cette base. En raison du caractère local de l'hamiltonien de champ moyen de Kohn-Sham, chaque coefficient des matrices h^V, J et F_{xc} est un coefficient de Fourier associé respectivement à $V_{\bar{x}}$, $\rho \star \frac{1}{|x|}$ et $\mu_{xc}(\rho)$. Une transformée de Fourier rapide permet le calcul de chaque matrice avec une complexité $\mathcal{O}\big(N_b log(N_b)\big)$. La construction de $F(D)$ est donc globalement de complexité $\mathcal{O}\big(N_b^2\big)$.
Pour une base d'orbitales atomiques localisées définies sur une grille, le calcul de S et F_{xc} s'effectue par intégration numérique. Le calcul de h^L se fait en approchant l'opérateur cinétique par différences finies. Soit $W = V_{\bar{x}} + \rho \star \dfrac{1}{|x|}$, W est solution de l'équation de Poisson $-\triangle W = 4\pi(\rho_{\bar{x}} - \rho)$ où $\rho_{\bar{x}}$ est la distribution des

charges nucléaires. L'utilisation d'une transformée de Fourier rapide de complexité $\mathcal{O}\big(N_b log(N_b)\big)$ (on peut opter aussi pour une méthode multigrille) permet de résoudre cette équation et d'obtenir par intégration numérique, en utilisant (1.11), les intégrales définies par

$$\int_{\mathbf{R}^3} W\chi_\mu\chi_\nu = h^V_{\mu\nu} + 2J_{\mu\nu}(D).$$

Ces bases étant localisées, les structures de $F(D)$ et S sont connues *a priori*, puisque l'hamiltonien de champ moyen de Kohn-Sham est local. La construction de $F(D)$ et S est de complexité $\mathcal{O}\big(N_b log(N_b)\big)$.
Pour une base de gaussiennes, les méthodes d'intégration numérique ont permis de traiter la matrice d'échange-corrélation LDA définie en (1.9) avec une complexité linéaire avec N_b [31].

Traitement des pseudo-potentiels

Le traitement du potentiel singulier $V_{\bar{x}}$ par transformée de Fourier, ou de façon équivalente par la résolution de l'équation de Poisson associée, n'est pas approprié pour les méthodes de grille qui exploitent la transformée de Fourier rapide. Par ailleurs, les fonctions ϕ_i sont très oscillantes près des noyaux, il est donc difficile de les approcher précisément avec une base d'ondes planes de taille raisonnable. C'est pourquoi, des pseudo-potentiels sont systématiquement introduits dans la résolution du problème électronique lorsqu'on emploie la base d'ondes planes et les bases d'orbitales atomiques numériques. La matrice h^{ps}_{loc}, définie en (1.10), se calcule de la même manière que F_{xc} comme V^{loc}_{ps} est local et régulier. J s'obtient ensuite soit en calculant la transformée de Fourier de $\rho \star \frac{1}{|x|}$ pour la base d'ondes planes, soit en résolvant l'équation de Poisson sans $V_{\bar{x}}$ pour les bases d'orbitales atomiques numériques. Les matrices h^{ps}_l, définies en (1.10), sont calculées par intégration numérique dans l'espace réel. La forme séparable de Kleinman-Bylander [75] est très employée comme on ramène le calcul de $N_b(N_b+1)/2$ intégrales au calcul de N_b intégrales.
Les bases de gaussiennes de très petite taille ne nécessitent pas l'introduction des pseudo-potentiels pour les atomes légers. En revanche, ils sont incontournables pour le traitement des atomes lourds. Afin de faciliter le calcul des matrices h^{ps}_{loc} et h^{ps}_l, les potentiels V^{ps}_{loc} et V^{ps}_l sont approchés au préalable *en un certain sens* par une somme de gaussiennes [34].

Enfin, il est important de noter que le caractère local de V^{ps}_{loc} et localisé des V^{ps}_l implique que la matrice h^{ps} est creuse lorsque les fonctions χ_μ sont localisées. C'est pourquoi, l'introduction des pseudo-potentiels ne détruit pas la complexité quasi-linéaire avec N_b obtenue pour la construction de $F(D)$. En revanche, le calcul des matrices h^{ps}_l pour la base d'ondes planes est de complexité $\mathcal{O}(N_b^3)$. Divers travaux ont conduit à une réduction de cette complexité à $\mathcal{O}\big(N_b^2 log(N_b)\big)$ [80]. Par ailleurs, on précise que la matrice $F(D)$ n'est jamais formée dans un code d'ondes planes. Comme $N \ll N_b$, on utilise des méthodes itératives de diagonalisation pour résoudre (1.8) dans lesquelles les produits *matrice-vecteur*, réalisés au niveau continu par des transformées de Fourier rapide, sont de complexité $\mathcal{O}(N_b log(N_b))$.

1.2.2 Méthode de résolution du problème discrétisé

La fonctionnelle à minimiser dans (1.4) n'étant pas quadratique, les équations (1.8) sont non linéaires. C'est pourquoi, une procédure itérative appelée *algorithme Self Consistent Field* (SCF) est nécessaire pour leur résolution. Pour des raisons de temps de calcul, la résolution itérative de (1.8) est très souvent préférée au traitement de (1.4) par une méthode directe [2] même si on n'est pas assuré de la décroissance de $\mathcal{E}_d$ à chaque itération, ni de la convergence vers un minimum global.

1.2.2.1 Les algorithmes SCF

D'une manière générale, une suite de matrices D_n est construite, le passage de D_n à D_{n+1} s'effectuant à partir de la matrice $F(D_n)$. Dans les méthodes les plus efficaces, le calcul de D_{n+1} est réalisé en gardant à l'esprit qu'on ne cherche pas un point critique de (1.4) mais le minimum global de (1.4) ou de façon équivalente de (1.7). Il est connu théoriquement pour le modèle de Hartree-Fock, qu'au minimum global $D^\star$ de (1.7), la N-ième valeur propre de $F(D^\star)$ est non dégénérée et les réels ϵ_i introduits en (1.8) sont nécessairement les N plus petites valeurs propres de $F(D^\star)$ [133].
En raison de ce résultat, D_{n+1} est construit à l'aide de la matrice $\tilde{D}_n = \tilde{C}\tilde{C}^t$ où $\tilde{C}$ est formée des N plus petits vecteurs propres de $F(D_n)$: c'est le principe *aufbau*. Si on choisit $D_{n+1} = \tilde{D}$, on retrouve le plus simple des algorithmes SCF, l'algorithme de Roothaan [96]. Des algorithmes plus évolués présentant de meilleures propriétés de convergence que l'algorithme de Roothaan ont été développés [13, 27, 127]. Ces algorithmes nécessitent eux aussi le calcul de $\tilde{D}_n$, et par suite la diagonalisation de $F(D_n)$.

1.2.2.2 Le sous-problème linéaire

Le sous-problème linéaire à résoudre à chaque itération d'un algorithme SCF est un problème aux valeurs propres généralisé. La complexité de résolution d'un tel problème par les méthodes directes standards est $\mathcal{O}\big(N_b^3\big)$ quel que soit le caractère creux de $F(D)$ et S (excepté lorsque $F(D)$ et S sont des matrices bandes où une complexité $\mathcal{O}(N_b^2)$ est possible [120]) [112]. Lorsque la base des χ_μ est localisée, $N_b = \mathcal{O}(N)$ et on récupère une complexité $\mathcal{O}(N^3)$. Pour une base d'ondes planes $N \ll N_b$ et une méthode itérative permet de réduire la complexité à $\mathcal{O}(N_b N^2)$ soit $\mathcal{O}(N^3)$.
Comme il a été dit plus haut, la construction de la matrice $F(D)$ est le plus souvent de complexité $\mathcal{O}\big(N^2\big)$ ou $\mathcal{O}\big(N_b log(N_b)\big)$ selon la base de galerkin considérée. Par conséquent, la résolution du sous problème linéaire à chaque itération de l'algorithme SCF constitue l'étape limitante de la résolution de (1.2).

1.2.3 Calcul des forces électroniques

Le calcul des forces électroniques utilise des formules analytiques [92]. Par exemple, pour le modèle RHF

$$\frac{\partial \mathcal{E}_d}{\partial \bar{x}} = \mathrm{Tr}\left(\frac{\partial h}{\partial \bar{x}} D\right) + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial G}{\partial \bar{x}}(D) D\right) - \mathrm{Tr}\left(D_E \frac{\partial S}{\partial \bar{x}}\right)$$

avec $D_E = CEC^t = S^{-1}F(D)D$ et $G(D) = 2J(D) - K(D)$. Les termes $\frac{\partial h}{\partial \bar{x}}$, $\frac{\partial G}{\partial \bar{x}}$, $\frac{\partial S}{\partial \bar{x}}$, se calculent facilement pour une base de gaussiennes (la dérivée d'une gaussienne est une gaussienne).
Lorsqu'un pseudo-potentiel est utilisé, un terme dépendant de $\frac{\partial h^{ps}}{\partial \bar{x}}$ est rajouté. Par ailleurs, on souligne que le calcul des forces est simplifié pour la base d'ondes planes puisque les fonctions χ_μ ne dépendent pas de $\bar{x}$ ($\frac{\partial G}{\partial \bar{x}} = \frac{\partial S}{\partial \bar{x}} = 0$).

1.3 Les méthodes de complexité linéaire avec N

Comme il a été dit plus haut, l'étape limitante d'un calcul *ab initio* est le plus souvent la résolution d'un problème aux valeurs propres généralisé de dimension N_b qui s'écrit

$$\inf\left\{\mathrm{Tr}(FCC^t),\ C \in \mathcal{M}(N_b, N),\ C^t S C = I_N\right\}. \tag{1.12}$$

Les méthodes de diagonalisation standards pour le calcul explicite des N plus petits éléments propres de F sont de complexité $\mathcal{O}(N^3)$. Une telle complexité est en fait loin d'être optimale pour notre problème. En effet seule la matrice $D^\star = C_{sol}C_{sol}^t$, où C_{sol} est une solution de (1.12), est utile pour le calcul de l'énergie et des forces électroniques. Le calcul des N plus petits éléments propres de F eux-mêmes n'est donc pas nécessaire. Comme (1.12) s'écrit de façon équivalente

$$\inf\left\{\mathrm{Tr}(FD),\ D \in \mathcal{M}_S(N_b),\ DSD = D,\ \mathrm{Tr}(SD) = N\right\}, \tag{1.13}$$

la matrice $D^\star$ peut s'obtenir en résolvant (1.13) par une méthode directe. Par conséquent, une diminution de la complexité du calcul de $D^\star$ semble possible si la matrice $D^\star$ solution de (1.13) est creuse. Dans la suite, on introduit le *gap* $\gamma = (\epsilon_{N+1} - \epsilon_N)/(\epsilon_{N_b} - \epsilon_1)$ et le niveau de Fermi $\epsilon_F = (\epsilon_N + \epsilon_{N+1})/2$ où $(\epsilon_i)_i$ désignent les valeurs propres rangées par ordre croissant de F.

1.3.1 Justification

Le principe de localité [66] et le principe de *nearsightedness* introduit par Kohn [78] postulent que certaines propriétés physiques d'un système à un endroit donné ne sont pas affectées par des modifications de l'environnement *suffisamment* lointaines. D'un point de vue formel, ces deux principes ont deux conséquences

- l'existence d'orbitales *localisées* ψ_i^w obtenues par transformation unitaire des orbitales ψ_i solution de (1.2), et appelées *orbitales de Wannier* dans le domaine de la physique du solide [14],
- la décroissance quand $|r - r'| \to +\infty$ de l'opérateur densité $\mathcal{D}$ défini par

$$\mathcal{D}(r,r') = 2\sum_{i=1}^{N} \psi_i(r)\psi_i(r') = 2\sum_{i=1}^{N} \psi_i^w(r)\psi_i^w(r') = 2\sum_{\mu,\nu=1}^{N_b} D^\star_{\mu\nu}\chi_\mu(r)\chi_\nu(r'). \tag{1.14}$$

Des travaux académiques sur des systèmes périodiques infinis ont estimé le taux de décroissance des orbitales de Wannier et de l'opérateur densité [17, 57, 69, 76, 77]. Il ressort les conclusions suivantes :

- La décroissance pour les isolants et semi-conducteurs ($\gamma \neq 0$) est exponentielle

$$\mathcal{D}(r,r') \simeq \exp\Big(-\gamma^\alpha |r-r'|\Big).$$

  où $\alpha = 1/2$ pour les isolants et $\alpha = 1$ pour les semi-conducteurs.
- La décroissance pour les systèmes métalliques ($\gamma = 0$) est exponentielle à température non nulle et polynômiale à température nulle.

Ces résultats ne sont pas étendus aux solides non périodiques et aux systèmes moléculaires. Néanmoins, on suppose que ces propriétés restent vraies [58]. Les ψ_i^w sont appelées alors *orbitales de Wannier généralisées* [83].

Pour des systèmes tels que $\gamma = \mathcal{O}(1)$, il existe un N-uplet de fonctions localisées ψ_i^w solution du problème électronique. Si la base des χ_μ est localisée, il existe une solution creuse $C^\star$ de (1.12) et par suite $D^\star = C^\star(C^\star)^t$ est creuse. De la même façon (1.14) conduit directement au caractère creux de $D^\star$ suite à la décroissance exponentielle de $\mathcal{D}$.

1.3.2 Les différentes méthodes

Suite aux arguments précédents lorsque les χ_μ sont localisées, la résolution de (1.12) dans le contexte du problème électronique se ramène en principe pour les isolants au calcul de $\mathcal{O}(N)$ coefficients (matrice $C^\star$ ou $D^\star$). Dans cette optique, différents travaux ont mené au développement de méthodes de complexité linéaire avec N pour le calcul de $C^\star$ ou $D^\star$. Ces méthodes pour la plupart se décomposent en deux étapes.

Une première étape consiste à caractériser la solution creuse $C^\star$ (respectivement $D^\star$) de (1.12) (respectivement (1.13)) de façon différente. En particulier, il est nécessaire de traiter les contraintes sur C et D autrement car leur prise en compte *explicite* exclut une complexité linéaire avec N.

Dans un second temps, suite à l'hypothèse sur le caractère creux de $C^\star$ et $D^\star$, les calculs sont effectués seulement sur un ensemble de coefficients définis *a priori*. On remarque qu'une telle stratégie permet de lever la dégénérescence du problème (1.12) si on suppose qu'il existe une *unique solution creuse* $C^\star$ à (1.12). Sans stratégie de

troncature, l'étude numérique des méthodes de complexité linéaire est simple. En revanche, aucune étude de leur comportement associé à une stratégie de troncature n'est disponible.

A l'issue de ces deux étapes, on est ramené à utiliser une méthode numérique basée sur des produits matriciels entre les matrices F, S et la matrice C ou D tronquée. Comme ces méthodes se placent dans le contexte d'une base localisée, F et S sont creuses, la complexité linéaire avec N est envisageable. Il reste toutefois à vérifier que le nombre d'itérations pour la convergence de ces méthodes est indépendant de la taille du système. Cette propriété n'est pas justifiée théoriquement mais est observée dans la pratique lorsque le gap γ est indépendant de la taille du système. Une méthode de complexité linéaire avec N est donc possible.

Dans le section suivante, on présente dans le cadre simplifié $S = I_{N_b}$ deux types de méthodes très répandues. Soulignons ici sommairement l'existence de deux autres types de méthodes :

- Une méthode rustique de *décomposition de domaine* qui consiste à partitionner le domaine en sous domaines se recouvrant deux à deux sur lesquels on résout un problème aux valeurs propres de plus petite taille par une méthode standard [104]. La matrice totale $D^\star$ est ensuite reconstruite par des pondérations appropriées sans imposer les contraintes d'orthogonalité et sans utiliser de principe variationnel.
- Des méthodes qui consistent à *pénaliser* de façon naturelle les contraintes sur D dans (1.13) [10, 64, 78, 99].

On laisse le lecteur se référer aux articles de revue suivant pour une description plus complète de l'état de l'art [38, 50, 58, 88]. Dans la suite, $\mathcal{G}_C$ (respectivement $\mathcal{G}_D$) désigne l'ensemble des matrices possédant la structure creuse, supposée *a priori*, de la matrice $C^\star$ (respectivement $D^\star$). Au niveau continu, $\mathcal{G}_C$ (respectivement $\mathcal{G}_D$) correspond au support des fonctions ψ_i^w (respectivement de l'opérateur $\mathcal{D}$).

1.3.2.1 Les méthodes variationnelles

Les méthodes variationnelles sont basées sur la reformulation de (1.12) et (1.13) en un problème de minimisation sans contrainte d'une nouvelle fonctionnelle :

$$\inf\Big\{\Omega_C(C),\ C \in \mathcal{M}(N_b, N)\Big\}, \quad \inf\Big\{\Omega_D(D),\ D \in \mathcal{M}_S(N_b)\Big\}. \tag{1.15}$$

Beaucoup de travaux ont été effectués autour du développement des fonctionnelles Ω_C et Ω_D [40,48,49,54]. Citons la fonctionnelle de D utilisée dans la méthode appelée *Density Matrix Minimization* (DMM) [81] et la fonctionnelle de C utilisée dans le méthode appelée *Orbital Minimization* (OM) [87] :

- pour la méthode DMM : $\Omega_D(D) = \mathrm{Tr}\Big((F - \epsilon_F I_{N_b})(3D^2 - 2D^3)\Big)$,
- pour la méthode OM : $\Omega_C(C) = \mathrm{Tr}\Big((F - \epsilon_F I_{N_b})C(2I_N - C^tC)C^t\Big)$.

On montre que la solution $D^\star$ de (1.13) est l'unique minimum local de Ω_D. De même les solutions de (1.12), dont $C^\star$, sont des minima locaux de Ω_C [1]. Afin de se

ramener à une méthode de complexité linéaire, on remplace ensuite les problèmes définis en (1.15) par

$$\left| \begin{array}{l} \inf \Big\{ \Omega_C(C),\ C \in \mathcal{M}(N_b, N),\ C \in \mathcal{G}_C \Big\}, \\ \inf \Big\{ \Omega_D(D),\ D \in \mathcal{M}_S(N_b),\ D \in \mathcal{G}_D \Big\}. \end{array} \right. \tag{1.16}$$

Sur le plan numérique, l'optimisation est réalisée par une méthode de gradient conjugué [113] avec une recherche linéaire adaptée [110]. Ces méthodes présentent l'avantage d'être variationnelles malgré l'introduction de $\mathcal{G}_{C,D}$. L'énergie approchée obtenue est toujours supérieure à l'énergie exacte. Les forces électroniques obtenues par des formules analytiques se révèlent en pratique suffisamment précises, et assurent une dynamique moléculaire stable [23].

Ces algorithmes nécessitent un initial guess de bonne qualité afin

- de converger vers le mimimum local de (1.16) de plus basse énergie (rappelons en effet que le minimum de (1.15), et par suite de (1.16), est $-\infty$). On montre que pour la méthode DMM, (1.16) admet un unique minimum local. Lorsque $S = I_{N_b}$, l'initialisation donnée par $D_0 = (1/2)I_{N_b}$ assure la convergence vers ce minimum ;
- d'éviter une convergence très lente.

Lorsqu'on considère un système dit ordonné (tel un enzyme ou une proteïne), la solution calculée au pas de temps précédent, proche de la solution cherchée, est un bon initial guess. Cependant, l'obtention d'un bon initial guess se révèle difficile pour la méthode OM car (1.16) admet de nombreux minima locaux d'énergie similaire suite à la localisation imposée.

Concernant la méthode OM, on observe que le nombre d'itérations requis à la convergence est d'autant plus grand que la dimension de $\mathcal{G}_C$ est grande [87]. En effet, plus la dimension de $\mathcal{G}_C$ est grande, plus le conditionnement du problème (1.16) est mauvais, car lié à l'invariance de Ω_C par rotation orthogonale des colonnes de C. Par ailleurs suite à la troncature des calculs, la solution approchée ne vérifie pas la contrainte $\mathrm{Tr}(CC^t) = N$ à l'inverse des méthodes basées sur la matrice D pour laquelle la contrainte linéaire $Tr(D) = N$ est plus simple à assurer [9]. La charge électronique totale ρ n'est donc pas conservée, ce qui est source de problème à l'intérieur de la boucle SCF [48].

Enfin, on souligne que les méthodes variationnelles se ramènent à des méthodes de pénalisation particulières [1].

1.3.2.2 Les méthodes de projection

Les méthodes de projection se basent sur la caractérisation de la matrice D solution de (1.15) :

$$D = q(F)\ \ \forall\, q \text{ tel que } \left| \begin{array}{ll} q(\epsilon_i) = 1 & \forall\ 1 \leq i \leq N \\ q(\epsilon_i) = 0 & \forall\ N+1 \leq i \leq N_b \end{array} \right. . \tag{1.17}$$

Parmi ces méthodes, on trouve :

- La méthode appelée *Fermi Operator Expansion* (FOE) [19, 58] où q est formée d'une expansion de polynômes de Chebyshev $(T_n)_n$,
- une version simple des méthodes de *purification de la matrice densité* (McW) [85, 90] où q est de la forme p^n avec $p : x \longmapsto 3x^2 - 2x^3$.

Des polynômes de Chebyshev-Jackson [33], des fonctions rationnelles [56] ont été aussi utilisés pour approcher la fonction q. Soulignons que la méthode DMM peut s'interpréter comme une méthode de projection si l'initial guess D_0 est une fonction de F.

L'algorithme correspondant à ces méthodes est itératif :

- initialisation : $D_0 = \tilde{F} = \alpha F + \beta$ où α et β dépendent de ϵ_1, ϵ_{N_b} et ϵ_F,
- passage de D_n à D_{n+1} :
  1. FOE : $D_{n+1} = D_n + c_n T_n(\tilde{F})$ où $(c_n)_n$ désignent les coefficients de Chebyshev de la fonction de Heaviside définie sur $[-1, 1]$.
  2. McW : $D_{n+1} = 3D_n^2 - 2D_n^3$.

A chaque itération, le calcul de D_n est tronqué suivant $\mathcal{G}_D$.

Ces méthodes sont plus simples à mettre en œuvre et généralement plus rapides que les méthodes variationnelles (moins d'opérations par itération, convergence quadratique de McW sans troncature). Cependant, les forces électroniques, et donc les simulations de dynamique moléculaire, sont moins précises que celles obtenues avec des méthodes variationnelles. Par ailleurs, leur sensibilité à la stratégie de troncature est plus importante. Notamment, on observe que la convergence du calcul n'est pas nécessairement plus rapide quand on augmente a dimension de $\mathcal{G}_D$ [42].

L'initial guess de ces algorithmes est fixé par la méthode de projection considérée. Par conséquent, aucune information donnée par le calcul au pas de temps précédent dans le contexte d'une simulation de dynamique moléculaire ne conduit à une accélération du calcul de $D^\star$.

Enfin, il existe une généralisation de la méthode FOE pour le calcul de $C^\star$ connue sous le nom de *Fermi Operator Projection* [58].

1.3.3 Insuffisances de ces méthodes

Les bases mathématiques des méthodes de complexité linéaire sont simples. Toutefois l'obtention d'une méthode réellement de complexité linéaire requiert une stratégie de troncature. Celle-ci se révèle difficile à mettre en œuvre et en l'état, n'a jamais été analysée rigoureusement. Notamment, on est amené à faire varier $\mathcal{G}_{C,D}$ au cours de l'algorithme [37] lorsque la structure de $C^\star$ ou $D^\star$ n'est pas connue *a priori*. En outre, ces méthodes ont de nombreuses imperfections.

Connaissance de ϵ_F

Toutes les méthodes développées nécessitent la connaissance de ϵ_F et pour certaines de ϵ_1 et ϵ_{N_b}. Les deux dernières quantités (où une bonne approximation de celles-ci) s'obtiennent avec une complexité $\mathcal{O}(N)$ par une méthode itérative comme F est creuse [116] ou à l'aide des formules de Gershgorin [119]. En revanche ϵ_F est

plus difficile à calculer. En pratique, ces méthodes sont couplées à une méthode itérative de Newton pour mettre à jour une approximation de ϵ_F [94].

Passage au cas $S \neq I_{N_b}$

Il s'effectue en introduisant des approximations supplémentaires. En effet, soit on se ramène au cas $S = I_{N_b}$ en calculant une approximation de la décomposition de Cholesky de S^{-1} [30], soit on utilise les généralisations des méthodes présentées.

- Les méthodes de projection se généralisent en remplaçant F par $\tilde{F} = S^{-1}F$. Cette matrice, dont on a remarqué le caractère creux [51], est calculée en résolvant $S\tilde{F} = F$ où le *masque* de $\tilde{F}$, représentant l'ensemble des coefficients non nuls de $\tilde{F}$, est défini *a priori* [89]. Par ailleurs, les méthodes de purification de la matrice densité nécessitent S^{-1} pour former l'initial guess généralisé. Une bonne approximation de S^{-1} dont le caractère creux n'est pas assuré est donc nécessaire. Enfin, la matrice calculée est $SD^\star$ et donc présente un caractère moins creux que $D^\star$.
- Les méthodes variationnelles se généralisent en préservant leur caractère variationnel au prix d'un plus grand nombre de produits matriciels. Aucune approximation de S^{-1} ou $S^{-1}F$ n'est en principe nécessaire. Cependant, le gradient de la fonctionnelle généralisée proposée dans [86] pour DMM est inconsistant numériquement [103], et déteriore les performances de DMM. Une adaptation basée sur une approximation de $S^{-1}F$ conduit à de meilleures performances de DMM. On aboutit à une conclusion similaire pour la généralisation de OM. Par ailleurs, l'obtention d'un initial guess de bonne qualité pour les méthodes variationnelles est encore plus difficile lorsque $S \neq I_{N_b}$.

Parallélisme

Les méthodes présentées ci-dessus, exceptée FOE, sont basées sur des produits matriciels et sont donc parallélisables. Dans la pratique, elles sont pénalisées par de nombreux échanges d'information entre processeurs [26, 70]. La méthode FOE fait intervenir seulement des produits *matrice-vecteur* et est donc plus efficacement parallélisable [79].

Les méthodes basées sur C

Les méthodes basées sur D se révèlent en pratique beaucoup plus performantes que les méthodes basées sur C. Toutefois, ces dernières présentent des aspects séduisants qui motivent une amélioration future. En effet, il est plus coûteux d'une part de stocker la matrice densité, même creuse. D'autre part, dans certaines situations (réactions chimiques, relaxation locale), la simulation de gros systèmes conduit à considérer une succession d'états perturbés localement. Il n'est donc pas nécessaire de recalculer entièrement la matrice $C^\star$ ou $D^\star$ mais seulement une petite partie de celle-ci. Comme les contraintes d'orthogonalité font intervenir des parties locales de la matrice C ($C_i^t C_j = \delta_{ij}$) au contraire de la contrainte $D^2 = D$, une telle procédure est plus simple à mettre en œuvre sur $C^\star$ que sur $D^\star$: il suffit de laisser invariantes les orbitales de Wannier (soit des colonnes de $C^\star$) dans le domaine qui ne varie pas

et de recalculer les orbitales de Wannier (les autres colonnes de $C^\star$) dans le domaine d'intérêt.

1.3.4 Le traitement des systèmes métalliques

Les méthodes de complexité linéaire ne sont pas adaptées au traitement des systèmes métalliques. En effet, sans stratégie de troncature, le nombre d'itérations nécessaires n_{it} pour converger à une précision donnée est $n_{it} \simeq \gamma^{-\alpha}$ avec α strictement positif. Par conséquent, le problème pour les systèmes métalliques ($\gamma \simeq 0$) est mal conditionné. D'autre part, les matrices $C^\star$ (pas d'existence d'orbitales de Wannier suffisamment localisées pour ces systèmes) et $D^\star$ sont pleines. Il semble donc impossible de les approcher correctement par un nombre d'opérations de complexité linéaire avec N. Le constat sur les approches spécifiques pour le traitement de ces systèmes [18, 53, 84] conforte ce point.

1.4 Vers une dynamique moléculaire *ab initio*

L'incapacité des méthodes de complexité linéaire au traitement précis des systèmes métalliques n'autorise pas des simulations de dynamique moléculaire *ab initio* pertinentes sur de tels systèmes. On rappelle qu'un pas de dynamique moléculaire est de l'ordre de 10^{-15} seconde. A ce jour, des simulations avec N de l'ordre de 10^3 sur un temps T de l'ordre de 10^{-12} seconde sont accessibles alors que des valeurs pertinentes seraient respectivement de l'ordre de 10^4 et 10^{-6}. Pour de tels ordres de grandeur, la dynamique introduite par Car-Parrinello est pour l'instant plus appropriée [28].

Le travail effectué dans la thèse a consisté à apporter de nouvelles idées numériques au cours des différentes étapes décrites dans la section précédente pour s'approcher de ces ordres de grandeur.

1.4.1 Les pseudo-potentiels (chapitre 2 de la thèse)

Les pseudo-potentiels permettent de réduire considérablement la taille du problème. Ils sont par conséquent couramment utilisés dans les codes de calcul de la physique du solide, et de chimie moléculaire pour le traitement des atomes lourds. Leur compréhension repose sur des arguments physico-chimiques simples. Toutefois, la construction traditionnelle d'un pseudo-potentiel avec de *bonnes* propriétés pour un nouveau système est difficile et relève plus d'un art que d'une approche systématique maîtrisée sur le plan mathématique et numérique. Il est notamment impossible de vérifier *a posteriori* la précision d'un pseudo-potentiel. Le chapitre 2, qui reprend une note publiée en interne à EDF [A1], est une tentative de compréhension et de formalisation mathématique de la construction d'un pseudo-potentiel. Même pour les systèmes les plus simples, une compréhension mathématique de la construction

traditionnelle d'un pseudo-potentiel semble encore hors de portée. Il en est autrement de la méthode *Plane Augmented Waves*, dont on propose une description dans un second temps, qui repose sur un principe variationnel susceptible de donner lieu à de futures études mathématiques et numériques.

1.4.2 Le problème électronique (chapitres 3-4 de la thèse)

L'étape limitante lors de la résolution du problème électronique est la résolution d'un problème aux valeurs propres généralisé associé à deux matrices creuses F et S (si on utilise une base localisée) à chaque cycle SCF. Les alternatives développées à ce jour conduisent à une complexité linéaire avec le nombre d'électrons du système considéré. Ces méthodes ne sont toutefois pas encore satisfaisantes. Limitées au traitement des isolants, elles présentent quelques insuffisances.
Par ailleurs, alors que les méthodes de décomposition de domaine [115] sont très répandues dans les autres domaines de la science du fait de leur capacité à utiliser le parallélisme grandissant des ordinateurs, peu d'effort est observé dans le développement de ces méthodes pour le calcul de structures électroniques.

1.4.2.1 Le traitement des isolants (chapitre 3 de la thèse)

En vue du développement ultérieur d'une méthode de complexité linéaire parallélisable sur C pour le traitement des systèmes métalliques, une méthode de décomposition de domaine est présentée pour le traitement des isolants dans le chapitre 3. Cette méthode basée sur le calcul de $C^\star$ présente une complexité linéaire avec N en terme de temps de calcul et d'encombrement mémoire. Elle consiste en une résolution du problème (1.12) sur l'ensemble des matrices C définies sur $\mathcal{G}_C$:

$$\inf \Big\{ \mathrm{Tr}(FCC^t),\ C \in \mathcal{M}(N_b, N),\ C^t S C = I_N,\ C \in \mathcal{G}_C \Big\}.$$

Cette méthode partage les propriétés des méthodes variationnelles et n'est pas soumise aux insuffisances des méthodes de complexité linéaire actuelles :

- un initial guess de bonne qualité n'est pas toujours nécessaire,
- aucune information *a priori* sur le spectre de F (telle la connaissance du niveau de Fermi) n'est nécessaire,
- le traitement du cas $S \neq I_{N_b}$ ne demande aucune approximation supplémentaire,
- à l'instar des méthodes dites de décomposition de domaine, la méthode est fortement parallélisable sur des machines à mémoire distribuée.

On observe dans les tests numériques effectués sur des polymères monodimensionnels simples (F et S bandes) que la méthode introduite converge très rapidement quel que soit l'initial guess vers une solution précise au contraire de DMM très peu robuste sur ce plan. Par suite, la méthode se montre plus performante en terme de temps CPU et en terme de précision que DMM (en tout cas dans notre implémentation) lorsque l'initial guess est mauvais. Toutefois, elle se révèle moins précise en terme

de précision que la méthode DMM pour un temps infini lorsque l'initial guess est de bonne qualité.

En terme de mémoire allouée, la taille mémoire requise croît plus rapidement pour la méthode DMM que pour la méthode de décomposition de domaine développée. En revanche, cette méthode requiert des ressources mémoire beaucoup plus importantes pour le traitement de systèmes physiques de petite taille.
Enfin, la mise à jour de quelques colonnes de $C^\star$ suite à une perturbation locale du système (et donc de F) peut en principe s'effectuer simplement dans l'algorithme présenté. Aucun test numérique n'a été toutefois effectué pour appuyer cette affirmation.

On reproduit en premier lieu dans le chapitre 3 un article soumis [A2]. Notons que des améliorations algorithmiques des étapes locale et globale conduiraient très certainement à une performance accrue de la méthode présentée. En second lieu, on donne une preuve de convergence de la méthode dans un cadre très simplifié. Bien entendu, des travaux plus approfondis d'analyse numérique et de nature plus algorithmique sont à réaliser pour une meilleure compréhension et une plus meilleur performance de l'algorithme développé.

1.4.2.2 Le traitement des systèmes métalliques (chapitre 4 de la thèse)

Dans le chapitre 4, on propose une adaptation des méthodes de projection pour le traitement des systèmes métalliques. Ce travail, commun avec Véronique Duwig et Guy Bencteux (EDF) a fait l'objet d'un rapport interne EDF [128] et de communications orales à des congrès [I1] [I2] [I3] [I4]. On se ramène au traitement d'un isolant caractérisé par $\gamma = \mathcal{O}(1)$ en créant un trou dans le spectre de F autour du niveau de Fermi. On calcule pour cela par l'algorithme *Implicit Restart Algorithm* quelques éléments propres de F autour de ce niveau. Dans un second temps, on applique une méthode de projection (FOE, McW, DMM) de façon astucieuse pour approcher seulement les coefficients $D^\star_{ij}$ de la matrice $D^\star$ solution de (1.13) tels que $F_{ij} \neq 0$. Une diminution de la complexité cubique avec N a été obtenue sur des tests préliminaires effectués sur une matrice tridiagonale. L'obtention d'une telle complexité suppose toutefois de connaître *a priori* le nombre d'éléments propres à calculer et de ne pas demander une précision trop importante à la solution.
La méthode de décomposition de domaine présentée peut s'adapter de la même façon sans difficulté pour tout initial guess au contraire de la méthode DMM qu'on sait adapter pour l'instant lorsque D_0 est un polynôme de F.

1.4.3 La dynamique moléculaire

Une simulation de dynamique moléculaire revient à calculer une inconnue $u_{\bar{x}}$ (ici les matrices $C_{\bar{x}}$ ou $D_{\bar{x}}$ du point de vue algébrique, ou les fonctions ϕ_i du point de vue continu) solution du problème (ici le problème electronique) noté $(\mathcal{P}_{\bar{x}})$ pour un très grand nombre de jeux de paramètres (ici positions des noyaux) $\bar{x}$. Sachant que

$\bar{x}$ évolue lentement au cours de la simulation, on peut penser qu'il en est de même de l'information physique contenue dans $u_{\bar{x}}$ et par conséquent que l'espace vectoriel engendré par les solutions $u_{\bar{x}}$ générées au cours de la simulation est de dimension très faible. Dans une telle situation, la méthode des *bases réduites* s'est montrée pertinente dans d'autres domaines [131,134,135,140,144]. Ces techniques, associées aux techniques d'obtention de bornes *a posteriori*, sont notamment très efficaces lorsqu'on est intéressé non pas par $u_{\bar{x}}$ lui-même mais par une fonctionnelle $s(u_{\bar{x}})$ (coefficients de trainée et de portance en aérodynamique par exemple) [136–138].

1.4.3.1 La méthode des bases réduites (chapitre 5 de la thèse)

La méthode des bases réduites se divise en deux étapes [126] :

- On effectue une étape préliminaire qui consiste à résoudre le problème $(\mathcal{P}_{\bar{x}})$ pour m configurations de référence $\bar{x}_k$, et à calculer des grandeurs $\mathcal{R}$ à partir des solutions $u_{\bar{x}_k}$. Ces calculs supplémentaires sont appelés plus communément *précalculs*. On note $\mathcal{B}$ la base, dite réduite, formée par la famille des $u_{\bar{x}_k}$.
- Pour tout nouveau jeu de paramètres $\bar{x}_l$, on approche $u_{\bar{x}_l}$ par $\tilde{u}_{\bar{x}_l}$ de la forme

$$\tilde{u}_{\bar{x}_l} = \sum_{k=1}^{m} \alpha_k u_{\bar{x}_k}. \tag{1.18}$$

  $\tilde{u}_{\bar{x}_l}$ est alors solution d'un problème $(\tilde{\mathcal{P}}_{\bar{x}})$ posé désormais sur les m coefficients réels α_k. $(\tilde{\mathcal{P}}_{\bar{x}})$ est défini à partir de $(\mathcal{P}_{\bar{x}})$, (1.18) et utilisant $\mathcal{R}$.

Excepté pour les configurations précalculées $\bar{x}_k$, la résolution de $(\mathcal{P}_{\bar{x}})$ est remplacée par la résolution de $(\tilde{\mathcal{P}}_{\bar{x}})$ de dimension m bien inférieure à la dimension du problème $(\mathcal{P}_{\bar{x}})$. Il en ressort un gain total notable dès que le nombre de configurations $\bar{x}_l$ à considérer est grand comme c'est le cas en dynamique moléculaire.
Comme le choix optimal des configurations de référence $\bar{x}_k$ assurant une précision suffisante n'est pas possible *a priori* excepté pour quelques problèmes académiques [139], ces méthodes comprennent des estimateurs *a posteriori* qui donnent une estimation de la précision de la solution approchée $\tilde{u}_{\bar{x}_l}$ [141, 142, 145] : si une précision suffisante n'est pas atteinte, $\mathcal{B}$ est enrichie de l'élément $u_{\bar{x}_l}$.

Le développement d'estimateurs *a posteriori* précis et peu coûteux est difficile. Lorsque la difficulté du problème est telle que leur obtention n'est pas possible, on doit définir *a priori* une base $\mathcal{B}$ alliant précision et petite dimension. Lorsque le domaine d'évolution des paramètres est connu, on peut choisir des configurations de référence pavant *plus ou moins* ce domaine.
Lorsqu'on réalise une simulation de dynamique moléculaire, on ne connaît pas *a priori* l'évolution du système. Par ailleurs, quelques tests effectués sur des systèmes simples montrent que le traitement d'une simulation dans son intégralité par une base $\mathcal{B}$ fixe se traduit par une trop grande dimension m. En revanche, des tests montrent que chaque solution $u_{\bar{x}}$ peut s'exprimer avec précision par une combinaison linéaire d'un petit nombre de solutions calculées à des pas de temps précédents. La

pertinence de l'approche des bases réduites se traduit donc par la gestion dynamique de $\mathcal{B}$. Dans cette optique, il reste à définir des critères pour réduire la base $\mathcal{B}$.

1.4.3.2 Les difficultés liées au problème électronique

La méthode des bases réduites exige de développer des estimateurs *a posteriori* mais en tout premier lieu de définir un problème $(\tilde{\mathcal{P}}_{\bar{x}})$ et un algorithme, qu'on appelle par la suite algorithme base réduite, pour la génération d'un solution approchée $\tilde{u}_{\bar{x}}$ de $u_{\bar{x}}$. Comme on le verra dans le chapitre 5, cette tâche est immédiate lorsqu'on considère le problème $(\mathcal{P}_{\bar{x}})$ suivant où $\bar{x}$ est un réel positif

$$(\mathcal{P}_{\bar{x}}) \left\{ \begin{array}{rcll} -\triangle u + \bar{x}u & = & f & \text{dans } \Omega, \\ u & = & 0 & \text{sur } \partial\Omega. \end{array} \right.$$

La situation est différente lorsqu'on considère un seul atome de charge z. Les équations d'Euler-Lagrange associées au problème électronique s'écrivent pour le modèle RHF sans pseudo-potentiel

$$\left\{ \begin{array}{c} -\frac{1}{2}\triangle\phi_i + V_{\bar{x}}\phi_i + \Big(\rho \star \frac{1}{|x|}\Big)\phi_i - \sum_{j=1}^{N} \Big(\phi_i\phi_j \star \frac{1}{|x|}\Big)\phi_j = \epsilon_i\phi_i \quad i = 1,\ N, \\ \rho(x) = 2\sum_{i=1}^{N} \phi_i^2(x),\ V_{\bar{x}} = \frac{z}{|x-\bar{x}|}, \\ \int_{\mathbf{R}^3} \phi_i\phi_j = \delta_{ij}. \end{array} \right. \tag{1.19}$$

Ces équations présentent quatre difficultés :

- elles sont non linéaires en les fonctions ϕ_i,
- elles sont non affines en les paramètres $\bar{x}$,
- le problème est vectoriel (N fonctions ϕ_i),
- les contraintes sont non linéaires et en grand nombre : $N(N+1)/2$.

Afin de surmonter les deux premières difficultés, une adaptation générale de la méthode des bases réduites est proposée dans un cadre général : une base réduite est définie pour approcher chaque terme non linéaire en l'inconnue et chaque terme non affine en les paramètres. Elle est présentée sous la forme d'une note CRAS publiée en 2004 [A3] reproduite dans la deuxième section du chapitre 5.

Les troisième et quatrième difficultés compliquent sensiblement la construction d'un algorithme de base réduite efficace. Nos investigations pour le traitement de ces difficultés se ramènent aux conclusions suivantes. Concernant le traitement de la troisième difficulté, deux stratégies sont possibles.

- Pour chaque ϕ_i, on définit une base réduite $\mathcal{B}_i$ formée de m fonctions $(\phi_i)_{\mu_k^i}$, on pose ensuite

$$\forall\ 1 \leq i \leq N, \quad \tilde{\phi}_i = \sum_{k=1}^{m_i} \alpha_{i,k}(\phi_i)_{\mu_k^i}.$$

Une telle approche conduit au mieux à un algorithme base réduite de complexité $\mathcal{O}(N)$ comme on se ramène à un problème sur les $\mathcal{O}(N)$ inconnues $\alpha_i, k)$.

- On définit une base réduite $(\phi_1^k, \ldots, \phi_N^k)_k$ formée de m N-uplets $(\phi_i^k)_{i=1,N}$, l'algorithme base réduite consiste ensuite à chercher $\tilde{\phi}_i$ de la forme

$$\forall\, 1 \leq i \leq N, \quad \tilde{\phi}_i = \sum_{k=1}^{m} \alpha_k \phi_i^k.$$

Comme on utilise les mêmes inconnues α_k pour approcher chaque ϕ_i, un algorithme base réduite de complexité indépendante de N semble possible. Des tests préliminaires montrent que la dimension m est bien supérieure aux dimensions m_i introduites dans la stratégie précédente. Suite à l'augmentation notable des précalculs, une telle démarche ne peut pas être mise en place pour des systèmes réalistes.

Quelle que soit la stratégie choisie pour traiter la troisième difficulté, le problème base réduite pour l'ordre de grandeur de N visé est de dimension très inférieure au nombre de contraintes à imposer. Il est donc nécessaire d'imposer les contraintes de façon plus faible tout en assurant la convergence vers une solution approchée $\tilde{u}_{\bar{x}}$ proche de la solution $u_{\bar{x}}$. Dans cette optique, on s'est inspiré des méthodes variationnelles de complexité linéaire en appliquant la méthode des bases réduites au problème (1.15) posé en terme de matrice densité. Ce travail préliminaire réalisé durant le CEMRACS 2001 a abouti à des résultats encourageants même si seule une amélioration du préfacteur des méthodes de complexité linéaire à défaut d'une amélioration de la complexité avec N semble envisageable.

Par ailleurs, les solutions ϕ_i présentent des pics en les positions des noyaux $\bar{x}_k$. La méthode des bases réduites n'est donc pas applicable directement sur les ϕ_i. Lorsque les χ_i ne sont pas centrées en les noyaux (Eléments Finis, différences finies, ondes planes), une première démarche consiste à définir dans un premier temps une transformation adaptée $\mathcal{T}_{\bar{x}}$ qui associe à toute solution ϕ_i une solution $\psi_i = \mathcal{T}_{\bar{x}}\phi_i$ définie sur un domaine où la position des noyaux est fixe. On écrit ensuite le problème électronique en les ψ_i, puis on utilise la méthode des bases réduites pour approcher les ψ_i.

Lorsque les χ_i sont centrés en les noyaux, une seconde stratégie consiste à définir un algorithme base réduite sur la quantité algébrique $C_{\bar{x}}$ ou $D_{\bar{x}}$. Cette approche est plus simple à mettre en œuvre en particulier lorsque la définition de la transformation $\mathcal{T}_{\bar{x}}$ est difficile, soit dès que le nombre de paramètres est supérieur à quelques unités (par exemple, quelques atomes dans $\mathbb{R}^3$). En revanche, elle est moins satisfaisante que l'approche continue pour laquelle on peut utiliser la stratégie présentée dans la deuxième section du chapitre 5 et développer plus classiquement des estimateurs *a posteriori*.

On rappelle que l'introduction des pseudo-potentiels a pour effet de régulariser les fonctions ϕ_i autour des noyaux. Il n'est donc plus nécessaire de définir une transformation $\mathcal{T}_{\bar{x}}$. Aucun travail dans cette direction n'a été réalisé.

Suite à ces difficultés, on s'est affranchi des troisième et quatrième difficultés en réalisant des tests "académiques" sur des "systèmes école". Ces systèmes écoles ont permis de tester l'approche sur l'ion H_2^+ et la molécule H_2 traitée par le modèle RHF ($N = 1$). Le paramètre μ est la distance entre les noyaux d'hydrogène. On a fait l'hypothèse d'une symétrie cylindrique et la fonction ϕ_μ solution de (1.19) est discrétisée sur une base χ_i d'Éléments Finis P_1. La définition de $\mathcal{T}_\mu$ est simple dans ce contexte. Ainsi, on se ramène à un problème posé sur une fonction $\psi_\mu = \mathcal{T}_\mu \phi_\mu$ à laquelle on applique la stratégie présentée en seconde partie du chapitre 5 pour le traitement du terme de Coulomb et du terme dépendant du potentiel nucléaire.

Dans la première section du chapitre 5, on présente dans un premier temps le principe général de la méthode des bases réduites sur un exemple simple. Dans un second temps, on décrit les résultats obtenus sur quelques systèmes école considérés en vue de la mise en pratique future de l'approche base réduite sur des systèmes moléculaires réels.

Chapitre 2

Les pseudo-potentiels

Les pseudo-potentiels basés sur les principes *ab initio* ont été introduits dans les années 40 en physique du solide par Herring [67] et en chimie quantique dans les années 60 par Phillips et Kleinman [73]. Leur introduction est motivée par le principe chimique connu sous la dénomination d'*approximation des cœurs gelés* : la plupart des phénomènes chimiques sont déterminés par les électrons de valence et non par les électrons de cœur [91]. Plus précisément, les grandeurs pertinentes d'un point de vue physico-chimique sont en fait les orbitales de valence *à l'extérieur* de régions entourant les noyaux du système. Ainsi, à chaque atome, on peut associer un cœur, qui coïncide avec la région de localisation des orbitales de cœur associées, dans lequel les orbitales sont peu influencées par le milieu externe. Il n'est donc d'aucun intérêt de calculer exactement les orbitales de cœur et les orbitales de valence dans les cœurs de chaque atome [102].

Dans cette optique, on introduit pour chaque atome un opérateur, appelé pseudo-potentiel, qui réalise l'approximation de l'interaction entre électrons de cœur et électrons de valence. Une fois ce potentiel ajouté au potentiel coulombien de chaque noyau, on est ramené à l'étude d'un système constitué de pseudo-atomes présentant un plus petit nombre d'électrons (il n'y a plus d'électrons de cœur) pour lesquels les orbitales de valence sont plus simples à représenter [65].

La première approche consiste à approcher le potentiel d'action des électrons de cœur de chaque atome du système par un potentiel obtenu par un calcul *ab initio* précis effectué sur chaque atome du système, pris isolé et dans un état électronique de référence. Les pseudo-orbitales associées à ces potentiels coïncident avec les orbitales de valence qu'on obtiendrait par un calcul *ab initio* complet du système. En particulier, ces pseudo-orbitales présentent dans les cœurs des atomes un grand nombre d'oscillations dont le traitement numérique est difficile pour une base d'ondes planes ou d'orbitales définies sur une grille et, dans une moindre mesure, pour une base de gaussiennes. Comme les informations relatives à la partie des orbitales de valence située dans le cœur sont peu pertinentes d'un point de vue physico-chimique, on développe, dans une seconde approche, des pseudo-potentiels, appelés *Effective Core Potentials*, dont découlent des pseudo-orbitales à la fois moins oscillantes dans les

cœurs des atomes, et qui coïncident avec les orbitales de valence exactes hors du cœur de chaque atome. Les pseudo-orbitales dans les cœurs des atomes étant moins oscillantes, elles admettent une représentation plus simple avec les bases usuelles, d'où une réduction du coût calcul.

Bien que cette démarche procède d'arguments physico-chimiques simples et légitimes, il n'en existe pas de validation rigoureuse, même pour les systèmes les plus simples. Une mesure de la qualité d'un pseudo-potentiel est sa *transférabilité* qui s'identifie à sa capacité à pouvoir être utilisé dans des situations autres que la situation de référence qui a servi à sa construction. On observe que seule la comparaison numérique permet de prédire *a posteriori* la transférabilité d'un pseudo-potentiel donné pour une situation donnée. La génération de bons pseudo-potentiels requiert un réel savoir-faire, qui semble hors de portée des arguments mathématiques dont on dispose à l'heure actuelle. Cependant, il est important de moduler ce constat selon le domaine dans lequel on se place. En effet, les domaines de la chimie moléculaire et de la physique du solide diffèrent par les bases et modèles utilisés, les systèmes physiques étudiés et les propriétés cherchées.

Dans les quatre premières sections, on précise ce qui est dit précédemment en suivant le même fil conducteur. Dans la dernière section, on présente la méthode PAW introduite par Blöchl en 1994 [22]. Cette méthode procède d'une approche variationnelle et pourrait donner lieu à des études mathématiques et numériques.

Enfin, les pseudo-potentiels sont incontournables pour le traitement des atomes lourds, les électrons de cœur étant soumis à des effets relativistes. La prise en compte de ces effets conduit à considérer des modèles relativistes de Dirac-Fock beaucoup plus complexes et coûteux que les modèles non relativistes [93]. Comme les électrons de valence responsables de la liaison chimique ne sont pas soumis à ces effets, on construit un pseudo-potentiel relativiste pour chaque atome en considérant un modèle relativiste sur l'atome isolé. On renvoit le lecteur aux références suivantes pour la construction de ces pseudo-potentiels [62, 74]. Dans la suite, on ne tient pas compte du spin et $(\mathcal{Y}_l^m)_{l\geq 0,-l\leq m\leq l}$ désignent les harmoniques sphériques.

2.1 La méthode des pseudo-potentiels

Pour plus de clarté, on présente dans un premier temps la méthode des pseudo-potentiels sur un système constitué d'un seul atome. L'influence du milieu extérieur sur cet atome est modélisé via un potentiel extérieur.
On indiquera sommairement dans un second temps la façon dont les pseudo-potentiels sont utilisés pour le traitement des systèmes moléculaires.

2.1.1 Un problème atomique type de la chimie quantique

On considère un noyau de charge z situé en $\bar{x} \in \mathbb{R}^3$ entouré de N électrons, le tout placé dans un potentiel extérieur V_{ex}. L'état électronique fondamental de ce

système s'obtient en résolvant les équations RHF ou KS

$$(\mathcal{P}) \left\| \begin{array}{l} \text{Les } (\psi_i, \epsilon_i)_i \text{ sont les } N \text{ plus petits modes propres de} \\ \forall\, 1 \le i \le N, \quad -\frac{1}{2}\triangle\psi_i + \big(V_{nuc} + V_{ex}\big)\psi_i + \mathcal{K}^{RHF,KS}(\Psi).\psi_i = \epsilon_i\psi_i, \\ \forall\, 1 \le i,j \le N, \quad \big(\psi_i, \psi_j\big) = \delta_{ij}, \end{array} \right. \tag{2.1}$$

avec

$$\epsilon_1 \le \ldots \le \epsilon_i \le \ldots \le \epsilon_N.$$

Le potentiel nucléaire V_{nuc} est défini par

$$\forall\, x \in \mathbb{R}^3 \quad V_{nuc}(x) = -\frac{z}{|x - \bar{x}|},$$

et, en notant Ψ la famille des fonctions $(\psi_i)_{i=1,N}$,

$$\forall\, \psi \in L^2(\mathbb{R}^3), \quad \Big(\mathcal{K}^{RHF}(\Psi).\psi\Big)(x) = J(\rho_\Psi)(x)\psi(x) - \frac{1}{2}\Big(K(\tau_\Psi).\psi\Big)(x),$$

$$\forall\, \psi \in L^2(\mathbb{R}^3), \quad \Big(\mathcal{K}^{KS}(\Psi).\psi\Big)(x) = J(\rho_\Psi)(x)\psi(x) + \mu_{xc}(\rho_\Psi)(x)\psi(x),$$

avec

$$\forall\, x \in \mathbb{R}^3, \quad J(\rho_\Psi)(x) = \int_{\mathbf{R}^3} \frac{\rho_\Psi(y)}{\mid x - y \mid}\, dy,$$

$$\forall\, x \in \mathbb{R}^3, \quad \Big(K(\tau_\Psi).\psi\Big)(x) = \int_{\mathbf{R}^3} \frac{\tau_\Psi(x,y)}{|x-y|}\psi(y)\, dy,$$

$$\forall\, x \in \mathbb{R}^3, \quad \rho_\Psi(x) = \sum_{i=1}^{N} |\psi_i(x)|^2,$$

$$\forall\, x, y \in \mathbb{R}^3, \quad \tau_\Psi(x,y) = \sum_{i=1}^{N} \psi_i(x)\psi_i^*(y).$$

Enfin, $\mu_{xc}(\rho)$ est le potentiel d'échange-corrélation (introduit à la section 1.2.1) dépendant de ρ. On note N_c, respectivement N_v, le nombre d'électrons de cœur, respectivement de valence, de l'atome. On introduit les notations suivantes :

- pour tout $1 \le i \le N_c$, $\psi_i^c = \psi_i$, (orbitales de cœur de l'atome),
- pour tout $1 \le i \le N_v$, $\psi_i^v = \psi_{i+N_c}$, (orbitales de valence de l'atome),
- pour tout $1 \le i \le N_c$, $\epsilon_i^c = \epsilon_i$ (énergies monoélectroniques des électrons de cœur),
- pour tout $1 \le i \le N_v$, $\epsilon_i^v = \epsilon_{i+N_c}$, (énergies monoélectroniques des électrons de valence).

Notons bien que ces quantités sont relatives à un calcul effectué sur le système perturbé par le potentiel extérieur. Dans la suite, $\mathcal{K}$ est une notation simplifiée pour $\mathcal{K}^{RHF,KS}$. Il faut toutefois garder à l'esprit que, contrairement à $\mathcal{K}^{KS}$, $\mathcal{K}^{RHF}$ n'est pas un potentiel local (i.e. un opérateur multiplicatif).

2.1.2 Introduction de problèmes de référence

On introduit un problème de référence $(\mathcal{P}_r)$ où l'atome est dans un état électronique donné. Généralement, on choisit le fondamental de l'atome neutre. Les équations d'Euler-Lagrange $(\mathcal{P}_r)$ associées à ce problème s'écrivent

$$(\mathcal{P}_r) \left\| \begin{array}{l} \text{Les } (\psi_i^r, \epsilon_i^r)_i \text{ sont les } N \text{ plus petits modes propres de} \\ \forall\ 1 \leq i \leq N, \quad -\frac{1}{2}\triangle\psi_i^r + V_{nuc}\psi_i^r + \mathcal{K}(\Psi_r).\psi_i^r = \epsilon_i^r\psi_i^r, \\ \forall\ 1 \leq i,j \leq N, \quad \left(\psi_i^r, \psi_j^r\right) = \delta_{ij}. \end{array} \right. \tag{2.2}$$

avec $\Psi_r = (\psi_i^r)_{1\leq i\leq N}$ et

$$\epsilon_1^r \leq \ldots \leq \epsilon_i^r \leq \ldots \leq \epsilon_N^r.$$

De la même façon que précédemment, on pose,

$$\begin{array}{lll} \forall\ 1 \leq i \leq N_c, & \psi_i^{r,c} & = \quad \psi_i^r, \\ \forall\ 1 \leq i \leq N_v, & \psi_i^{r,v} & = \quad \psi_{i+N_c}^r, \\ \forall\ 1 \leq i \leq N_c, & \epsilon_i^{r,c} & = \quad \epsilon_i^r, \\ \forall\ 1 \leq i \leq N_v, & \epsilon_i^{r,v} & = \quad \epsilon_{i+N_c}^r. \end{array}$$

2.1.3 Approximation des cœurs gelés

L'approximation des cœurs gelés à la base de l'introduction des pseudo-potentiels s'écrit

$$\forall\ 1 \leq i \leq N_c, \quad \forall\ x \in \mathbb{R}, \quad \psi_i^c(x) \simeq \psi_i^{r,c}(x), \tag{2.3}$$

$$\forall\ 1 \leq i \leq N_c, \quad \epsilon_i^c \simeq \epsilon_i^{r,c}, \tag{2.4}$$

$$\forall\ 1 \leq i \leq N_v, \quad \forall\ x \in \mathcal{S}^c, \quad \psi_i^v(x) \simeq \psi_i^{r,v}(x), \tag{2.5}$$

où $\mathcal{S}^c$ désignent le cœur de l'atome où sont localisées les orbitales de cœur $\psi_i^{r,c}$ correspondantes. C'est en pratique une sphère de rayon r_c centrée sur le noyau.

2.1.4 Récriture du problème modèle

Le problème $(\mathcal{P})$ se récrit

$$\left\{ \begin{array}{ll} \text{Les } (\psi_i, \epsilon_i)_i \text{ sont les } N \text{ plus modes propres de} & (2.1a) \\ \forall\ 1 \leq i \leq N_c, \quad -\frac{1}{2}\triangle\psi_i^c + \left(V_{nuc} + V_{ex}\right)\psi_i^c + \mathcal{K}(\Psi^c).\psi_i^c + \mathcal{K}(\Psi^v).\psi_i^c = \epsilon_i^c\psi_i^c, & (2.1b) \\ \forall\ 1 \leq i \leq N_v, \quad -\frac{1}{2}\triangle\psi_i^v + \left(V_{nuc} + V_{ex}\right)\psi_i^v + \mathcal{K}(\Psi^c).\psi_i^v + \mathcal{K}(\Psi^v).\psi_i^v = \epsilon_i^v\psi_i^v, & (2.1c) \\ \forall\ 1 \leq i,j \leq N_c, \quad \left(\psi_i^c, \psi_j^c\right) = \delta_{ij}, & (2.1d) \\ \forall\ 1 \leq i,j \leq N_v, \quad \left(\psi_i^v, \psi_j^v\right) = \delta_{ij}, & (2.1e) \\ \forall\ 1 \leq i \leq N_v,\ \forall\ 1 \leq j \leq N_c, \quad (\psi_i^v, \psi_j^c) = 0. & (2.1f) \end{array} \right.$$

avec $\Psi^c = (\psi_i^c)_{1\leq i\leq N_c}$ et $\Psi^v = (\psi_i^v)_{1\leq i\leq N_v}$. Le seul ingrédient pour passer de (2.1) à $(2.1a) - (2.1f)$ est l'utilisation de la linéarité de $\mathcal{K}$ comme fonction de la matrice densité à N corps. Cette propriété est vérifiée par $\mathcal{K}^{RHF}$. En revanche, elle n'est plus vraie pour $\mathcal{K}^{KS}$. On la justifie en faisant l'hypothèse d'une séparation spatiale totale entre les densités de cœur et de valence. Lorsque cette hypothèse se révèle trop forte, il est nécessaire d'introduire des corrections de cœur non linéaires [35, 82]. Ainsi, le calcul de $(\psi_i^v, \epsilon_i^v)_i$ se ramène à la résolution du système d'équations

$$\left\{\begin{array}{l} \text{Les } (\psi_i^v, \epsilon_i^v)_i \text{ sont les } N_v \text{ plus petites solutions de} \\ \forall\ 1\leq i\leq N_v, \quad -\dfrac{1}{2}\triangle\psi_i^v + \big(V_{nuc}+V_{ex}\big)\psi_i^v + \mathcal{K}(\Psi^c).\psi_i^v + \mathcal{K}(\Psi^v).\psi_i^v = \epsilon_i^v\psi_i^v, \\ \forall\ 1\leq i,j\leq N_v, \quad \big(\psi_i^v,\psi_j^v\big) = \delta_{ij}, \\ \forall\ 1\leq i\leq N_v,\ \forall\ 1\leq j\leq N_c, \quad (\psi_i^v,\psi_j^c) = 0, \end{array}\right.$$

avec

$$\epsilon_1^v \leq \ldots \leq \epsilon_i^v \leq \ldots \leq \epsilon_{N_v}^v.$$

Il s'agit quasiment de la résolution d'un système d'équations RHF-KS modélisant un atome constitué de N_v électrons et d'un noyau dont l'action sur les électrons est donnée par l'addition du potentiel extérieur V_{ex} et du potentiel total $V^{tot} = V_{nuc} + \mathcal{K}(\Psi^c)$ qu'on ne connaît pas *a priori*. De plus, il reste à s'affranchir des contraintes $(2.1f)$ si l'on veut retrouver exactement un problème de type RHF-KS. On introduit alors un pseudo-potentiel V^{ps} et on approche le problème $(\mathcal{P})$ à $N = N_c + N_v$ électrons par le problème $(\mathcal{P}^{ps})$ à N_v électrons défini par

$$(\mathcal{P}^{ps}) \left\|\begin{array}{l} \text{Les } (\phi_i^{ps}, \epsilon_i^{ps})_i \text{ sont les } N_v \text{ plus petites solutions de} \\ \forall\ 1\leq i\leq N_v, \quad -\dfrac{1}{2}\triangle\phi_i^{ps} + \big(V_{nuc}+V_{ex}\big)\phi_i^{ps} + V^{ps}\phi_i^{ps} + \mathcal{K}(\Phi^{ps}).\phi_i^{ps} = \epsilon_i^{ps}\phi_i^{ps}, \\ \forall\ 1\leq i,j\leq N_v, \quad \big(\phi_i^{ps},\phi_j^{ps}\big) = \delta_{ij}. \end{array}\right.$$

La résolution de $(\mathcal{P}^{ps})$ est moins coûteuse que la résolution de $(\mathcal{P})$ parce que $N_v < N$ et surtout parce que, les pseudo-orbitales ϕ_i^{ps}, nous le verrons plus loin, sont plus régulières que les orbitales ψ_i^v et peuvent donc être développées sur une base d'ondes planes de plus petite taille. Comme on ne s'intéresse d'un point de vue physico-chimique qu'aux énergies monoélectroniques et à l'expression des orbitales de valence *en dehors de la régions du cœur* (car c'est là que se manifestent les phénomènes chimiques comme la création de liaison), on cherchera V^{ps} de sorte que

$$\forall\ 1\leq i\leq N_v, \quad \forall\ x\notin\mathcal{S}^c \quad \phi_i^{ps}(x) = \psi_i^v(x), \tag{2.6}$$

$$\forall\ 1\leq i\leq N_v, \quad \epsilon_i^{ps} = \epsilon_i^v. \tag{2.7}$$

De plus, si possible, on cherchera à obtenir des pseudo-orbitales ϕ_i^{ps} qu'on puisse correctement représenter dans $\mathcal{S}^c$ par des bases d'ondes planes de petite taille afin de diminuer le coût de la résolution de $(\mathcal{P}^{ps})$.

La construction d'un pseudo-potentiel V^{ps} nécessite *a priori* la connaissance des orbitales exactes ψ_i dans le cœur $\mathcal{S}^c$ de l'atome, qu'on veut éviter de calculer. Il semble donc naturel, à la vue des propriétés (2.3)-(2.5), de poser

$$V^{ps} = V_r^{ps} \tag{2.8}$$

où le potentiel V_r^{ps} est un pseudo-potentiel introduit de la même manière que le potentiel V^{ps} ci-dessus mais pour le problème de référence $(\mathcal{P}_r)$ défini en (2.2) et qu'on résoudra une fois pour toutes. On fait correspondre à $(\mathcal{P}_r)$ le problème $(\mathcal{P}_r^{ps})$ défini par

$$(\mathcal{P}_r^{ps}) \left\| \begin{array}{l} \text{Les } (\phi_i^{r,ps}, \epsilon_i^{r,ps})_i \text{ sont les } N_v \text{ plus petites solutions de} \\ \forall\ 1 \le i \le N_v, \quad -\frac{1}{2}\triangle\phi_i^{r,ps} + V_{nuc}\phi_i^{r,ps} + V_r^{ps}\phi_i^{r,ps} + \mathcal{K}(\Phi_r^{ps}).\phi_i^{r,ps} = \epsilon_i^{r,ps}\phi_i^{r,ps}, \\ \forall\ 1 \le i,j \le N_v, \quad \left(\phi_i^{r,ps}, \phi_j^{r,ps}\right) = \delta_{ij}, \end{array} \right.$$

avec

$$\epsilon_1^{r,ps} \le \ldots \le \epsilon_j^{r,ps} \le \ldots \le \epsilon_{N_v}^{r,ps},$$

tel que

$$\forall\ 1 \le i \le N_v, \quad \forall\ x \notin \mathcal{S}^c, \quad \phi_i^{r,ps}(x) = \psi_i^{r,v}(x), \tag{2.9}$$

$$\forall\ 1 \le i \le N_v, \quad \epsilon_i^{r,ps} = \epsilon_i^{r,v}, \tag{2.10}$$

avec des pseudo-orbitales $\phi_i^{r,ps}$ qu'on puisse correctement représenter dans $\mathcal{S}^c$ par des bases d'ondes planes de petite taille.

On a donc ramené le calcul du pseudo-potentiel V^{ps} au calcul d'un pseudo-potentiel V_r^{ps} adapté à un système de référence, calculé une fois pour toutes. On propose de décrire sommairement les principes de construction de ces pseudo-potentiels V_r^{ps} dans la section 2.2. Dans un deuxième temps, on s'interrogera sur la validité, au sens des propriétés (2.6-2.7), du choix (2.8).

2.2 Les méthodes générales de construction

On appelle $(2.1f^r)$ l'analogue de la condition $(2.1f)$ pour le problème $(\mathcal{P}_r)$, soit

$$\forall\ 1 \le i \le N_v,\ \forall\ 1 \le j \le N_c, \quad \left(\psi_i^{r,v}, \psi_j^{r,c}\right) = 0. \tag{$2.1f^r$}$$

Cette condition permet de ne pas retomber sur les orbitales $\psi_i^{r,c}$ lorsqu'on résout $(\mathcal{P}_r^{ps})$ avec $V_r^{ps} = \mathcal{K}(\Psi_r^c)$. S'affranchir de $(2.1f^r)$ nécessite de perturber le choix de V_r^{ps} de façon à faire coïncider des couples $(\phi_i^{r,ps}, \epsilon_i^{r,ps})$, vérifiant (2.9) et (2.10), avec les plus bas niveaux du potentiel $-\frac{1}{2}\triangle + V_{nuc} + V_r^{ps} + \mathcal{K}(\Phi_r^{ps})$.
La première approche, à la base des *Potentiels Modèles* [20], consiste à remonter suffisamment les niveaux d'énergie associés aux orbitales $\psi_i^{r,c}$. Les pseudo-orbitales $\phi_i^{r,ps}$ solutions de $(\mathcal{P}_r^{ps})$ s'identifient aux orbitales $\psi_i^{r,v}$ sur tout $\mathbb{R}^3$. La relation $(2.1f^r)$, vérifiée par les pseudo-orbitales $\phi_i^{r,ps}$, n'est plus nécessaire.

Dans la seconde approche, à la base des *Effective Core Potentials* [46], on perturbe le potentiel d'une façon telle que les orbitales ψ_i^r ne soient plus des fonctions propres de l'opérateur $-\frac{1}{2}\triangle + V_{nuc} + V_r^{ps} + \mathcal{K}(\Psi_r^v)$. On cherche en effet des pseudo-orbitales $\phi_i^{r,ps}$ solutions de $(\mathcal{P}_r^{ps})$ qui, d'une part s'identifient aux orbitales $\psi_i^{r,v}$ hors de $\mathcal{S}^c$, et d'autre part, oscillent beaucoup moins dans $\mathcal{S}^c$ que les orbitales $\psi_i^{r,v}$. Les pseudo-orbitales $\phi_i^{r,ps}$ ne coïncident plus dans $\mathcal{S}^c$ avec les orbitales $\psi_i^{r,v}$. Ainsi, la relation $(2.1f^r)$, non vérifiée par les pseudo-orbitales $\phi_i^{r,ps}$, n'a plus à être prise en compte.

Suite à l'approximation des cœurs gelés, il est naturel de penser que l'action du pseudo-potentiel V_r^{ps} sur la partie des orbitales ψ_i^v située dans $\mathcal{S}^c$ s'identifie à son action sur la partie des orbitales $\psi_i^{r,v}$ située dans $\mathcal{S}^c$. Ainsi, en pratique, on néglige ensuite la condition $(2.1f)$ lors du calcul sur le système moléculaire dans son ensemble avec le pseudo-potentiel V^{ps}. Bien sûr, une telle démarche reste à éclaircir d'un point de vue mathématique.

2.2.1 Les Potentiels Modèles

Les Potentiels Modèles ont été introduits pour des modèles de Hartree-Fock et sont donc très répandus dans le domaine de la chimie moléculaire. Leur principe est facilement applicable à des modèles de Kohn-Sham plus communs en physique du solide.

Dans un premier temps, on approche $\mathcal{K}(\Psi_r^c)$ par un potentiel local V_r^{pm} de la forme suivante

$$V_r^{pm} = \frac{N_c}{r} - \sum_{j=1}^{n_a} A_j \frac{e^{-\alpha_j r^2}}{r}.$$

Ensuite, on rajoute à l'opérateur $V_{nuc} + V_r^{pm}$ un terme supplémentaire V_r^{add} défini par

$$V_r^{add} = \sum_{i=1}^{N_c} \epsilon |\psi_i^{r,c}\rangle\langle\psi_i^{r,c}|,$$

avec ϵ tel que

$$\epsilon + \epsilon_1^{r,c} \gg \epsilon_{N_v}^{r,v}.$$

de manière à remonter suffisamment les niveaux d'énergie $\epsilon_i^{r,c}$, et donc, de s'affranchir de $(2.1f^r)$. Les paramètres $(A_j, \alpha_j)_{j=1,n_a}$ et n_a sont ajustés de façon à retrouver les orbitales $\psi_i^{r,v}$ et les niveaux $\epsilon_i^{r,v}$ lors de la résolution de $(\mathcal{P}_r^{ps})$ avec

$$V_r^{ps} = V_r^{pm} + V_r^{add}.$$

Par construction, les relations (11) et (12) sont vérifiées puisque les pseudo-orbitales $\phi_i^{r,ps}$ s'identifient aux orbitales $\psi_i^{r,v}$ partout. En outre, comme on ajuste les paramètres à partir de solutions issues de la résolution de $(\mathcal{P}_r^{ps})$, cette procédure est consistante avec le calcul de référence.

2.2.2 Effective Core Potentials

Les pseudo-orbitales $\phi_i^{r,ps}$, calculées par les méthodes à base de Potentiels Modèles, gardent le caractère oscillant des orbitales $\psi_i^{r,v}$ dans $\mathcal{S}^c$, ce qui, on le rappelle, entraîne un coût numérique important inutile. On se propose donc de construire un pseudo-potentiel V_r^{ps} de façon à ce que les pseudo-orbitales $\phi_i^{r,ps}$ solutions de $(\mathcal{P}_r^{ps})$ oscillent le moins possible dans $\mathcal{S}^c$, et coïncident avec les orbitales $\psi_i^{r,v}$ seulement en dehors de $\mathcal{S}^c$ [46].

2.2.2.1 Forme générale

Le problème $(\mathcal{P}_r)$ est associé à un système à symétrie sphérique. C'est pourquoi, on cherche en pratique V_r^{ps} sous la forme de l'opérateur de noyau

$$V_r^{ps}(r,r') = V_{loc}\big(|r|\big)\delta(r-r') + \sum_{l=0}^{l_{max}} \sum_{m=-l}^{l} |\mathcal{Y}_l^m\rangle \Big(V_l(|r|) - V_{loc}\big(|r|\big)\Big)\delta\Big(|r|-|r'|\Big)\langle \mathcal{Y}_l^m|, \tag{2.11}$$

L'entier l_{max} est généralement égal à l_0+1 où l_0 est le nombre cinétique le plus élevé atteint par un électron de valence de l'atome. La forme (2.11) se justifie lorsqu'on considère des électrons de valence appartenant à une même couche électronique [46]. Dans le cas contraire, on développe des méthodes plus complexes (Vanderbilt par exemple) de mise en œuvre très difficile [101]. C'est pourquoi, il est courant, notamment en chimie moléculaire, de garder la forme (13), qui conduit cependant à des pseudo-orbitales $\phi_i^{r,ps}$ plus oscillantes.

2.2.2.2 D'un point de vue numérique

On distingue trois approches pour l'obtention de pseudo-potentiels de type *ECP*.

La première approche repose sur un paramétrage de V_r^{ps} [15]. Les paramètres sont ajustés de façon à recouvrer les orbitales $\psi_i^{r,v}$ hors de $\mathcal{S}^c$ et les niveaux d'énergie $\epsilon_i^{r,v}$ par la résolution de $(\mathcal{P}_r^{ps})$. Cette approche est différente de l'approche des Potentiels Modèles où on cale le pseudo-potentiel V_r^{ps} de façon à recouvrer les orbitales partout. On garde toutefois la consistance avec le calcul de référence.

La seconde approche repose sur un paramétrage des pseudo-orbitales $\phi_i^{r,ps}$ [61, 72, 95, 100]. Les paramètres sont ajustés de façon à recouvrer les orbitales $\psi_i^{r,v}$ hors de $\mathcal{S}^c$, et, à minimiser la complexité numérique des pseudo-orbitales $\phi_i^{r,ps}$ dans $\mathcal{S}^c$. Ensuite, le pseudo-potentiel V_r^{ps} est obtenu en inversant les équations radiales du modèle (RHF, KS) à partir des pseudo-orbitales $\phi_i^{r,ps}$ construites et des niveaux d'énergie $\epsilon_i^{r,v}$ (voir ci-dessous). Une telle procédure est plus simple à mettre en œuvre que la précédente. Cependant, on n'est aucunement assuré que les couples $(\phi_i^{r,ps}, \epsilon_i^{r,v})$ sont solutions de $(\mathcal{P}_r^{ps})$, bien qu'ils vérifient les équations de $(\mathcal{P}_r^{ps})$.

Ces deux approches sont qualifiées "Shape Consistent" car les pseudo-orbitales $\phi_i^{r,ps}$ sont forcées numériquement à coïncider avec les orbitales $\psi_i^{r,v}$ hors de $\mathcal{S}^c$.

Enfin, des ECP dits "Energy Consistent" ont été développés dans le domaine de la chimie moléculaire [41]. La forme paramétrée du pseudo-potentiel V_r^{ps} est ajustée de façon à retrouver des différences d'énergie totale entre différentes configurations électroniques de l'atome. Le pseudo-potentiel V_r^{ps} construit par cette approche conduit à des pseudo-orbitales très similaires aux pseudo-orbitales issues des ECP "Shape Consistent".

Les Potentiels ECP "Shape Consistent" ont été développés indépendamment pour les modèles Hartree-Fock [34, 63] et les modèles DFT [15, 101]. Les deux techniques de construction présentées peuvent s'appliquer aux modèles HF et DFT. Dans la pratique, pour une modélisation Hartree-Fock, l'approche basée sur un fittage de V_r^{ps} est la plus répandue. Par contre, pour une modélisation DFT, l'approche basée sur le fittage des pseudo-orbitales est la plus utilisée.

2.2.2.3 Résumé schématique dans l'exemple de la génération d'un ECP "shape consistent" par fittage du pseudo-potentiel

- (i) On calcule les orbitales ψ_i^r avec précision, sur base grande, comme les solutions de $(\mathcal{P}_r)$.
- (ii) On choisit une forme paramétrée pour V_{loc} et V_l pour tout $0 \leq l \leq l_{max}$.
- (iii) On ajuste les paramètres introduits de façon à ce que les solutions $(\phi_i^{r,ps}, \epsilon_i^{r,ps})$ du problème $(\mathcal{P}_r^{ps})$ avec

$$V_r^{ps}\big(|r|,|r'|\big) = V_{loc}\big(|r|\big)\delta(r-r') + \sum_{l=0}^{l_{max}} \sum_{m=-l}^{l} |\mathcal{Y}_l^m\rangle \Big(V_l\big(|r|\big) - V_{loc}\big(|r|\big)\Big)\delta\Big(|r|-|r'|\Big)\langle \mathcal{Y}_l^m|.$$

vérifient

$$\begin{cases} \phi_i^{r,ps} = \psi_i^{r,v} \text{ à l'extérieur du cœur,} \\ \phi_i^{r,ps} \text{ a peu d'oscillations dans le cœur,} \\ \epsilon_i^{r,ps} = \epsilon_i^{r,v}. \end{cases}$$

Le détail de la pratique de cette étape (iii) reste un peu mystérieux.

- (iv) On résout $(\mathcal{P}^{ps})$ avec V^{ps} ainsi défini.

Rappelons que cette procédure de construction de V_r^{ps} est consistante avec le calcul de référence $(\mathcal{P}_r^{ps})$ car on ajuste les paramètres à partir des solutions de $(\mathcal{P}_r^{ps})$.

2.2.2.4 Résumé schématique dans l'exemple de la génération d'un ECP "shape consistent" par fittage des pseudo-orbitales

- (i) On calcule les orbitales ψ_i^r avec précision, sur base grande, comme les solutions de $(\mathcal{P}_r)$.

- (ii) On se donne une base petite, on cherche (par méthode des moindres carrés) des $\phi_i^{r,ps}$ sur cette petite base qui approchent au mieux

$$\begin{cases} \psi_i^{r,v} \text{ à l'extérieur du cœur,} \\ \phi_i^{r,ps} \text{ a peu d'oscillations dans le cœur.} \end{cases}$$

- (iii) On inverse les équations radiales du modèle adopté pour trouver V_r^{ps}, c'est-à-dire (dans un cas simple)
  - (a) pour chaque l, on note $i(l)$ l'orbitale de valence de nombre cinétique l, et de nombre quantique n le plus élevé.
  - (b) On pose, pour tout $0 \leq l \leq l_{max}$,

$$V_l\big(|r|\big) = \frac{\epsilon_{i(l)}^{r,v}\phi_{i(l)}^{r,ps} + \frac{1}{2}\triangle\phi_{i(l)}^{r,ps} - V_{nuc}\phi_{i(l)}^{r,ps} - \mathcal{K}(\Phi_r^{ps}).\phi_{i(l)}^{r,ps}}{\phi_{i(l)}^{r,ps}},$$

    où $\mathcal{K}$ peut indifféremment désigner $\mathcal{K}^{KS}$ ou $\mathcal{K}^{RHF}$. Le membre de droite est bien une fonction de $|r|$ (et non de r) car on considère seulement la partie radiale de la pseudo-orbitale $\phi_{i(l)}^{r,v}$ dans la formule donnant $V_l\big(|r|\big)$.
  - (c) On pose

$$V_{loc} = V_{l_{max}}.$$

  - (d) On construit

$$V_r^{ps}\big(|r|,|r'|\big) = V_{loc}\big(|r|\big)\delta(r-r') + \sum_{l=0}^{l_{max}}\sum_{m=-l}^{l} |\mathcal{Y}_l^m\rangle\Big(V_l\big(|r|\big) - V_{loc}\big(|r|\big)\Big)\delta\Big(|r|-|r'|\Big)\langle\mathcal{Y}_l^m|.$$

    Au membre de droite, le premier terme est local, le second est non local.
- (iv) On résout $(\mathcal{P}^{ps})$ avec V^{ps} ainsi défini.

Rappelons que cette procédure de construction de V_r^{ps} n'est pas consistante avec le calcul de référence $(\mathcal{P}_r^{ps})$, bien que les équations de $(\mathcal{P}_r^{ps})$ soient vérifiées par les couples $(\phi_i^{r,ps}, \epsilon_i^{r,v})$. En aucun cas, on ne sait montrer que $(\phi_i^{r,ps}, \epsilon_i^{r,ps})$ sont les solutions de $(\mathcal{P}_r^{ps})$.

2.2.2.5 La forme non séparable de Kleinman-Bylander

La partie semi-locale du pseudo-potentiel V_r^{ps} défini en (2.11) mène lors de la résolution du problème $(\mathcal{P}^{ps})$ après discrétisation sur une base $(\chi_\mu)_{i=1,N_b}$ au calcul de $N_b(N_b+1)/2$ intégrales de la forme

$$\forall\ 1 \leq \mu \leq \nu \leq N_b \quad \int_r \Big(\int_{\theta,\varphi} \mathcal{Y}_l^m \chi_\mu\Big)\Big(\int_{\theta,\varphi} \mathcal{Y}_l^m \chi_\nu\Big)(V_l - V_{loc})(r)\, dr.$$

Cette voie empruntée dans un premier temps [21] se révèle très coûteuse. Afin de réduire le coût de calcul de ces intégrales, Kleinman-Bylander ont proposé de

transformer la partie semi-locale du pseudo-potentiel V_r^{ps} [75] par

$$\sum_{l=0}^{l_{max}} \sum_{m=-l}^{l} \frac{\left|\phi_{lm}^{r,ps}(V_l - V_{loc})\right\rangle\left\langle\phi_{lm}^{r,ps}\left(V_l - V_{loc}\right)\right|}{\left\langle\phi_{lm}^{r,ps}\middle|V_l - V_{loc}\middle|\phi_{lm}^{r,ps}\right\rangle}$$

où chaque fonction $\phi_{lm}^{r,ps}$ désigne la fonction $\phi_i^{r,ps}\mathcal{Y}_l^m$, $\phi_i^{r,ps}$ correspondant aux nombres quantiques l et m. On est donc ramené au calcul de N_b intégrales de la forme

$$\forall\ 1 \leq \mu \leq N_b, \quad \int_{\mathbf{R}^3} \phi_{lm}^{r,ps}(V_l - V_{loc})\chi_\mu.$$

Le problème majeur d'une telle transformation réside dans la non localité du pseudo-potentiel généré. On observe l'apparition d'états, appelés *ghost states* correspondant à des fonctions propres nodales associées à des valeurs propres inférieures à celles associées aux $\phi_i^{r,ps}$ [59]. Afin de s'assurer de l'absence de tels états, l'alternative la plus répandue consiste à choisir la partie locale V_{loc} du pseudo-potentiel total égale à l'opérateur local V_l associé à l'orbitale de moment cinétique le plus élevé, soit V_{lmax} [46]. Ensuite, on ajuste le rayon de cœur r_c de manière à vérifier la non-existence d'états *fantômes* par une procédure développée dans [60].

2.3 Le traitement des systèmes moléculaires

On considère dans cette section un système moléculaire isolé comportant M atomes $(A_k)_{k=1,M}$ de numéro atomique $(z_k)_{1\leq k\leq M}$. Soient N le nombre total d'électrons du système et $(\bar{x}_k)_{k=1,M}$ les positions des noyaux. L'état fondamental électronique de ce système s'obtient en résolvant les équations RHF ou KS

$$(\mathcal{P}) \left\| \begin{array}{l} \text{Les } (\psi_j, \epsilon_j)_j \text{ sont les } N \text{ plus petits modes propres de} \\ \forall\ 1 \leq j \leq N, \quad -\frac{1}{2}\triangle\psi_j + V_{nuc}\psi_j + \mathcal{K}^{RHF,KS}(\Psi).\psi_j = \epsilon_j\psi_j, \\ \forall\ 1 \leq i,j \leq N, \quad (\psi_i, \psi_j) = \delta_{ij}, \end{array} \right. \tag{2.12}$$

où Ψ désigne la famille des fonctions $(\psi_j)_{1\leq j\leq N}$ et

$$\epsilon_1 \leq \ldots \leq \epsilon_j \leq \ldots \leq \epsilon_N.$$

Le potentiel nucléaire V_{nuc} est défini par

$$\forall\ x \in \mathbb{R}^3, \quad V_{nuc}(x) = \sum_{k=1}^{M} V_k(x) = -\sum_{k=1}^{M} \frac{z_k}{|x - \bar{x}_k|},$$

et

$$\forall\ x \in \mathbb{R}^3, \quad \rho_\Psi(x) = \sum_{j=1}^{N} |\psi_j(x)|^2,$$

$$\forall\, x, y \in \mathbb{R}^3, \quad \tau_\Psi(x,y) = \sum_{j=1}^{N} \psi_j(x)\psi_j^*(y).$$

Désormais, les orbitales ψ_j ne sont plus localisées autour d'un seul noyau. Toutefois lorsque les cœurs des atomes du système ne se recouvrent pas, on remarque qu'une partie des orbitales ψ_j (les orbitales de plus basses énergie) appartiennent *quasiment* au sous-espace vectoriel engendré par les orbitales de cœur des atomes calculées dans l'état de référence.
Afin d'illustrer cette propriété, on considère le système moléculaire formé de deux atomes de carbone. On désigne par μ la distance internucléaire et par ψ_{1s}^r l'orbitale de cœur $1s$ de l'atome de carbone isolé placé à l'origine. On note r_c le support de ψ_{1s}^r. Pour des valeurs de μ supérieures à $2r_c$ les deux orbitales ψ_1 et ψ_2 de plus basse énergie du problème $(\mathcal{P})$ sont représentées sur la Figure 2.1.

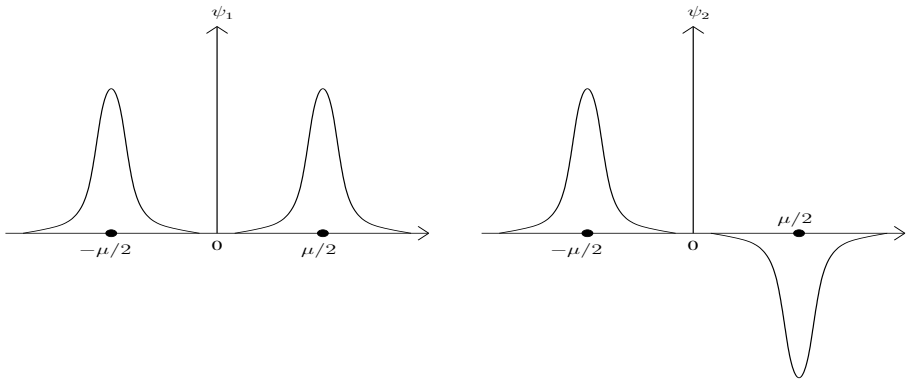


FIG. 2.1 – Profil des fonctions ψ_1 et ψ_2.

On observe numériquement pour tout $x \in \mathbb{R}^3$,

$$\begin{cases} \psi_1(x) & \simeq & \Big(\psi_{1s}^r(x+\mu/2) + \psi_{1s}^r(x-\mu/2)\Big)/\sqrt{2} \\ \psi_2(x) & \simeq & \Big(\psi_{1s}^r(x+\mu/2) - \psi_{1s}^r(x-\mu/2)\Big)/\sqrt{2} \end{cases}.$$

Soient $\phi_1 = (\psi_1 + \psi_2)/\sqrt{2}$ et $\phi_2 = (\psi_1 - \psi_2)/\sqrt{2}$ les fonctions représentées sur la Figure 2.2, on a pour tout $x \in \mathbb{R}^3$,

$$\begin{cases} \phi_1(x) & \simeq & \psi_{1s}^r(x+\mu/2) \\ \phi_2(x) & \simeq & \psi_{1s}^r(x-\mu/2) \end{cases}.$$

Par suite lorsqu'on considère un système moléculaire, on définit en pratique pour pour chaque atome :

- *un nombre d'électrons de cœur* N_k^c,
- un rayon de cœur $r_{c,k}$,
- un problème de référence $(\mathcal{P}_{r,k})$ (posé sur l'atome A_k isolé) pour lequel on calcule un pseudo-potentiel $V_{r,k}^{ps}$.

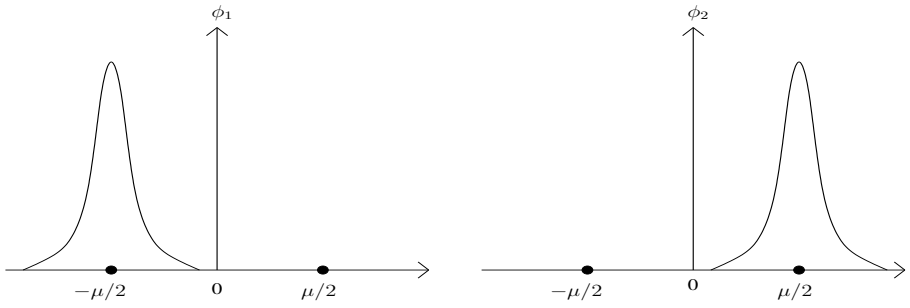


FIG. 2.2 – Profil des fonctions ϕ_1 et ϕ_2.

On approche ensuite le problème $(\mathcal{P})$ à N électrons par le problème $(\mathcal{P}^{ps})$ à $N^v = N - \sum_{k=1}^{M} N_k^c$ électrons défini par

$$(\mathcal{P}^{ps}) \left\| \begin{array}{l} \text{Les } (\phi_j^{ps}, \epsilon_j^{ps})_j \text{ sont les } N^v \text{ plus modes propres de} \\ \forall\ 1 \leq j \leq N_v, \quad -\frac{1}{2}\triangle\phi_j^{ps} + V_{nuc}\phi_j^{ps} + V^{ps}\phi_j^{ps} + \mathcal{K}(\Phi^{ps}).\phi_j^{ps} = \epsilon_j^{ps}\phi_j^{ps}, \\ \forall\ 1 \leq i,j \leq N_v, \quad (\phi_i^{ps}, \phi_j^{ps}) = \delta_{ij}, \end{array} \right.$$

avec

$$\epsilon_1^{ps} \leq \ldots \leq \epsilon_j^{ps} \leq \ldots \leq \epsilon_{N_v}^{ps}$$

et

$$V^{ps} = \sum_{k=1}^{M} V_{r,k}^{ps}. \tag{2.13}$$

Les configurations de référence pour chaque atome A_k sont choisies de sorte que

$$\forall\ 1 \leq j \leq N_v,\ \forall\ x \notin \mathcal{S}^c \quad \phi_j^{ps}(x) \simeq \psi_{j+N_c}^v(x), \tag{2.14}$$

$$\forall\ 1 \leq j \leq N_v, \quad \epsilon_j^{ps} \simeq \epsilon_{j+N_c}^v \tag{2.15}$$

avec $\mathcal{S}^c = \bigcup_{k=1}^{M} \mathcal{S}_k^c$.

2.4 Transférabilité des pseudo-potentiels construits

La notion de transférabilité introduite au début de ce chapitre s'identifie à la validité de (2.8), respectivement (2.13), au sens des propriétés (2.6-2.7), respectivement (2.14-2.15). A l'heure actuelle, le manque de formulation du problème ne permet pas de valider de façon rigoureuse, *a priori* ou *a posteriori*, l'introduction d'un pseudo-potentiel même pour les systèmes les plus simples. Toutefois, on peut citer certaines précautions à prendre en vue d'une bonne transférabilité du pseudo-potentiel construit.

2.4.1 Précautions indispensables

En premier lieu, la méthode des pseudo-potentiels est basée sur l'approximation des cœurs gelés, dont la validité nécessite de vérifier, pour chaque atome A_k,

- (i) la séparation claire entre électrons de cœur et de valence sur un plan spatial (peu de recouvrement entre les densités de valence et de cœur),
- (ii) la séparation claire entre électrons de cœur et de valence sur un plan énergétique (séparation des niveaux associés),
- (iii) la rigidité du cœur considéré (soit sa faible polarisabilité sous l'action du milieu extérieur).

Les conditions $(i-iii)$ constituent un premier guide très instructif sur les partitions cœur/valence envisageables pour les atomes du tableau périodique. En effet, on peut toujours considérer pour chaque atome un petit nombre d'électrons de cœur et ainsi rentrer parfaitement dans le cadre posé par $(i-iii)$. Cependant, une telle démarche reste peu satisfaisante du fait du grand nombre d'électrons de valence restants. La difficulté de la méthode des pseudo-potentiels réside dans le choix optimal (dans le sens où, pour des raisons de coût de calcul, on ne veut retenir qu'un minimum d'électrons de valence) de la partition cœur/valence en fonction des propriétés électroniques du système qu'on veut décrire avec précision. Un tel choix nécessite une connaissance physico-chimique *a priori* du système.

On dénomme par SC (*Small Core*) les pseudo-potentiels où les électrons de valence sont les électrons occupant les deux couches électroniques les plus externes de l'atome. De la même manière, on dénomme par LC (*Large Core*) les pseudo-potentiels où les électrons de valence sont les électrons occupant la couche électronique la plus externe de l'atome. Les pseudo-potentiels SC sont beaucoup plus transférables mais induisent un coût numérique souvent trop élevé (méthodes post-HF en $\mathcal{O}(N^7)$, or entre LC et SC, on double, voire plus, le nombre d'électrons explicitement pris en compte).

Généralement, pour les atomes de la première ligne du tableau périodique, les pseudo-potentiels LC respectent les conditions $(i-iii)$. En revanche, ce n'est plus vrai pour les métaux de transition (pas de séparation spatiale entre les orbitales $3d$ et $4s$), les métaux f (pas de séparation spatiale entre les orbitales $5f$ et $6s$) et les alcalins (ions très polarisables). Pour traiter ces systèmes, les seuls remèdes sont d'augmenter le nombre d'électrons de valence explicitement pris en compte, ou, d'ajouter des ingrédients supplémentaires issus d'un raisonnement de nature chimique (ajout d'un potentiel de polarisation [43], ...).

Parallèlement au choix de la partition cœur/valence, on constate une grande importance de la base de gaussienne considérée lors de la résolution de $(\mathcal{P}^{ps})$. En effet, du fait de la petite taille de la base (provenant de son caractère spectral), il est important de réoptimiser la base servant à la résolution de $\mathcal{P}^{ps}$ en fonction de V^{ps}. Généralement, les pseudo-orbitales $\phi_{ik}^{r,ps}$ sont approchées par des gaussiennes [63]. Ces gaussiennes définissent ensuite une base de résolution associée au pseudo-

potentiel V^{ps}. Lorsque la base d'ondes planes ou les bases d'orbitales atomiques numériques sont considérées, on a juste à diminuer le cut-off ou la finesse de la grille lors de l'utilisation de V^{ps}.

2.4.2 Chimie moléculaire

La chimie moléculaire a pour but de décrire des phénomènes très fins et localisés qui nécessitent un traitement électronique précis. Dans cette optique, les bases d'orbitales atomiques localisées sont très adaptées.

Ces bases sont des bases spectrales, de très petite taille. On constate que les pseudo-potentiels utilisés dépendent peu des rayons de cœur r_c^k et des configurations de référence des atomes A_k. Il en est de même pour la taille de la base qu'on utilise pour le calcul des pseudo-orbitales ϕ_{ik}^{ps}. Il n'y a donc aucun intérêt à optimiser un pseudo-potentiel pour une situation donnée de manière à diminuer de seulement quelques éléments la base atomique considérée.

En revanche, du fait de la complexité des méthodes utilisées ($\mathcal{O}(N^7)$), il est rarement envisageable de considérer des pseudo-potentiels SC pour décrire avec précision les phénomènes chimiques mis en jeu. Comme les partitions cœur/valence de type LC, gérables sur le plan numérique, sont très souvent des approximations trop fortes (violation des précautions d'usage (i)-(iii)), on s'intéresse au développement de stratégies supplémentaires autorisant le choix d'une partition LC dans de telles situations.

2.4.3 Physique du solide

La physique du solide a pour but de calculer des grandeurs physiques sur des systèmes périodiques de grande taille présentant généralement une délocalisation du nuage électronique. Traditionnellement, on développe les orbitales cristallines sur une base d'ondes planes.

Les pseudo-potentiels générés, et donc la taille de la base nécessaire au calcul des pseudo-orbitales, sont très sensibles aux variations des paramètres de définition (rayon de coupure, configuration de référence). Il semble, en revanche, que la justification de l'approche soit qu'en pratique, les grandeurs macroscopiques (donc moyennes) à l'équilibre qu'on cherche à calculer dépendent peu de ces paramètres. Il s'agit seulement d'approcher correctement la nature chimique des liaisons du système à l'équilibre.

Il est important de garder à l'esprit que plus on optimise le calcul d'un pseudo-potentiel pour bien décrire une situation particulière, plus il est difficile d'obtenir un pseudo-potentiel présentant de bonnes propriétés de transférabilité. Ainsi, on peut seulement envisager la transférabilité de ces pseudo-potentiels pour des situations chimiquement équivalentes au niveau des liaisons entre atomes.

2.5 Principe de la méthode PAW

L'utilisation des pseudo-potentiels traditionnels est très simple et très largement répandue. Une autre façon de traiter le caractère oscillant des orbitales près des noyaux consiste à considérer une base mixte : des orbitales atomiques sont utilisées dans le cœur de chaque atome et la base d'ondes planes sert de base d'approximation hors de ces régions. La méthode PAW fait partie de cette seconde famille de méthodes de génération de pseudo-potentiels. Elle est d'une grande précision mais reste difficile à mettre en œuvre. Introduite par Blöchl [22], cette technique combine la simplicité des pseudo-potentiels et la précision des méthodes mixtes. En particulier, les méthodes traditionnelles décrites dans la section précédente dérivent de cette approche sous certaines approximations. Dans cette section, on se propose donc de clarifier les hypothèses et les étapes du calcul d'un pseudo-potentiel par la méthode PAW. Cette étude est un préliminaire nécessaire et important en vue d'une future étude numérique et mathématique de cette méthode.

Dans le modèle de Kohn-Sham, le potentiel effectif

$$W_\rho(x) = -\sum_{k=1}^{M} \frac{z_k}{|x-\bar{x}_k|} + \left(\rho \star \frac{1}{|x|}\right)(x) + \mu_{xc}[\rho](x)$$

comporte des singularités en $1/|x|$ situées en les positions $\bar{x}_k$ des noyaux. En conséquence, les $N_o \geq N$ orbitales occupées[1] de Kohn-Sham $\{\psi_i\}_{1\leq i\leq N_o}$ solutions de

$$\left\{\begin{array}{l} H_\rho\psi_i = \epsilon_i\psi_i \qquad\qquad \epsilon_1 < \epsilon_2 \leq \epsilon_3 \leq \cdots \\ \\ (\psi_i,\psi_j)_{L^2} = \delta_{ij} \\ \\ \rho(x) = \displaystyle\sum_{i=1}^{+\infty} f_i|\psi_i(x)|^2 \\ \\ \left|\begin{array}{ll} f_i = 1 & \text{si } \epsilon_i < \varepsilon_F \\ 0 \leq f_i \leq 1 & \text{si } \epsilon_i = \varepsilon_F \\ f_i = 0 & \text{si } \epsilon_i > \varepsilon_F \end{array}\right. \\ \\ \displaystyle\sum_{i=1}^{+\infty} f_i = N \end{array}\right. \tag{2.16}$$

[1]Notons que N_o est *a priori* une inconnue du problème. C'est par définition le cardinal de l'ensemble des nombres d'occupation f_i apparaissant dans (2.16) qui sont non nuls. Il est égal à N si la HOMO est saturée ; dans le cas contraire N_o est majoré par (et en général égal à) $N-1+n_{\varepsilon_F}$ où n_{ε_F} désigne la dégénérescence du niveau de Fermi ε_F. Pour les systèmes neutres ou chargés positivement, il est plus ou moins prouvé que (2.16) a une solution et que N_o est bien défini car l'opérateur H_ρ admet une suite infinie de valeurs propres strictement négatives qui converge vers 0 (qui est l'infinum du spectre essentiel de cet opérateur).

où $H_\rho = -\frac{1}{2}\Delta + W$, sont susceptibles de présenter des *cusps* en les points $\{\bar{x}_k\}$, ce qui rend inefficace l'utilisation d'une base d'ondes planes. La méthode PAW repose sur la remarque suivante : si T un opérateur continu sur $L^2(\mathbb{R}^3)$ tel que $I+T$ soit inversible, les solutions $\{\psi_i\}_{1\leq i\leq N_0}$ de (2.16) s'obtiennent à partir des solutions $\left\{\widetilde{\psi}_i\right\}_{1\leq i\leq N_0}$ de

$$\left\{\begin{array}{l} \widetilde{H}_\rho\widetilde{\psi}_i = \epsilon_i \widetilde{S}\widetilde{\psi}_i \qquad\qquad \epsilon_1 < \epsilon_2 \leq \epsilon_3 \leq \cdots \\ \\ (\widetilde{\psi}_i, \widetilde{S}\widetilde{\psi}_j)_{L^2} = \delta_{ij} \\ \\ \rho(x) = \displaystyle\sum_{i=1}^{N} f_i \left(\left[(I+T)\widetilde{\psi}_i \right](x) \right)^2 \\ \\ \left| \begin{array}{ll} f_i = 1 & \text{si } \epsilon_i < \varepsilon_F \\ 0 \leq f_i \leq 1 & \text{si } \epsilon_i = \varepsilon_F \\ f_i = 0 & \text{si } \epsilon_i > \varepsilon_F \end{array} \right. \\ \\ \displaystyle\sum_{i=1}^{+\infty} f_i = N \end{array}\right. \tag{2.17}$$

où $\widetilde{H}_\rho = (I+T^T)H_\rho(I+T)$ et $\widetilde{S} = (I+T^T)(I+T)$, par la transformation

$$\psi_i = (I+T)\widetilde{\psi}_i. \tag{2.18}$$

Pour résoudre (2.16), on peut donc de façon équivalente résoudre (2.17) puis appliquer (2.18).
Le principe de la méthode PAW consiste à construire un opérateur T tel que

[C1] $(I+T)$ soit inversible,

[C2] le problème (2.17) possède des solutions $\widetilde{\psi}_i$ "régulières" (ce qui fait qu'il se prête mieux à une résolution par méthode spectrale).

Pour ce faire, on procède de la façon suivante :

1. pour chacun des éléments du tableau périodique, on construit un opérateur T à partir des orbitales de Kohn-Sham de cet élément (qu'on peut calculer numériquement avec une grande précision en résolvant un problème 3D radial) ; on note T_z l'opérateur T relatif à un atome isolé de numéro atomique z dont le noyau est placé à l'origine de l'espace ; l'opérateur T_z sera choisi tel que

$$\forall u \in L^2(\mathbb{R}^3), \qquad \forall x \in \mathbb{R}^3 \setminus B_{r_c^z}(0), \qquad (T_z u)(x) = 0, \tag{2.19}$$

où r_c^z désigne un rayon de coupure (dans la formule ci-dessus et dans la suite de ce document, la notation $B_r(\bar{x})$ désigne la boule de $\mathbb{R}^3$ de centre $\bar{x}$ et de

rayon r). Autrement dit, l'opérateur $(I+T_z)$ ne modifie les fonctions que dans la "région d'augmentation" $B_{r_c^z}(0)$. Notons que cette étape est effectuée une fois pour toutes : les caractéristiques des opérateurs T_z sont générées par un code atomique indépendant (tel celui décrit dans [68]), et stockées dans un fichier ;

2. pour un système moléculaire quelconque comprenant M noyaux de charges $z_1, \cdots, z_M$ situés en les points $\bar{x}_1, \cdots, \bar{x}_M$ de l'espace, on définit l'opérateur T par la formule

$$T = \sum_{k=1}^{M} \tau_{\bar{x}_k} T_{z_k} \tag{2.20}$$

où $\tau_{\bar{x}_k} T_{z_k}$ désigne le translaté de T_{z_k} en $\bar{x}_k$, c'est-à-dire l'opérateur défini sur $L^2(\mathbb{R}^3)$ par

$$\forall\, u \in L^2(\mathbb{R}^3),\ \forall\, x \in \mathbb{R}^3, \qquad \Big(\big(\tau_{\bar{x}_k} T_{z_k}\big)u\Big)(x) = \Big(T_{z_k}\big(u(\cdot + \bar{x}_k)\big)\Big)(x - \bar{x}_k). \tag{2.21}$$

Remarquons que tant que la condition

$$1 \leq k < l \leq M, \qquad B_{r_c^{z_k}}(\bar{x}_k) \cap B_{r_c^{z_l}}(\bar{x}_l) = \emptyset \tag{2.22}$$

de non-recouvrement des régions d'augmentation est vérifiée, il est facile de voir en utilisant (2.19) que la condition [C1] est satisfaite pour l'opérateur T défini par (2.20) dès que cette condition est vérifiée pour chacun des opérateurs T_{z_k}.

En revanche, il n'est pas vrai en général que la condition [C2] est vérifiée pour le système moléculaire si elle l'est pour chacun des atomes qui le composent. Tout le problème est de construire des opérateurs atomiques T_z possèdant une bonne "transférabilité" vis-à-vis de la condition [C2].

2.5.1 Construction des opérateurs atomiques T_z

Considérons un atome comportant $N = z$ électrons. Un opérateur T_z réalisant les conditions [C1] et [C2] pour un calcul sur cet atome peut être construit de la façon suivante :

1. on se donne
   (a) un réel r_c^z qui représente le rayon d'une sphère autour de l'atome considéré (cette sphère représente la zone dans laquelle les grandeurs électroniques sont supposées être peu influencées par l'environnement extérieur).
   (b) une suite $\{\phi_{z,\mu}\}_{1 \leq \mu \leq N_p}$ de N_p fonctions de $L^2(\mathbb{R}^3)$ telle que pour tout $1 \leq \mu \leq N_p$

$$\phi_{z,\mu}(x) = \sum_{\nu=1}^{\mu} \alpha_{\mu\nu} \phi^0_{z,\nu}(x) \tag{2.23}$$

où les coefficients $\alpha_{\mu\nu}$ sont des réels et où les $\left\{\phi^0_{z,\nu}\right\}_{1\leq\nu\leq N_p}$ désignent les N_p orbitales de Kohn-Sham de plus basse énergie, solutions de

$$\left\{\begin{array}{l} H^z_\rho \phi^0_{z,\nu} = \epsilon^0_{z,\nu}\phi^0_{z,\nu} \qquad\qquad \epsilon^0_{z,1} < \epsilon^0_{z,2} \leq \epsilon^0_{z,3} \leq \cdots \\ \\ (\phi^0_{z,\mu}, \phi^0_{z,\nu})_{L^2} = \delta_{\mu\nu} \\ \\ \rho(x) = \displaystyle\sum_{\nu=1}^{+\infty} f_\nu |\phi^0_{z,\nu}(x)|^2 \\ \\ \left|\begin{array}{ll} f_\nu = 1 & \text{si } \epsilon^0_{z,\nu} < \varepsilon^0_{z,F} \\ 0 \leq f_\nu \leq 1 & \text{si } \epsilon^0_{z,\nu} = \varepsilon^0_{z,F} \\ f_\nu = 0 & \text{si } \epsilon^0_{z,\nu} > \varepsilon^0_{z,F} \end{array}\right. \\ \\ \displaystyle\sum_{\nu=1}^{+\infty} f_\nu = N \end{array}\right. \tag{2.24}$$

avec

$$H^z_\rho = -\frac{1}{2}\Delta - \frac{z}{|x|} + \left(\rho \star \frac{1}{|x|}\right) + \mu_{xc}[\rho];$$

(c) une suite $\left\{\widetilde{\phi}_{z,\nu}\right\}_{1\leq\nu\leq N_p}$ de fonctions de $L^2(\mathbb{R}^3)$ "régulières" (i.e. développables sur une petite base d'ondes planes) telle que

$$\forall\, 1 \leq \nu \leq N_p,\ \forall\, x \in \mathbb{R}^3 \setminus B_{r^z_c}(0), \quad \widetilde{\phi}_{z,\nu}(x) = \phi_{z,\mu}(x); \tag{2.25}$$

(d) une suite $\{\widetilde{p}_{z,\nu}\}_{1\leq\nu\leq N_p}$ de fonctions de $L^2(\mathbb{R}^3)$ "duale" de la suite $\left\{\widetilde{\phi}_{z,\nu}\right\}_{1\leq\nu\leq N_p}$, autrement dit vérifiant

$$1 \leq \mu, \nu \leq N_p, \qquad (\widetilde{p}_{z,\mu}, \widetilde{\phi}_{z,\nu})_{L^2} = \delta_{\mu\nu};$$

2. on définit l'opérateur T_z (qu'on suppose inversible, voir remarque 1 ci-dessous) par

$$\forall\, u \in L^2(\mathbb{R}^3), \qquad T_z u = \sum_{\nu=1}^{N_p} (\widetilde{p}_{z,\nu}, u)_{L^2}\, (\phi_{z,\nu} - \widetilde{\phi}_{z,\nu}). \tag{2.26}$$

Il est clair que pour tout $1 \leq \nu \leq N_p$,

$$(I + T_z)\widetilde{\phi}_{z,\nu} = \phi_{z,\nu}.$$

Il en résulte que les solutions $\left\{\widetilde{\psi}_i\right\}_{1\leq i\leq N}$ de

$$\left\{\begin{array}{l} \widetilde{H}^z_\rho\widetilde{\psi}_i = \epsilon_i\widetilde{S}\widetilde{\psi}_i \qquad\qquad \epsilon_1 < \epsilon_2 \leq \epsilon_3 \leq \cdots \\ \\ (\widetilde{\psi}_i, \widetilde{S}\widetilde{\psi}_j)_{L^2} = \delta_{ij} \\ \\ \rho(x) = \displaystyle\sum_{i=1}^{+\infty} f_i \left(\left[(I+T_z)\widetilde{\psi}_i\right](x)\right)^2, \\ \\ \left|\begin{array}{ll} f_i = 1 & \text{si } \epsilon_i < \varepsilon_{z,F} \\ 0 \leq f_i \leq 1 & \text{si } \epsilon_i = \varepsilon_{z,F} \\ f_i = 0 & \text{si } \epsilon_i > \varepsilon_{z,F} \end{array}\right. \\ \\ \displaystyle\sum_{i=1}^{+\infty} f_i = N \end{array}\right. \tag{2.27}$$

avec

$$\widetilde{H}^z_\rho = (I+T_z^T)H^z_\rho(I+T_z) \qquad \text{et} \qquad \widetilde{S} = (I+T_z^T)(I+T_z),$$

sont des combinaisons linéaires des fonctions $\left\{\widetilde{\phi}_{z,\nu}\right\}_{1\leq\nu\leq N}$. Plus précisément,

$$\widetilde{\psi}_i(x) = \sum_{\nu=1}^{i} \beta_{i\nu}\widetilde{\phi}_{z,\nu}(x)$$

et

$$\forall\, 1 \leq i \leq \infty, \quad \epsilon_i = \epsilon_{z,i}.$$

où la matrice triangulaire inférieure $[\beta]$ est l'inverse de la matrice triangulaire inférieure $[\alpha]$ apparaissant dans (2.23). Le problème (2.27) est donc facile à résoudre sur une base d'ondes planes puisque les solutions $\left\{\widetilde{\psi}_i\right\}_{1\leq i\leq N}$ de ce problème sont des combinaisons linéaires des fonctions "régulières" (car choisies comme telles) $\left\{\widetilde{\phi}_{z,\nu}\right\}_{1\leq\nu\leq N}$.

La méthode de construction d'un opérateur T_z relatif au cas atomique que nous venons de décrire est plus un cadre conceptuel qu'un algorithme. Nous décrivons ci-dessous le procédé de construction proposé dans [68] (qui s'inspire très fortement de celui décrit dans [22]). Selon les auteurs, ce procédé fournit des opérateurs atomiques T_z "transférables" : à partir de ces opérateurs atomiques, on peut construire par la formule (2.20) un opérateur T pour un système moléculaire quelconque (vérifiant toutefois la condition (2.22) de non-recouvrement des régions d'augmentation) tel que les conditions [C1] et [C2] soient vérifiées. Ce procédé de construction se décompose en quatre étapes

– **1ère étape : construction des fonctions $\phi^{\mathbf{0}}_{\mathbf{z},\nu}$.**

On résout le problème de Kohn-Sham (2.24) en mettant à profit la symétrie sphérique pour se ramener à un calcul 3D radial. On obtient ainsi des orbitales de la forme

$$\phi^0_{z,\nu}(x) = \frac{\phi^0_{z,n_\nu l_\nu}(r)}{r}\, \mathcal{Y}^{m_\nu}_{l_\nu}(\theta,\varphi) \tag{2.28}$$

En pratique, on s'arrange pour maintenir la symétrie sphérique au cours du calcul SCF en imposant des contraintes sur les nombres d'occupation : supposons qu'à la k-ième itération SCF, la densité ρ_k soit à symétrie sphérique ; l'opérateur $H^z_{\rho_k}$ est alors à symétrie sphérique et on obtient une base de fonctions propres de la forme

$$\phi^{k+1}_{n,l,m}(x) = \frac{\phi^{k+1}_{n,l}(r)}{r}\, \mathcal{Y}^m_l(\theta,\varphi).$$

Si on impose que pour tout (n,l), les nombres d'occupation $f_{n,l,m}$ sont indépendants de m, on obtient une densité

$$\begin{aligned}
\rho_{k+1}(x) &= \sum_{n,l,m} f_{nlm} |\phi^{k+1}_{n,l,m}(x)|^2 \\
&= \sum_{n,l} f_{nl} \left| \frac{\phi^{k+1}_{n,l}(r)}{r} \right|^2 \sum_{-l\le m\le l} |\mathcal{Y}^m_l(\theta,\varphi)|^2 \\
&= \sum_{n,l} (2l+1)\, f_{nl} \frac{|\phi^{k+1}_{n,l}(r)|^2}{4\pi r^2}
\end{aligned}$$

qui est encore à symétrie sphérique. A la convergence, la densité s'écrit

$$\rho_{SCF}(x) = \sum_{\nu=1}^{+\infty} w_{n_\nu l_\nu} \frac{|\phi^0_{z,n_\nu l_\nu}(r)|^2}{4\pi r^2}. \tag{2.29}$$

– **2ème étape : construction des fonctions $\widetilde{\phi}^{\mathbf{0}}_{\mathbf{z},\nu}$.**

On choisit

1. un réel positif r^z_c (le rayon de la sphère d'augmentation) ;
2. une fonction régulière $k : \mathbb{R}^+ \to \mathbb{R}^+$ de classe C^1 vérifiant

$$\left\{ \begin{array}{l} k(0) = 1 \\ k'(0) = 0 \end{array} \right. , \qquad \left\{ \begin{array}{l} k(r^z_c) = 0 \\ k'(r^z_c) = 0 \end{array} \right. , \qquad \text{et} \qquad \forall\, r \ge r^z_c, \quad k(r) = 0.$$

La fonction k utilisée dans [68] est définie par

$$\forall\, 0 \le r \le r^z_c, \qquad k(r) = \left[\frac{\sin(\pi\, r/r^z_c)}{(\pi\, r/r^z_c)} \right]^2 ;$$

3. une constante réelle $\mathcal{V}_0$,

et on cherche les fonctions $\left\{\widetilde{\phi}^0_{z,\nu}\right\}_{1\leq\nu\leq N_p}$ de la forme

$$\widetilde{\phi}^0_{z,\nu}(x) = \frac{\widetilde{\phi}^0_{z,n_\nu l_\nu}(r)}{r}\,\mathcal{Y}^{m_\nu}_{l_\nu}(\theta,\varphi) \tag{2.30}$$

solutions du problème

$$\begin{cases} \left(H^{PS}_{\widetilde{\rho}} - \epsilon^0_{z,\nu}\right)\widetilde{\phi}^0_{z,\nu} = C_\nu\, k(r)\,\widetilde{\phi}^0_{z,\nu} \\ \widetilde{\phi}^0_{z,\nu}(r^z_c) = \phi^0_{z,\nu}(r^z_c) \\ \dfrac{\partial\widetilde{\phi}^0_{z,\nu}}{\partial r}(r^z_c) = \dfrac{\partial\phi^0_{z,\nu}}{\partial r}(r^z_c) \\ \widetilde{\rho}(x) = \widetilde{\rho}(r) = \displaystyle\sum_{\nu=1}^{+\infty} w_{n_\nu l_\nu}\frac{|\widetilde{\phi}^0_{z,n_\nu l_\nu}(r)|^2}{4\pi r^2}. \end{cases} \tag{2.31}$$

avec

$$H^{PS}_{\widetilde{\rho}} = -\frac{1}{2}\Delta + \widetilde{v}^{eff}_{\widetilde{\rho}}$$

$$\widetilde{v}^{eff}_{\widetilde{\rho}} = \widetilde{v}_{loc} + \left((\widetilde{\rho}+\hat{\rho}_{\widetilde{\rho}})\star\frac{1}{|x|}\right) + \mu_{xc}(\widetilde{\rho})$$

$$\widetilde{v}_{loc}(x) = \mathcal{V}_0\, k(|x|)$$

$$\hat{\rho}_{\widetilde{\rho}}(x) = Q^{00}_{\widetilde{\rho}}\, g_{00}(x)$$

$$g_{00}(x) = \frac{k(|x|)}{4\pi\int_0^{+\infty} r^2\,k(r)\,dr} \qquad \left(\text{de sorte que } \int_{\mathbf{R}^3} g_{00} = \int_{B_{r^z_c}(0)} g_{00} = 1\right)$$

$$Q^{00}_{\widetilde{\rho}} = -z + \int_{B_{r^z_c}(0)} (\rho_{SCF} - \widetilde{\rho}).$$

Notons que les nombres d'occupation $w_{n_\nu l_\nu}$ intervenant dans (2.31) sont fixés : ce sont ceux intervenant dans l'expression (2.29) de la densité solution du problème de Kohn-Sham considéré lors de la construction des $\phi^0_{z,\nu}$.

Comme la fonction k est nulle en dehors de la boule $B_{r^z_c}(0)$, et qu'il résulte du théorème de Gauss que

$$\begin{aligned} \forall\, x\in\mathbb{R}^3\setminus B_{r^z_c}(0), \qquad \left((\widetilde{\rho}|_{B_{r^z_c}(0)} + \hat{\rho}_{\widetilde{\rho}})\star\frac{1}{|x|}\right)(x) &= \frac{-z+\int_{B_{r^z_c}(0)}\rho_{SCF}}{|x|} \\ &= -\frac{z}{|x|} + \left(\rho_{SCF}|_{B_{r^z_c}(0)}\star\frac{1}{|x|}\right)(x), \end{aligned}$$

toute solution de (2.31) vérifie sur $\mathbb{R}^3 \setminus B_{r_c^z}(0)$

$$\begin{cases} \left(-\frac{1}{2}\Delta - \frac{z}{|x|} + \left(\left(\rho_{SCF}|_{B_{r_c^z}(0)} + \widetilde{\rho}|_{\mathbf{R}^3\setminus B_{r_c^z}(0)}\right) \star \frac{1}{|x|}\right)(x) + \mu_{xc}[\widetilde{\rho}](x)\right)\widetilde{\phi}^0_{z,\nu} = \epsilon^0_{z,\nu}\widetilde{\phi}^0_{z,\nu} \\ \widetilde{\phi}^0_{z,\nu}(r_c^z) = \phi^0_{z,\nu}(r_c^z) \\ \frac{\partial\widetilde{\phi}^0_{z,\nu}}{\partial r}(r_c^z) = \frac{\partial\phi^0_{z,\nu}}{\partial r}(r_c^z) \\ \widetilde{\rho}(x) = \widetilde{\rho}(r) = \sum_{\nu=1}^{+\infty} w_{n_\nu l_\nu}\frac{|\widetilde{\phi}^0_{z,n_\nu l_\nu}(r)|^2}{4\pi r^2}. \end{cases} \tag{2.32}$$

Si on se place dans le cadre d'une fonctionnelle d'échange-corrélation *locale* (type LDA ou GGA), on vérifie que la restriction des $\phi^0_{z,\nu}$ à l'ensemble $\mathbb{R}^3 \setminus B_{r_c^z}(0)$ fournit une solution de (2.32). Par conséquent, (2.25) est vérifiée. Il reste à résoudre sur $B_{r_c^z}(0)$ le problème (2.31). En effectuant le changement de variable

$$u_\nu(r) = r\widetilde{\phi}_{z,n_\nu l_\nu}(r)$$

ce problème s'écrit

$$\begin{cases} -\frac{1}{2}u_\nu''(r) + \frac{l_\nu(l_\nu+1)}{r^2}u_\nu(r) + \widetilde{v}^{eff}_{\widetilde{\rho}}(r)u_\nu(r) - \epsilon^0_{z,\nu}u_\nu(r) = C_\nu\, k(r)\, u_\nu(r) \\ u_\nu(0) = 0 \\ u_\nu(r_c^z) = r_c^z\phi^0_{z,n_\nu l_\nu}(r_c^z) \\ u_\nu'(r_c^z) = r_c^z\frac{d\phi^0_{z,n_\nu l_\nu}}{dr}(r_c^z) + \phi^0_{z,n_\nu l_\nu}(r_c^z) \\ \widetilde{\rho}(r) = \frac{1}{4\pi}\sum_{\nu=1}^{+\infty} w_{n_\nu l_\nu}|u_\nu(r)|^2. \end{cases} \tag{2.33}$$

Notons que les solutions de

$$\begin{cases} -\frac{1}{2}u''(r) + W(r)u(r) = C_\nu\, k(r)\, u(r) \\ u(0) = 0 \\ u(r_c^z) = r_c^z\phi^0_{z,n_\nu l_\nu}(r_c^z) \\ u'(r_c^z) = r_c^z\frac{d\phi^0_{z,n_\nu l_\nu}}{dr}(r_c^z) + \phi^0_{z,n_\nu l_\nu}(r_c^z) \end{cases} \tag{2.34}$$

sont de la forme

$$u(r) = r_c^z \phi^0_{z,n_\nu l_\nu}(r_c^z)\; v(r)/v(r_c^z)$$

où v est une solution non nulle du problème aux valeurs propres généralisé

$$\begin{cases} -\frac{1}{2}v''(r) + W(r)v(r) = \lambda\, k(r)\, v(r) \\ \\ v(0) = 0 \\ \\ v'(r_c^z) = \left(\dfrac{1}{\phi^0_{z,n_\nu l_\nu}(r_c^z)} \dfrac{d\phi^0_{z,n_\nu l_\nu}}{dr}(r_c^z) + \dfrac{1}{r_c^z} \right)\; v(r_c^z) \end{cases}$$

qui possède pour tout $n \in \mathbb{N}$, une et une seule solution[2] ayant exactement n zéros dans $]0, r_c^z[$. Lors des itérations SCF intervenant dans la résolution de (2.33), on choisira la solution u_ν qui a "le bon nombre de zéros" : pour chaque valeur du nombre quantique l, on classe les $\epsilon^0_{z,\nu}$ correspondant aux fonctions $\phi^0_{z,\nu}$ telles que $l_\nu = l$ par valeurs croissantes et on impose que la fonction u_ν correspondant à la n_ν-ième plus petite valeur $\epsilon^0_{z,\nu}$ ait exactement $n-1$ nœuds dans $]0, r_c^z[$.

– **3ème étape : construction des fonctions $\widetilde{\mathbf{p}}^{\mathbf{0}}_{\mathbf{z},\nu}$.**

On pose

$$\widetilde{p}^0_{z,\nu}(x) = \frac{k(|x|)\, \widetilde{\phi}^0_{z,\nu}(x)}{\langle \widetilde{\phi}^0_{z,\nu} | k | \widetilde{\phi}^0_{z,\nu} \rangle}$$

autrement dit

$$\widetilde{p}^0_{z,\nu}(x) = \frac{k(r)\, \widetilde{\phi}^0_{z,n_\nu l_\nu}(r)}{r \int_0^{+\infty} 4\pi s^2\, k(s)\, |\widetilde{\phi}^0_{z,n_\nu l_\nu}(s)|^2\, ds}\; \mathcal{Y}_l^m(\theta,\varphi) \tag{2.35}$$

de façon à avoir en particulier

$$\left(H^{PS}_{\widetilde{\rho}} - \epsilon^0_{z,\nu}\right) \widetilde{\phi}^0_{z,\nu} = \widetilde{p}^0_{z,\nu}\; \langle \widetilde{\phi}^0_{z,\nu} | H^{PS}_{\widetilde{\rho}} - \epsilon^0_{z,\nu} | \widetilde{\phi}^0_{z,\nu} \rangle \qquad \text{et} \qquad \langle \widetilde{\phi}^0_{z,\nu} | \widetilde{p}^0_{z,\nu} \rangle = 1.$$

Les fonctions $\widetilde{p}^0_{z,\nu}$ ainsi définies sont à support dans $B_{r_c^z}(0)$.

– **4ème étape : orthonormalisation.**

On pose

$$\phi_{z,1} = \phi^0_{z,1}, \qquad \widetilde{\phi}_{z,1} = \widetilde{\phi}^0_{z,1}, \widetilde{p}_{z,1} = \widetilde{p}^0_{z,1},$$

puis on orthonormalise les fonctions d'indice $\nu \geq 2$ de façon à obtenir des fonctions $\widetilde{\phi}_{z,\nu}$ et $\widetilde{p}_{z,\nu}$ vérifiant

$$1 \leq \mu, \nu \leq N_p, \qquad (\widetilde{p}_{z,\mu}, \widetilde{\phi}_{z,\nu})_{L^2} = \delta_{\mu\nu}.$$

[2] cette solution étant définie bien entendu à une constante multiplicative près

On utilise pour cela un algorithme de type Gramm-Schmidt :

$$\widetilde{p}_{z,\mu} = \mathcal{F}_{z,\mu}\ \left(\widetilde{p}^0_{z,\mu} - \sum_{\nu=1}^{\mu-1} (\widetilde{p}^0_{z,\mu}, \widetilde{\phi}_{z,\nu})_{L^2}\, \widetilde{p}_{z,\nu}\right)$$

($\widetilde{p}_{z,\mu}$ est ainsi orthogonal à tous les $\widetilde{\phi}_{z,\nu}$ pour $1 \leq \nu < \mu$)

$$\widetilde{\phi}_{z,\mu} = \mathcal{F}_{z,\mu}\ \left(\widetilde{\phi}^0_{z,\mu} - \sum_{\nu=1}^{\mu-1} (\widetilde{p}_{z,\nu}, \widetilde{\phi}^0_{z,\mu})_{L^2}\, \widetilde{\phi}_{z,\nu}\right)$$

($\widetilde{\phi}_{z,\mu}$ est ainsi orthogonal à tous les $\widetilde{p}_{z,\nu}$ pour $1 \leq \nu < \mu$)

$$\phi_{z,\mu} = \mathcal{F}_{z,\mu}\ \left(\phi^0_{z,\mu} - \sum_{\nu=1}^{\mu-1} (\widetilde{p}_{z,\nu}, \widetilde{\phi}^0_{z,\mu})_{L^2}\, \phi_{z,\nu}\right),$$

ceci pour préserver l'égalité

$$\forall\ x \in \mathbb{R}^3 \setminus B_{r_c^z}(0), \qquad \widetilde{\phi}_{z,\nu}(x) = \phi_{z,\nu}(x). \tag{2.36}$$

Le coefficient $\mathcal{F}_{z,\mu}$ est ajusté pour avoir

$$(\widetilde{p}_{z,\mu}, \widetilde{\phi}_{z,\mu})_{L^2} = 1.$$

Cette expression donne pour $\mu = 2$

$$\mathcal{F}_{z,\mu} = \left(1 - (\widetilde{p}_1, \widetilde{\phi}^0_2)_{L^2} (\widetilde{p}^0_2, \widetilde{\phi}_1)_{L^2}\right)^{-1/2}$$

comme indiqué dans [68].

Notons qu'en raison de la forme (2.28), (2.30), (2.35) des fonctions $\phi_{z,\nu}$, $\widetilde{\phi}_{z,\nu}$ et $\widetilde{p}_{z,\nu}$, il suffit évidemment d'orthogonaliser entre elles les fonctions radiales correspondant à une valeur donnée de l.

On retiendra que

- les fonctions $\widetilde{p}_{z,\nu}$ sont à support dans $B_{r_c^z}(0)$;
- les fonctions $\widetilde{\phi}_{z,\nu}$ et $\phi_{z,\nu}$ sont égales en dehors de $B_{r_c^z}(0)$.

Notamment, la construction de ces fonctions dans $B_{r_c^z}(0)$ et non dans $\mathbb{R}^3$ est nécessaire pour définir l'opérateur T_z.

Remarque 1. L'opérateur T_z défini par (2.26) est de rang fini, donc compact. Il en résulte que l'opérateur $I+T_z$ est inversible si et seulement si il est injectif (alternative de Fredholm). Or il est facile de vérifier que si les deux conditions suivantes sont satisfaites :

1. les N_p fonctions $(\phi_{z,\nu} - \widetilde{\phi}_{z,\nu})$ sont linéairement indépendantes ;

2. la matrice $[(\widetilde{p}_{z,\mu}, \phi_{z,\nu})]$ est inversible,

alors l'opérateur $I+T_z$ est injectif. Remarquons enfin que si 1 n'est pas valeur propre de la matrice $[(\widetilde{p}_{z,\mu}, \phi_{z,\nu})]$, alors les $(\phi_{z,\nu} - \widetilde{\phi}_{z,\nu})$ sont linéairement indépendantes.

Remarque 2. De manière générale, il est indispensable de choisir un procédé de construction pour les $\widetilde{\phi}_{z,\nu}$ qui mène à des fonctions beaucoup plus régulières que les $\phi_{z,\nu}$ dans $B_{r_c^z}(0)$. D'une part, ceci n'est pas immédiat pour le procédé décrit ci-dessus et donc demande un réel savoir-faire dans le choix des divers paramètres introduits ($k(r)$, $\mathcal{V}_0$, ...). D'autre part, on pourrait penser à utiliser les procédés de construction utilisés lors de la construction des pseudo-potentiels à norme conservée ou des procédés de construction basés sur la résolution de problèmes de nature très différente du problème (2.31).

2.5.2 Méthode PAW pour les systèmes moléculaires

Considérons maintenant le système moléculaire dans son intégralité, on suppose que la condition de non recouvrement (2.22) est satisfaite.

Pour des opérateurs atomiques T_z de la forme (2.19), les opérateurs translatés $\tau_{\bar{x}_k} T_{z_k}$ définis par (2.21) se calculent selon la formule

$$\forall\, u \in L^2(\mathbb{R}^3),\ \forall\, x \in \mathbb{R}^3, \qquad \left((\tau_{\bar{x}_k} T_{z_k})u\right)(x) = \sum_{\nu=1}^{N_{p,k}} (\widetilde{p}_{k,\nu}(\cdot - \bar{x}_k), u)_{L^2} \left(\phi_{k,\nu}(x-\bar{x}_k) - \widetilde{\phi}_{k,\nu}(x-\bar{x}_k)\right).$$

Pour simplifier les notations, on pose $N_k = N_{p,k}$

$$\phi_\nu^k = \phi_{z_k,\nu}(\cdot - \bar{x}_k), \qquad \widetilde{\phi}_\nu^k = \widetilde{\phi}_{z_k,\nu}(\cdot - \bar{x}_k), \qquad \widetilde{p}_\nu^k = \widetilde{p}_{z_k,\nu}(\cdot - \bar{x}_k), \qquad \bar{\phi}_\nu^k = \phi_\nu^k - \widetilde{\phi}_\nu^k,$$

$$J_k u = \sum_{\nu=1}^{N_{p,k}} (\widetilde{p}_\nu^k, u)_{L^2}\, \phi_\nu^k, \qquad \widetilde{J}_k u = \sum_{\nu=1}^{N_{p,k}} (\widetilde{p}_\nu^k, u)_{L^2}\, \widetilde{\phi}_\nu^k, \qquad \text{et} \qquad T_k u = \tau_{\bar{x}_k} T_{z_k} u = J_k u - \widetilde{J}_k u,$$

de sorte que

$$T = \sum_{k=1}^{M} (J_k - \widetilde{J}_k) = \sum_{k=1}^{M} T_k.$$

Les adjoints de J_k, $\widetilde{J}_k$ et T_k sont donnés respectivement par

$$J_k^T = \sum_{\nu=1}^{N_k} (\phi_\nu^k, \cdot)\, \widetilde{p}_\nu^k, \qquad \widetilde{J}_k^T = \sum_{\nu=1}^{N_k} (\widetilde{\phi}_\nu^k, \cdot)\, \widetilde{p}_\nu^k, \qquad \text{et} \qquad T_k^T = \sum_{\nu=1}^{N_k} (\bar{\phi}_\nu^k, \cdot)\, \widetilde{p}_\nu^k.$$

Par la condition (2.25), on a

$$\forall\, 1 \le k \le M, \quad \forall\, x \in \mathbb{R}^3 \setminus B_{r_c^{z_k}}(\bar{x}_k), \quad \left(\widetilde{J}_k u\right)(x) = \left(J_k u\right)(x). \tag{2.37}$$

Par ailleurs, on fait à ce niveau l'hypothèse que les restrictions à $B_{r_c^{z_k}}(\bar{x}_k)$ des $\widetilde{\phi}_\nu^k$ forment une base de $L^2(B_{r_c^{z_k}}(\bar{x}_k))$ [22]. Cela revient à faire l'

$$\text{Hypothèse de complétude :} \qquad \forall\, 1 \leq k \leq M,\ \forall\, x \in B_{r_c^{z_k}}, \quad \left(\widetilde{J}_k u\right)(x) = u(x). \tag{2.38}$$

Il n'est pas clair que cette hypothèse soit incontournable et nous ne la retiendrons donc pas pour l'instant.

Décomposition de la densité

L'équation (2.18) s'écrit

$$\psi_i = (I+T)\widetilde{\psi}_i = \widetilde{\psi}_i + \sum_{k=1}^{M} \left(J_k\widetilde{\psi}_i - \widetilde{J}_k\widetilde{\psi}_i\right). \tag{2.39}$$

Si l'on ne fait pas l'hypothèse (2.38), un quatrième terme dû à l'incomplétude de la base apparaît dans la décomposition de la densité ρ associée à une famille d'orbitales de Kohn-Sham ψ_i proposée par Blöchl [22]. On a en effet

$$\rho(x) = \sum_{i=1}^{N} |\psi_i(x)|^2 = \widetilde{\rho}(x) + \rho_1(x) - \widetilde{\rho}_1(x) + \rho_{inc}(x) \tag{2.40}$$

avec

$$\widetilde{\rho}(x) = \sum_{i=1}^{N} |\widetilde{\psi}_i(x)|^2$$

et

$$\rho_1(x) = \sum_{k=1}^{M} \rho_1^k(x), \qquad \widetilde{\rho}_1(x) = \sum_{k=1}^{M} \widetilde{\rho}_1^k(x), \qquad \rho_{inc}(x) = \sum_{k=1}^{M} \rho_{inc}^k(x),$$

avec

$$\rho_1^k(x) = \sum_{i=1}^{N} |(J_k\widetilde{\psi}_i)(x)|^2,$$

$$\widetilde{\rho}_1^k(x) = \sum_{i=1}^{N} |(\widetilde{J}_k\widetilde{\psi}_i)(x)|^2,$$

$$\rho_{inc}^k(x) = 2\sum_{i=1}^{N} \left[\widetilde{\psi}_i(x)\,(J_k\widetilde{\psi}_i)(x) - (J_k\widetilde{\psi}_i)(x)\,(\widetilde{J}_k\widetilde{\psi}_i)(x) - \widetilde{\psi}_i(x)\,(\widetilde{J}_k\widetilde{\psi}_i)(x) + |(\widetilde{J}_k\widetilde{\psi}_i)(x)|^2\right].$$

Par (2.37) la densité ρ_{inc}^k est localisée dans la région d'augmentation $B_{r_c^{z_k}}(\bar{x}_k)$. Si on utilise l'hypothèse de complétude (2.38), on retrouve $\rho_{inc}^k = 0$ dans $B_{r_c^{z_k}}(\bar{x}_k)$.

Soulignons que la décomposition (2.40) a été introduite pour mettre en évidence l'erreur introduite par l'hypothèse de complétude. Il est peu probable que ce soit la décomposition qui facilite le plus les calculs numériques.

Opérateurs monoélectroniques

Si A est un opérateur auto-adjoint sur $L^2(\mathbb{R}^3)$,

$$\langle \psi, A\psi \rangle = \langle \widetilde{\psi}, \widetilde{A}\widetilde{\psi} \rangle$$

où $\widetilde{A}$ désigne l'opérateur auto-adjoint

$$\widetilde{A} = A + \sum_{k=1}^{M} \left[T_k^T A + A T_k + T_k^T A T_k \right]. \tag{2.41}$$

Une autre expression de l'opérateur $\widetilde{A}$ est fournie par la formule

$$\begin{aligned} \widetilde{A} &= A + \sum_{k=1}^{M} \left[J_k^T A J_k - \widetilde{J}_k^T A \widetilde{J}_k \right] \\ &\quad + \sum_{k=1}^{M} \left[A J_k + J_k^T A - J_k^T A \widetilde{J}_k - \widetilde{J}_k^T A J_k - A\widetilde{J}_k - \widetilde{J}_k^T A + 2\widetilde{J}_k^T A \widetilde{J}_k \right]. \end{aligned}$$

Si A est un opérateur local multiplicatif (i.e. de type $(A\phi)(x) = v(x)\phi(x)$ où $v(x)$ est un potentiel donné), en utilisant (2.37) on vérifie que si on fait l'hypothèse de complétude (2.38), alors le dernier terme de l'expression ci-dessus est nul et on obtient

$$\begin{aligned} \widetilde{A} &= A + \sum_{k=1}^{M} \left[J_k^T A J_k - \widetilde{J}_k^T A \widetilde{J}_k \right] \\ &= A + \sum_{k=1}^{M} \left[\sum_{\mu,\nu=1}^{N_k} |\widetilde{p}_\mu^k\rangle \left(\langle \phi_\mu^k, A\phi_\nu^k \rangle - \langle \widetilde{\phi}_\mu^k, A\widetilde{\phi}_\nu^k \rangle \right) \langle \widetilde{p}_\nu^k| \right] ; \end{aligned}$$

on retrouve ainsi l'expression donnée dans [22]. Pour pouvoir effectuer la même manipulation avec un opérateur comme le laplacien, il faudrait s'assurer qu'aucune singularité n'apparaît aux interfaces $\partial B_{r_c^k}(\bar{x}_k)$.

Remarque 3. Si on considère une base de résolution pour (2.17) appartenant à l'espace vectoriel formé des fonctions $\widetilde{\psi}_i$, les termes liés à l'incomplétude de la base de résolution sont nuls. Dans le cas contraire, la négligence de ces termes est peut être la cause des oscillations obtenues sur l'énergie du système par exemple lorsqu'on augmente le cut-off de la base d'ondes planes.

Approximation des cœurs gelés et états fantômes

Lorsqu'on gèle des orbitales de cœur, la démarche décrite précédemment se généralise seulement sur les orbitales de valence. Aucune condition d'orthogonalité liée aux

orbitales de cœur n'est imposée lors du calcul sur le système moléculaire comme lors de l'utilisation des pseudo-potentiels traditionnels. Par conséquent, suite à l'incomplétude de la base des pseudo-orbitales dans les régions de cœur, on n'est pas assuré de respecter les contraintes d'orthogonalité des orbitales de valence calculées après reconstruction avec les orbitales de cœur gelées si on ne considère pas une base de résolution pour (2.17) appartenant à l'espace vectoriel formé des fonctions $\widetilde{\psi}_i$,

La méthode PAW correspond à l'introduction d'un potentiel séparable non local dans les équations de Kohn-Sham. Comme lors de l'utilisation des pseudo-potentiels traditionnels sous la forme de Kleinman et Bylander, on observe l'apparition d'états fantômes. Un choix adapté du potentiel v_{loc} permet de palier à ce problème, par exemple en augmentant la valeur de $\mathcal{V}_0$.

2.5.3 Quelques pistes de réflexion

En vue de l'étude théorique et numérique de la méthode PAW, une première étape consiste à donner un éclairage sur le calcul des opérateurs atomiques T_z et le passage aux systèmes moléculaires. D'un point de vue numérique, une étude de la sensibilité des résultats au choix des rayons r_c^z, de la fonction $k(r)$ et de $\mathcal{V}_0$ permettrait de comprendre le rôle de chacun de ces paramètres. Par ailleurs, la méthode PAW utilise des méthodes d'approximation du terme coulombien et de l'énergie d'échange-corrélation [71]. Une voie d'amélioration du calcul du terme coulombien consisteraient à utiliser des méthodes intégrales sur les sphères délimitant les régions d'augmentation.
D'un point de vue méthodologique, il serait utile d'étudier l'influence de la prise en compte des termes supplémentaires liés à l'incomplétude de la base. Enfin, la résolution de (2.17) étant indépendante du calcul des $\phi_{z,\nu}$, $\widetilde{\phi}_{z,\nu}$ et $\widetilde{p}_{z,\nu}$, on peut s'interroger sur un choix optimal de ces fonctions.

Chapitre 3

Une méthode de décomposition de domaine

Ce chapitre reproduit dans un premier temps un article soumis [A2] présentant une méthode de décomposition de domaine pour le calcul de structures électroniques. En second lieu, on donne une preuve de convergence de la méthode dans un cadre très simplifié.

Multilevel domain decomposition for electronic structure calculations

M. Barrault[1], E. Cancès[1], W. Hager [2] and C. Le Bris[1]

[1] *CERMICS, École Nationale des Ponts et Chaussées, 6 et 8 avenue Blaise Pascal, 77455 Marne-la-Vallée Cedex 2*
[2] *University of Florida, USA.*

We introduce a new domain decomposition method for electronic structure calculations within semi-empirical and Density Functional Theory (DFT) frameworks. This method iterates between a local fine solver and a global coarse solver, in the spirit of domain decomposition methods used in many other fields of the engineering sciences. Using this approach, calculations have been successfully performed on several linear polymer chains containing up to 40,000 atoms and 200,000 atomic orbitals. Both the computational cost and the memory requirement scale linearly with the number of atoms. Additional speed-up can easily be obtained by parallelization. We show that this domain decomposition method outperforms the Density Matrix Minimization (DMM) method for poor initial guesses. It is, in any case, an efficient preconditioner for DMM and other linear scaling methods, variational in nature, such as the Orbital Minimization (OM) procedure.

3.1 Introduction and motivation

A central issue in computational quantum chemistry is the determination of the electronic ground state of a molecular system. For completeness and self-consistency, we now briefly introduce the problem. In particular, we present it in a mathematical way.

3.1.1 Standard electronic structure calculations

A molecular system is composed of N electrons, modelled quantum mechanically, and a given number of nuclei, the latter being considered as classical point-like particles clamped at known positions (Born-Oppenheimer approximation). We refer to [1] for a general mathematical exposition and to [3,5] for the chemical background. Determining the electronic ground state amounts to solving a time-independent Schrödinger equation in $\mathbb{R}^{3N}$. This goal is out of reach for large values of N. In fact it is already infeasible for values of N exceeding three or four, unless dedicated techniques are employed. Examples are stochastic-like techniques such as Diffusion

Monte-Carlo approaches, or emerging techniques, such as sparse tensor products techniques [132]. Approximations of the Schrödinger equation have been developed, such as the widely used *tight-binding*, *Hartree-Fock* and *Kohn-Sham* models. For these three models, the numerical resolution of a problem of the following type is required : given H and S, respectivement an $N_b \times N_b$ symmetric matrix and an $N_b \times N_b$ symmetric positive definite matrix (with $N_b > N$), compute a solution $D_\star$ of the problem

$$\left\{ \begin{array}{ll} Hc_i = \epsilon_i S c_i, & \epsilon_1 \leq \ldots \leq \epsilon_N \leq \epsilon_{N+1} \leq \ldots \leq \epsilon_{N_b}, \\ c_i^t S c_j = \delta_{ij}, & \\ D_\star = \displaystyle\sum_{i=1}^{N} c_i c_i^t. & \end{array} \right. \tag{3.1}$$

Let us mention that most electronic structure calculations are performed with closed shell models [3], and that, consequently, the integer N in (3.1) then is the number of electron pairs. We remark that when S is the identity matrix, a solution $D_\star$ to (3.1) is a solution to the problem

$$\left\{ \begin{array}{l} \text{Find the } \textit{orthogonal projector} \text{ on the space spanned by the } N \text{ eigenvectors} \\ \text{associated with the lowest } N \text{ eigenvalues of } H. \end{array} \right. \tag{3.2}$$

In (3.2), and throughout this article, the eigenvalues are counted with their multiplicities. The N eigenvectors c_i, called *generalized* eigenvectors in order to emphasize the presence of the matrix S, represent the expansion in a given Galerkin basis $\{\chi_i\}_{1 \leq i \leq N_b}$ of the N one-electron wavefunctions. The matrix H is a mean-field Hamiltonian matrix. For instance, for the Kohn-Sham model, we have

$$H_{ij} = \frac{1}{2} \int_{\mathbf{R}^3} \nabla \chi_i \cdot \nabla \chi_j + \int_{\mathbf{R}^3} V \chi_i \chi_j \tag{3.3}$$

where V is a mean-field local potential. The matrix S is the overlap matrix associated with the basis $\{\chi_i\}_{1 \leq i \leq N_b}$:

$$S_{ij} = \int_{\mathbf{R}^3} \chi_i \chi_j. \tag{3.4}$$

In this article, we focus on the *Linear Combination of Atomic Orbitals* (LCAO) approach. This is a very efficient discretization technique, using localized basis functions $\{\chi_i\}$, compactly supported [97] or exhibiting a gaussian fall-off [3].

It is important to emphasize what makes the electronic structure problem, discretized with the LCAO approach, specific as compared to other linear eigenvalue problems encountered in other fields of the engineering sciences (see [124, 130] for instance). First, N_b is proportional to N, and not much larger than it (say $N_b \sim 2N$

to fix the ideas). Hence, the problem is not finding a few eigenvectors of the generalized eigenvalue problem (3.1). Second, although the matrices H and S are sparse for large molecular systems (see section 3.1.2 for details), they are not as sparse as the stiffness and mass matrices usually encountered when using finite difference or finite element methods. For example, the bandwith of H and S is of the order of 10^2 in the numerical examples reported in section 3.4. Note that, in contrast, for plane wave basis set discretizations (which will not be discussed here), the parameter N_b is much larger than N (say $N_b \sim 100\,N$), the matrix S is the identity matrix and the matrix H est full. Third, and this is a crucial point, the output of the calculation is the matrix $D_\star$ and not the generalized eigenvectors c_i themselves. This is the fundamental remark allowing the construction of linear scaling methods (see section 3.1.2).

A solution $D_\star$ of (3.1) is

$$D_\star = C_\star C_\star^t \tag{3.5}$$

where $C_\star$ is a solution to the minimization problem

$$\inf\left\{ \mathrm{Tr}\Big(HCC^t\Big), \quad C \in \mathcal{M}^{N_b,N}(\mathbb{R}),\ C^tSC = I_N \right\}. \tag{3.6}$$

Note that the energy functional $\mathrm{Tr}\Big(HCC^t\Big)$ can be given the more symmetric form $\mathrm{Tr}\Big(C^tHC\Big)$. Here and below, $\mathcal{M}^{k,l}$ denotes the vector space of the $k\times l$ real matrices. Notice that (3.6) has many minimizers : if $C_\star$ is a minimizer, so is $C_\star U$ for any orthogonal $N \times N$ matrix U. However, under the standard assumption that the N-th eigenvalue of H is strictly lower than the $(N+1)$-th one, the matrix D^* defined by (3.5) does not in fact depend on the choice of the minimizer $C_\star$ of (3.6). Notice also that (3.1) are not the Euler-Lagrange equations of (3.6) but that any critical point of (3.6) is obtained from a solution of (3.1) by an orthogonal transformation of the columns of $C_\star = (c_1|\cdots|c_N)$.

The standard approach to compute $D_\star$ is to solve the generalized eigenvalue problem (3.1) and then construct $C_\star$ thus $D_\star$ by collecting the lowest N generalized eigenvectors of H. This approach is employed when the number N of electrons (or electron pairs) is not too large, say smaller than 10^3.

3.1.2 Linear scaling methods

One of the current challenges of Computational Chemistry is to lower the computational complexity N^3 of this solution procedure. A linear complexity N is the holy grail. There are various existing methods designed for this purpose. Surveys on such methods are [24, 58]. Our purpose here is to introduce a new method, based on the *domain decomposition* paradigm. We remark that the method introduced here is not the first occurrence of a method based on a decomposition of the matrix H [104], but a significant methodological improvement is fulfilled with the present

method. To the best of our knowledge, such methods only consist of local solvers complemented by a crude global step. The method introduced below seems to be the first one really exhibiting the local/global paradigm in the spirit of methods used in other fields of the engineering sciences. Numerical obervations confirm the major practical interest methodological improvement.

Why is a *linear scaling* plausible for computing $D_\star$? To justify the fact that the cubic scaling is an estimate by excess of the computational task required to solve (3.1), we argue that the matrix does not need to be diagonalized. As mentioned above, only the *orthogonal projector* on the subspace generated by the lowest N eigenvectors is to be determined and *not* the *explicit* values of these lowest N eigenvectors. But in order to reach a linear complexity, appropriate assumptions are necessary, both on the form of the matrices H and S, and on the matrix $D_\star$ solution to (3.1) :

- (H1). The matrices H and S are assumed sparse, in the sense that, for large systems, the number of non-zero coefficients scales as N. This assumption is not restrictive. In particular, it follows from (3.3) and (3.4) that it is automatically satisfied for Kohn-Sham models as soon as the basis functions are localized in real space, which is in particular the case for the widely used atomic orbital basis sets [1] ;
- (H2). A second assumption is that the matrix $D_\star$ built from the solution to (3.1) is also sparse. This condition seems to be fulfilled as soon as the relative gap

$$\gamma = \frac{\epsilon_{N+1} - \epsilon_N}{\epsilon_{N_b} - \epsilon_1}. \tag{3.7}$$

  deduced from the solution of (3.1) is large enough. As explained in section 3.2 below, this observation can be supported by qualitative physical arguments. On the other hand, we are not aware of any mathematical argument of linear algebra that would justify assumption (H2) in a general setting.

We assume (H1)-(H2) in the following. Current efforts aim at treating cases when the second assumption is not fulfilled, which in particular corresponds to the case of conducting materials. The problem (3.2) is then extremely difficult because the gap γ in (3.7) being very small, the matrix D is likely to be dense. Reaching linear complexity is then a challenging issue, unsatisfactorily solved to date. State of the art linear scaling methods presented in the literature experience tremendous difficulties (to say the least) in such cases. It is therefore reasonable to improve in a first step the existing methods in the setting of assumption (H2), before turning to more challenging issues.

Before we get to the heart of the matter, we would like to point out the following feature of the problem under consideration.

In practice, Problem (3.1) has to be solved *repeatedly*. For instance, it is the inner loop in a nonlinear minimization problem where H depends self-consistently on $D_\star$. We refer to [27, 127] for efficient algorithms to iterate on this nonlinearity and to [1]

for a review on the subject. Alternatively, or in addition to the above, problem (3.1) is parametrized by the positions of the nuclei (both the mean-field operator H and the overlap matrix S indeed depend on these positions), and these positions may vary. This is the case in molecular mechanics (find the optimal configuration of nuclei that gives the lowest possible energy to the molecular system), and in molecular dynamics as well (the positions of nuclei follow the Newton law of motion in the mean-field created by the electrons). In either case, problem (3.1) is not be solved *from scratch*. Because of previous calculations, we may consider we have at our disposal a good initial guess for the solution. The latter comes from e.g. previous positions of nuclei, or previous iterations in the outer loop of determination of H. In difficult cases it may even come from a previous computation with a coarse grained model. In other words, the question addressed reads *solving Problem (3.1) for some $H + \delta H$ and $S + \delta S$ that are small perturbations of previous H and S for which the solution is known*. This specific context allows for a speed up of the algorithm when the initial guess is sufficiently good. This is the reason why, in the following, we shall frequently make distinctions between bad and good initial guesses.

3.2 Localization in Quantum Chemistry

The physical system we consider is a long linear molecule (for instance a one-dimensional polymer or a nanotube). Let us emphasize that we do not claim a particular physical relevance of this system. This is for the purpose of illustration. We believe the system considered to be a good representative of a broad class of large molecular systems that may be encountered practically. Each atomic orbital χ_i is centered on one nucleus. Either it is supported in a ball of small radius [87] (in comparison to the size of the macromolecule under study), or it has a rapid exponential-like or Gaussian-like [52] fall-off. The atomic orbitals are numbered following the orientation of the molecule. Then, the mean-field Hamiltonian matrix H whose entries are defined by (3.3) has the band structure shown in Figure 3.1.

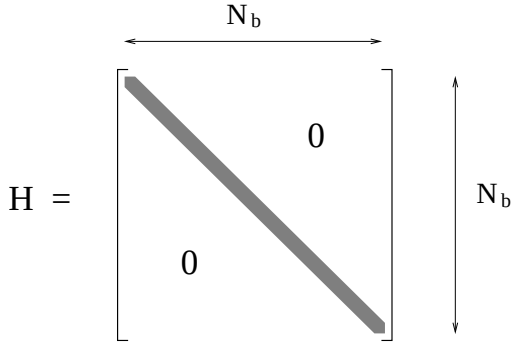


FIG. 3.1 – Band structure of the symmetric matrix H.

Although the eigenvectors of H are *a priori* delocalized (most of their coefficients

do not vanish), it seems to be possible to build a S-orthonormal basis of the subspace generated by the lowest N eigenvectors of H, consisting of *localized* vectors (only a few consecutive coefficients are non zero). This is motivated by a physical argument of locality of the interactions [78]. For periodic systems, the localized vectors correspond to the so-called Wannier orbitals [14]. It can be proven that in this case, the larger the band gap, the better the localization of the Wannier orbitals [76]. For insulators, the Wannier orbitals indeed enjoy an exponential fall-off rate proportional to the band gap. For conductors, the fall-off is only algebraic. As mentioned in the introduction, we only consider here the former case. This allows us to assume that there exists some integer $q \ll N_b$, such that N_b/q is an integer, for which all of these localized functions can be essentially expanded on q consecutive atomic orbitals. Denoting by $n = 2q$, we can therefore assume a good approximation of a solution $C_\star$ to (3.6) exists, with the block structure displayed on Figure 3.4. Note that each block C_i only overlaps with its nearest neighbors. Correspondingly, we introduce the block structure of H displayed on Figure 3.2. The matrix D constructed from a block matrix C using (3.5) has the structure represented in Figure 3.3 and satisfies the constraints $D = D^t$, $D^2 = D$, $\mathrm{Tr}(D) = N$.

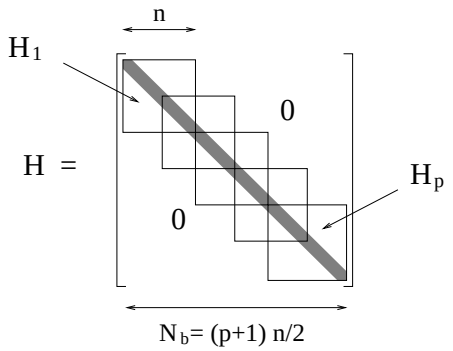


FIG. 3.2 – Block structure of the matrix H.

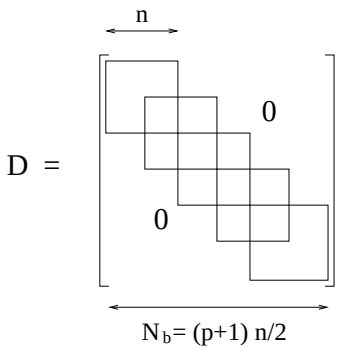


FIG. 3.3 – Block structure of the matrix D.

Let us point out that the integers q and $n = 2q$ depend on the band gap, *not* on

the size of the molecule. The condition $n = 2q$ is only valid for $S = I_{N_b}$. For $S \neq I_{N_b}$, it is replaced by $n = 2q + nbs$ where $2\,nbs - 1$ is the bandwidth of the matrix S.

The domain decomposition algorithm we propose aims at searching an approximate solution to (3.6) that has the block structure described above.

For simplicity, we now present our method assuming that $S = I_{N_b}$, i.e. that the Galerkin basis $\{\chi_i\}_{1 \leq i \leq N_b}$ is orthonormal. The extension of the method to the case when $S \neq I_{N_b}$ is straightforward. Problem (3.6) then reads

$$\inf\left\{ \mathrm{Tr}\Big(HCC^t\Big), \quad C \in \mathcal{M}^{N_b,N}(\mathbb{R}),\ C^tC = I_N \right\}. \tag{3.8}$$

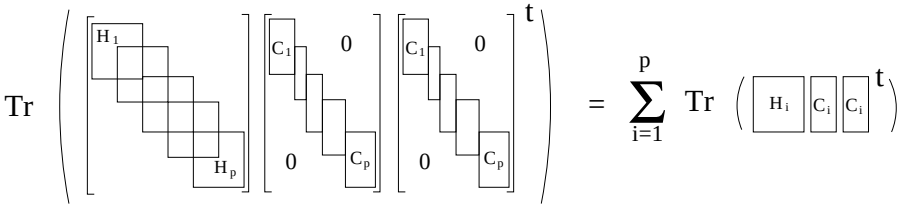
Our approach consists in solving an approximation of problem (3.8) obtained by minimizing the exact energy $\mathrm{Tr}\Big(HCC^t\Big)$ on the set of the matrices C which have the block structure displayed on Figure 3.4 and satisfy the constraint $C^tC = I_N$. The resulting minimization problem can be recast as

$$\inf\left\{ \sum_{i=1}^{p} \mathrm{Tr}\left(H_i C_i C_i^t\right), \qquad C_i \in \mathcal{M}^{n,m_i}(\mathbb{R}), \quad m_i \in \mathbb{N}, \quad C_i^t C_i = I_{m_i} \quad \forall\ 1 \leq i \leq p, \right.$$
$$\left. C_i^t T C_{i+1} = 0 \quad \forall\ 1 \leq i \leq p-1, \quad \sum_{i=1}^{p} m_i = N \right\}. \tag{3.9}$$

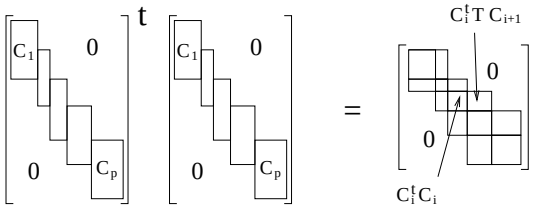
In the above formula, $T \in \mathcal{M}^{n,n}(\mathbb{R})$ is the matrix defined by

$$T_{kl} = \begin{cases} 1 & \text{if } k - l = q \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

and $H_i \in \mathcal{M}^{n,n}(\mathbb{R})$ is a symmetric submatrix of H (see Figure 3.2). Indeed,



and



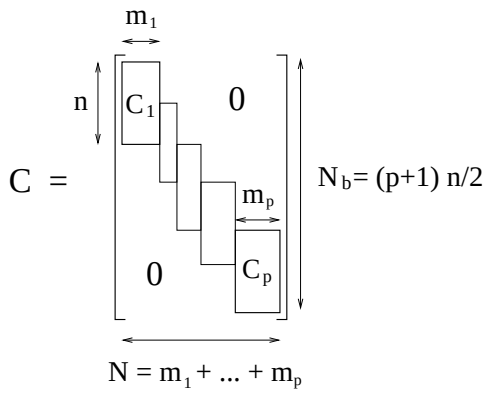


FIG. 3.4 – Block structure of the matrices C. Note that by construction each block only overlaps with its nearest neighbors.

In this way, we replace the $\frac{N(N+1)}{2}$ *global* scalar constraints $C^tC = I_N$ involving vectors of size N_b, by the $\sum_{i=1}^{p} \frac{m_i(m_i+1)}{2}$ *local* scalar constraints $C_i^tC_i = I_{m_i}$ and the $\sum_{i=1}^{p-1} m_i m_{i+1}$ *local* scalar constraints $C_i^tTC_{i+1} = 0$, involving vectors of size n. We would like to emphasize that we can obtain in this way a basis of the vector space generated by the lowest N eigenvectors of H, but not the eigenvectors themselves. This method is therefore not directly applicable to standard diagonalization problems.

Our algorithm searches for the solution to (3.9), not to (3.8). More rigorously stated, we search for the solution to the Euler-Lagrange equations of (3.9) :

$$\left\{ \begin{array}{rcll} H_iC_i & = & C_iE_i + T^tC_{i-1}\Lambda_{i-1,i} + TC_{i+1}\Lambda_{i,i+1}^t & 1 \leq i \leq p, \\ C_i^tC_i & = & I_{m_i} & 1 \leq i \leq p, \\ C_i^tTC_{i+1} & = & 0 & 1 \leq i \leq p-1, \end{array} \right. \tag{3.11}$$

where by convention

$$C_0 = C_{p+1} = 0. \tag{3.12}$$

The matrices $(E_i)_{1\leq i\leq p}$ and $(\Lambda_{i,i+1})_{1\leq i\leq p-1}$ respectively denote the matrices of Lagrange multipliers associated with the orthonormality constraints $C_i^tC_i = I_{m_i}$ and $C_i^tTC_{i+1} = 0$. The $m_i \times m_i$ matrix E_i is symmetric. The matrix $\Lambda_{i,i+1}$ is of size $m_i \times m_{i+1}$. The above equations can be easily derived by considering the Lagrangian

$$\begin{aligned} \mathcal{L}\Big(\{C_i\}, \{E_i\}, \{\Lambda_{i,i+1}\} \Big) &= \sum_{i=1}^{p} \mathrm{Tr}\Big(H_iC_iC_i^t \Big) + \sum_{i=1}^{p} \mathrm{Tr}\Big(\left(C_i^tC_i - I_{m_i}\right) E_i \Big) \\ &\quad + \sum_{i=1}^{p-1} \mathrm{Tr}\Big(C_i^tTC_{i+1}\Lambda_{i,i+1}^t \Big). \end{aligned}$$

The block structure imposed on the matrices clearly lowers the dimension of the search space we have to explore. However, this simplification comes at a price.

First, problem (3.9) only *approximates* problem (3.8). Second, (3.9) may have local, non global, minimizers, whereas all the local minimizers of (3.8) are global. There are thus *a priori* many spurious solutions of the Euler Lagrange equations (3.11) associated with (3.9).

A point is that the sizes $(m_i)_{1\leq i\leq p}$ are not *a priori* prescribed. In our approach, they are ajusted during the iterations. We shall see how in the sequel.

3.3 Description of the domain decomposition algorithm

3.3.1 Description of a simplified form

For pedagogic purpose, we first consider the following problem

$$\inf\left\{\langle H_1Z_1,Z_1\rangle+\langle H_2Z_2,Z_2\rangle,\quad Z_i\in\mathbb{R}^{N_b},\ \langle Z_i,Z_i\rangle=1,\ \langle Z_1,Z_2\rangle=0\right\}. \tag{3.13}$$

Problem (3.13) is a particular occurence of (3.9). We have denoted by $\langle\cdot,\cdot\rangle$ the standard Euclidean scalar product on $\mathbb{R}^{N_b}$.

For (3.13), the algorithm is defined in the following simplified form. Choose (Z_1^0,Z_2^0) satisfying the constraints and construct the sequence $(Z_1^k,Z_2^k)_{k\in\mathbb{N}}$ by the following iteration procedure. Assume (Z_1^k,Z_2^k) is known, then

– Local step. Solve

$$\left\{\begin{array}{l}\widetilde{Z}_1^k=\mathrm{arginf}\left\{\langle H_1Z_1,Z_1\rangle,\ Z_1\in\mathbb{R}^{N_b},\ \langle Z_1,Z_1\rangle=1\ \langle Z_1,Z_2^k\rangle=0\right\},\\ \widetilde{Z}_2^k=\mathrm{arginf}\left\{\langle H_2Z_2,Z_2\rangle,\ Z_2\in\mathbb{R}^{N_b},\ \langle Z_2,Z_2\rangle=1\ \langle\widetilde{Z}_1^k,Z_2\rangle=0\right\};\end{array}\right. \tag{3.14}$$

– Global step. Solve

$$\alpha^*=\mathrm{arginf}\left\{\langle H_1Z_1,Z_1\rangle+\langle H_2Z_2,Z_2\rangle,\ \alpha\in\mathbb{R}\right\} \tag{3.15}$$

where

$$Z_1=\frac{\widetilde{Z}_1^k+\alpha\widetilde{Z}_2^k}{\sqrt{1+\alpha^2}},\ Z_2=\frac{-\alpha\widetilde{Z}_1^k+\widetilde{Z}_2^k}{\sqrt{1+\alpha^2}}, \tag{3.16}$$

and set

$$Z_1^{k+1}=\frac{\widetilde{Z}_1^k+\alpha^*\widetilde{Z}_2^k}{\sqrt{1+(\alpha^*)^2}},\quad Z_2^{k+1}=\frac{-\alpha^*\widetilde{Z}_1^k+\widetilde{Z}_2^k}{\sqrt{1+(\alpha^*)^2}}. \tag{3.17}$$

In the k-th iteration of the local step, we first fix $Z_2=Z_2^k$ and optimize over Z_1 to obtain $\tilde{Z}_1^k$. Then we fix $Z_1=\tilde{Z}_1^k$ and optimize over Z_2 to obtain $\tilde{Z}_2^k$. This local step monotonically reduces the objective function, however, it may not converge to the global optimum. The technical problem is that the Lagrange multipliers associated with the constraint $\langle Z_1,Z_2\rangle=0$ may converge to different values in the two subproblems associated with the local step. In the global step, we optimize the

sum $\langle H_1 Z_1, Z_1\rangle + \langle H_2 Z_2, Z_2\rangle$ over the subspace spanned by $\tilde{Z}_1^k$ and $\tilde{Z}_2^k$, subject to the constraints in (3.13). The global step again reduces the value of the objective function since $\tilde{Z}_1^k$ and $\tilde{Z}_2^k$ are feasible in the global step. It can be shown that the combined algorithm (local step + global step) monotonically decreases the objective function and globally converges to an optimal solution of (3.13).

This algorithm operates at two levels : a fine level where we solve two problems of dimension N_b rather than one problem of dimension $2N_b$; a coarse level where we solve a problem of dimension 2. Left by itself, the fine step converges to a suboptimal solution of (3.13). Combining the fine step with the global step yields convergence to a global optimum.

In addition to providing a pedagogic view on the general algorithm presented in the following section, the simplified form (3.14)-(3.17) has a theoretical interest. In contrast to the general algorithm for which we cannot provide a convergence analysis, the simplified form (3.14)-(3.17) may be analyzed mathematically, at least in the particular situation when $H_1 = H_2 = H$. Then solving (3.13) amounts to searching for the lowest two eigenelements of the matrix H. Notice that the global step (3.15)-(3.17) is then unnecessary because the functional to minimize in (3.15) does not depend on α.

However, we can show that the iterations (3.14) converge in the following sense. The 2-dimensional vector space spanned by the lowest two eigenvalues of H is reached asymptotically. This occurs under an appropriate condition on the matrix H. The latter is a condition of separation of the eigenvalues, namely $\epsilon_2 - \epsilon_1 < \epsilon_3 - \epsilon_2$ with obvious notation. The gap $\epsilon_3 - \epsilon_2$ gives the speed of convergence. For brevity, we do not detail the proof here (see the section 3.6). Future work on the numerical analysis of more general cases is in progress.

3.3.2 Description of the algorithm

We define, for all p-tuple $(C_i)_{1\le i\le p}$,

$$\mathcal{E}\Big((C_i)_{1\le i\le p}\Big) = \sum_{i=1}^{p} \mathrm{Tr}\Big(H_i C_i C_i^t\Big), \tag{3.18}$$

and set by convention

$$U_0 = U_p = 0. \tag{3.19}$$

We introduce an integer ϵ, initialized to one, that will alternate between the values zero and one during the iterations.

At iteration k, we have at hand a set of block sizes $(m_i^k)_{1\le i\le p}$ and a set of matrices $(C_i^k)_{1\le i\le p}$ such that $C_i^k \in \mathcal{M}^{n,m_i^k}(\mathbb{R})$, $[C_i^k]^t C_i^k = I_{m_i^k}$, $[C_i^k]^t T C_{i+1}^k = 0$. We now explain how to compute the new iterate $(m_i^{k+1})_{1\le i\le p}$, $(C_i^{k+1})_{1\le i\le p}$.

Multilevel Domain Decomposition (MDD) algorithm

- **Step 1 : Local fine solver**.

(a) For each i, diagonalize the matrix $H_{2i+\epsilon}$ in the subspace

$$V^k_{2i+\epsilon} = \Big\{ x \in \mathbb{R}^n, \quad \big[C^k_{2i+\epsilon-1}\big]^t T x = 0, \quad x^t T C^k_{2i+\epsilon+1} = 0 \Big\},$$

i.e. diagonalize $P^k_{2i+\epsilon} H_{2i+\epsilon} P^k_{2i+\epsilon}$ where $P^k_{2i+\epsilon}$ is the orthogonal projector on $V^k_{2i+\epsilon}$. This provides (at least) $n - m^k_{2i+\epsilon-1} - m^k_{2i+\epsilon+1}$ real eigenvalues $\lambda^k_{2i+\epsilon,1} \le \lambda^k_{2i+\epsilon,2} \le \cdots$ and associated orthonormal vectors $x^k_{2i+\epsilon,j}$. The latter are T-orthogonal to the column vectors of C^k_{i-1} and C^k_{i+1}.

(b) Sort the eigenvalues $(\lambda^k_{2i+\epsilon,j})_{i,j}$ in increasing order, and select the lowest $\sum_i m_{2i+\epsilon}$ of them. For each i, collect in block $\#2i+\epsilon$ the eigenvalues $\lambda^k_{2i+\epsilon,j}$ selected. New intermediate block sizes $\bar{m}^k_{2i+\epsilon}$ are defined.

(c) For each i, collect the lowest $\bar{m}^k_{2i+\epsilon}$ vectors $x^k_{2i+\epsilon,j}$ in the $n \times \bar{m}^k_{2i+\epsilon}$ matrix $\overline{C}^k_{2i+\epsilon}$.

(d) For each i, diagonalize the matrix $H_{2i+\epsilon+1}$ in the subspace

$$V^k_{2i+\epsilon+1} = \left\{ x \in \mathbb{R}^n, \quad \left[\overline{C}^k_{2i+\epsilon}\right]^t T x = 0, \quad x^t T \overline{C}^k_{2i+\epsilon+2} = 0 \right\}$$

in order to get eigenvalues $\lambda^k_{2i+\epsilon+1,1} \le \lambda^k_{2i+\epsilon+1,2} \le \cdots$ and associated orthonormal vectors $x^k_{2i+\epsilon+1,j}$. The latter are T-orthogonal to the column vectors of $\overline{C}^k_{2i+\epsilon}$ and $\overline{C}^k_{2i+\epsilon+2}$.

(e) Sort all the eigenvalues $\big\{(\lambda^k_{2i+\epsilon+1,j})_{i,j}, (\lambda^k_{2i+\epsilon,j})_{i,j}\big\}$ in increasing order. Select the lowest N. For each l, collect in block $\#l$ the eigenvalues $\lambda^k_{l,j}$ selected. New intermediate block sizes $(m^{k+1}_l)_{1\le l\le p}$ are thus defined.

(f) Set $\widetilde{C}^k_l = \left[x^k_{l,1} | \cdots | x^k_{l,m^{k+1}_l} \right]$.

(g) Replace ϵ by $1-\epsilon$ and proceed to step 2 below.

• **Step 2 : global coarse solver**. Solve

$$\mathcal{U}^* = \mathrm{arginf}\Big\{ f(\mathcal{U}),\ \mathcal{U} = (U_i)_i,\ \forall 1 \le i \le p-1 \ \ U_i \in \mathcal{M}^{m_{i+1},m_i}(\mathbb{R}) \Big\}, \quad (3.20)$$

where

$$f(\mathcal{U}) = \mathcal{E}\left(\Big(C_i(\mathcal{U}) \big(C_i(\mathcal{U})^t C_i(\mathcal{U}) \big)^{-\frac{1}{2}} \Big)_i \right), \quad (3.21)$$

and

$$C_i(\mathcal{U}) = \widetilde{C}^k_i + T\widetilde{C}^k_{i+1} U_i \Big([\widetilde{C}^k_i]^t T T^t \widetilde{C}^k_i \Big) - T^t \widetilde{C}^k_{i-1} U^t_{i-1} \Big([\widetilde{C}^k_i]^t T^t T \widetilde{C}^k_i \Big). \quad (3.22)$$

Next set, for all $1 \le i \le p$,

$$C^{k+1}_i = C_i(\mathcal{U}^*) \ \Big(C_i(\mathcal{U}^*)^t \, C_i(\mathcal{U}^*) \Big)^{-1/2} . \quad (3.23)$$

Note that $\left[C^{k+1}_i\right]^t T C^{k+1}_{i+1} = 0$ (this follows from $T^2 = 0$).

We think of the even indexed unknowns C_{2i} as the black variables and the odd indexed unknowns C_{2i+1} as the white variables. In the first phase of the local fine solver, we optimize over the white variables while holding the black variables fixed. In the second phase of the local fine solver, we optimize over the black variables while holding the white variables fixed. In the global step, we perturb each variable by a linear combination of the adjacent variables. The matrices $\mathcal{U} = (U_i)_i$ in (3.20) play the same role as the real parameter α in (3.15). The perturbation is designed so that the constraints are satisfied. The optimization is performed over the matrices generating the linear combinations. In the next iteration, we interchange the order of the optimizations : first optimize over the black variables while holding the white variables fixed, then optimize over the white variables while holding the black variables fixed.

Let us point out that an accurate solution to (3.20) is not needed. In practice, we reduce the computational cost of the global step, by using again a domain decomposition method. The blocks $(C_i)_{1\leq i\leq p}$ are collected in r overlapping groups $(G_l)_{1\leq l\leq r}$ as shown in Figure 3.5. Problem (3.20) is solved first for the blocks (G_{2l+1}), next for the blocks (G_{2l}). Possibly, this procedure is repeated a few times. The advantage of this strategy is that the computational time of the global step scales linearly with N. In addition, it is parallel in nature. The solution of (3.20) for a given group is performed by a few steps of a Newton-type algorithm. Other preconditioned iterative methods could also be considered.

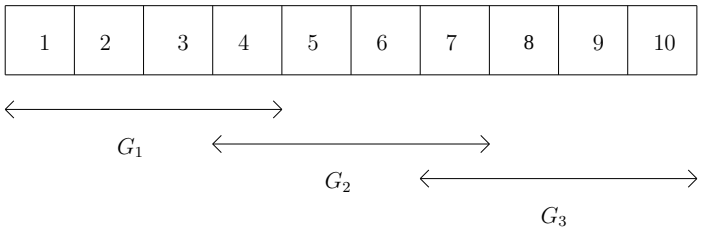


FIG. 3.5 – Collection of $p = 10$ blocks into $r = 3$ groups.

3.3.3 Comments on the local step

The local step is based on a checkerboard iteration technique.

When $\epsilon = 1$, steps 1a-1c search for a solution $(\bar{m}^k_{2i+1}, \overline{C}^k_{2i+1})_i$ to the problem

$$\inf\Bigg\{ \sum_i \mathrm{Tr}\left(H_{2i+1} C_{2i+1} C^t_{2i+1}\right), \quad C_{2i+1} \in \mathcal{M}^{n,m_{2i+1}}(\mathbb{R}), \quad C^t_{2i+1} C_{2i+1} = I_{m_{2i+1}},$$
$$[C^k_{2i}]^t T C_{2i+1} = 0, \quad C^t_{2i+1} T C^k_{2i+2} = 0,$$
$$m_{2i+1} \in \mathbb{N}, \quad \sum_i m_{2i+1} = \sum_i m^k_{2i+1} \Bigg\}.$$

During steps 1a-1c, the "white" blocks C_{2i}^k are kept fixed. The "black" blocks C_{2i+1}^k are optimized under the orthogonality constraints imposed by the "white" blocks. A point is that most of the computational effort can be done *in parallel*. Indeed, for p even, say, performing step 1a amounts to solving $p/2$ *independent* diagonalisation problems of size n.

Likewise, steps 1d-1f solve

$$\inf\Big\{ \sum_{i=1}^{p} \mathrm{Tr}\left(H_i C_i C_i^t\right), \qquad C_i \in \mathcal{M}^{n,m_i}(\mathbb{R}), \quad C_i^t C_i = I_{m_i}, \quad m_i \in \mathbb{N}, \quad \sum_i m_i = N$$
$$[\overline{C}_{2j-1}^k]^t T C_{2j} = 0, \quad [C_{2j}]^t T [\overline{C}_{2j+1}]^k = 0,$$
$$0 \le m_{2j+1} \le \bar{m}_{2j+1}^k, \quad C_{2j+1} \subset \overline{C}_{2j+1}^k \Big\},$$

where the notation $C_{2j+1} \subset \overline{C}_{2j+1}^k$ means that each column of C_{2j+1} is a column of $\overline{C}_{2j+1}^k$. Here again, most of the computational effort can be performed in parallel.

When $\epsilon = 1$, "black" vectors (i.e. vectors belonging to blocks with odd indices) are allowed to become "white" vectors, but the reverse is forbidden. In order to symmetrize the process, ϵ is replaced by $1 - \epsilon$ in the next iteration.

We wish to emphasize that, although called *local*, this step already accounts for some global concern. Indeed, and it is a key point of the local step, substeps (b) and (e) sort the *complete* set of eigenvalues generated locally. This, together with the update of the size m_i of the blocks, allows for a preliminary propagation of the information throughout the whole system. The global step will complement this.

Finally, let us mention that in the local steps, (approximate) T-orthogonality is obtained by a Householder orthonormalization process. The required orthonormality criterion is

$$\forall\, 1 \le i \le p-1, \quad \left\| [\tilde{C}_i^k]^t T \tilde{C}_{i+1}^k \right\| \le \epsilon_L, \tag{3.24}$$

where $\epsilon_L > 0$ is a threshold to be chosen by the user.

3.3.4 Comments on the global step

Let us briefly illustrate the role played by the global step. For simplicity, we consider the case of two blocks of same initial size $m_1 = m_2 = m$ and we assume that m_1 and m_2 do not vary during the iterations. If only the local step is performed, then the new iterate

$$(C_1^{k+1}, C_2^{k+1}) = (\widetilde{C}_1^k, \widetilde{C}_2^k)$$

does not necessarily satisfies (3.11). Indeed, there is no reason why the Lagrange multipliers corresponding to the two contraints $C^t T C_2^k = 0$ (step 1a when $\epsilon = 1$) on the one hand and $[\widetilde{C}_1^k]^t T C = 0$ (step 1d when $\epsilon = 1$) on the other hand should be the

same. The global step *asymptotically* enforces the equality of Lagrange multipliers. This is a way to account for a global feature of the problem.

Let us emphasize this specific point. Assume $U^* = 0$ in the global step of the k-th iteration of the algorithm, or in other words that the global step is not effective at the k-th iteration. Then it implies that the output $(\widetilde{C}_1, \widetilde{C}_2) = (\widetilde{C}_1^k, \widetilde{C}_2^k)$ of the local step already satisfies (3.11). Indeed,

$$f(U) = \mathrm{Tr}\Big(J_1(U)C_1(U)^t H_1 C_1(U)\Big) + \mathrm{Tr}\Big(J_2(U)C_2(U)^t H_2 C_2(U)\Big) \tag{3.25}$$

with $J_i(U) = \Big(C_i(U)^t C_i(U)\Big)^{-1}$ for $i = 1, 2$. Since

$$\begin{aligned}
\Big(J_1(U)\Big)^{-1} &= I_m + \Big(\widetilde{C}_1^t T T^t \widetilde{C}_1\Big) U^t \Big(\widetilde{C}_2^t T^t T \widetilde{C}_2\Big) U \Big(\widetilde{C}_1^t T T^t \widetilde{C}_1\Big), \\
\Big(J_2(U)\Big)^{-1} &= I_m + \Big(\widetilde{C}_2^t T^t T \widetilde{C}_2\Big) U \Big(\widetilde{C}_1^t T T^t \widetilde{C}_1\Big) U^t \Big(\widetilde{C}_2^t T^t T \widetilde{C}_2\Big),
\end{aligned}$$

we have $\nabla J_1(0) = \nabla J_2(0) = 0$. The matrix U being a square matrix of dimension m, for all $1 \le i, j \le m$,

$$\begin{aligned}
\frac{1}{2}\frac{\partial f}{\partial U_{ij}}(0) &= \mathrm{Tr}\Big(\Big[\frac{\partial C_1}{\partial U_{ij}}(0)\Big]^t H_1 \widetilde{C}_1 \Big) + \mathrm{Tr}\Big(\Big[\frac{\partial C_2}{\partial U_{ij}}(0)\Big]^t H_2 \widetilde{C}_2 \Big) \\
&= \Big(\Big(\widetilde{C}_1^t T T^t \widetilde{C}_1\Big) \widetilde{C}_1^t H_1 T \widetilde{C}_2 \Big)_{ji} - \Big(\widetilde{C}_1^t T H_2 \widetilde{C}_2 \Big(\widetilde{C}_2^t T^t T \widetilde{C}_2\Big) \Big)_{ji} \\
&= \Big(\Big(\widetilde{C}_1^t T T^t \widetilde{C}_1\Big) (\Lambda_1 - \Lambda_2) \Big(\widetilde{C}_2^t T^t T \widetilde{C}_2\Big) \Big)_{ji},
\end{aligned} \tag{3.26}$$

where Λ_1 and Λ_2 are defined by

$$\begin{cases} H_1 \widetilde{C}_1 = \widetilde{C}_1 E_1 + T \widetilde{C}_2 \Lambda_1^t, \\ H_2 \widetilde{C}_2 = \widetilde{C}_2 E_2 + T^t \widetilde{C}_1 \Lambda_2. \end{cases} \tag{3.27}$$

As $U^* = 0$ implies

$$\forall\ 1 \le i, j \le m \quad \frac{\partial f}{\partial U_{ij}}(0) = 0, \tag{3.28}$$

we conclude that $\Lambda_1 = \Lambda_2$ if the matrices $\Big(\widetilde{C}_1^t T T^t \widetilde{C}_1\Big)$ and $\Big(\widetilde{C}_2^t T^t T \widetilde{C}_2\Big)$ are invertible, which is generally the case when $n \gg 2m$. Consequently, (3.11) is satisfied by $(\widetilde{C}_1, \widetilde{C}_2)$.

On the other hand, when n is not much larger that $2m$, the above matrices are not invertible and (3.11) is usually not satisfied. In this case, the global step is slightly modified in order to recover (3.11) and thus improve the efficiency of the global step. We replace (3.22) by

$$\forall\ 1 \le i \le p, \quad C_i(\mathcal{U}) = \tilde{C}_i^k + T \widehat{C}_{i+1}^k U_i \Big([\widehat{C}_i^k]^t T T^t \widehat{C}_i^k \Big) - T^t \widehat{C}_{i-1}^k U_{i-1}^t \Big([\widehat{C}_i^k]^t T^t T \widehat{C}_i^k \Big) \tag{3.29}$$

where $\widehat{C}_i^k$ is a block formed by vectors collected in the vector space defined by $\widetilde{C}_i^k$. These vectors are selected using a modified Gram-Schmidt orthonormalization process. The size of the blocks $\widehat{C}_i^k$ is appropriately chosen. The larger the blocks $\widehat{C}_i^k$, the more precise the global step but the worse the conditioning of the optimization problem. In addition, since the global step is the most demanding step of the algorithm, considerations both on the computational time and in terms of memory are accounted for when fixing the sizes of the blocks $\widehat{C}_i^k$.

Our numerical experiments show that when the global step is performed (using (3.22) or (3.29), depending on n and m), the blocks $(C_i^{k+1})_i$ do not exactly satisfy the orthonormality constraint, owing to evident round-off errors. All the linear scaling algorithms have difficulties in ensuring this constraint and our MDD approach is no exception. The tests performed however show that the constraint remains satisfied throughout the iterations within a good degree of accuracy.

3.4 Numerical tests

An extensive set of numerical tests was performed to illustrate the important features of the domain decomposition algorithm introduced above, and to compare it with a standard scheme, commonly used in large scale electronic structure calculations.

3.4.1 Setting of the algorithm and of the tests

3.4.1.1 Molecular systems used for the tests

Numerical tests on the algorithm presented above were performed on three chemical systems. The first two systems both have formula COH-$(CO)_{n_m}$-COH. They differ in their Carbon-Carbon interatomic distances. For system $\mathcal{P}_1$, this distance is fixed to 5 atomic units, while it is fixed to 4 for system $\mathcal{P}_2$. On the other hand, our third system, denoted by $\mathcal{P}_3$ has formula CH_3-$(CH_2)_{n_m}$-CH_3.

For each of the three systems $\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$, several numbers n_m of monomers were considered. A geometry optimization was performed using the GAUSSIAN package [45] in order to fix the internal geometrical parameters of the system. The only exception to this is the Carbon-Carbon distance for $\mathcal{P}_1$ and $\mathcal{P}_2$, which, as said above, is fixed *a priori*. Imposing the Carbon-Carbon distance allows to control the sparsity of the matrices H and S (the larger the distance, the sparser the matrices). Although not physically relevant, fixing the Carbon-Carbon distance is therefore useful for the purpose of numerical tests.

3.4.1.2 Data, parameters and initialization

For an extremely large number n_m of monomers, the matrices H, S, and $D_\star$ cannot be generated directly with the GAUSSIAN package. We therefore make a

	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$
n	130	200	308
q	50	80	126
Bandwith of S	59	79	111
Bandwith of H	99	159	255
Cut-off for entries of H	10^{-12}	10^{-12}	10^{-10}
Cut-off for entries of D	10^{-11}	10^{-11}	10^{-7}
Size of first block	$m_1 = 67$	$m_1 = 105$	$m_1 = 136$
Size of last block	$m_p = 67$	$m_p = 106$	$m_p = 137$
Size of a generic block	$m_i = 56$	$m_i = 84$	$m_i = 104$

TAB. 3.1 – Localization parameters and initial size of the blocks used in the tests

periodicity assumption. For large values of n_m, these matrices approach a periodic pattern (leaving apart, of course, the "boundary layer", that is the terms involving orbitals close to one end of the linear molecule). So, we first fix some n_m sufficiently large, but for which a direct calculation with Gaussian is feasible, and construct H, S. The matrices H and S, as well as the ground-state density matrix $D_\star$, and the ground-state energy E_0, are then obtained for arbitrary large n_m assuming periodicity out of the "boundary layer". Likewise, the gap γ in the eigenvalues of H is observed to be constant, for each system, irrespective of the number n_m of polymers, supposedly large. Proceeding so, the gap for systems $\mathcal{P}_1$, $\mathcal{P}_2$, and $\mathcal{P}_3$ is respectively evaluated to 0.00104, 0.00357, and 0.0281.

For our MDD approach, localization parameters are needed. They are shown in Table 3.4.1.2 below. Additionally, we need to provide the algorithm with an initial guess on the size m_i of the blocks. Based on physical considerations on the expected repartition of the electrons in the molecule and on the expected localization of the orbitals, the sizes were fixed to values indicated in the table Tab 3.4.1.2. The specific block C_i is then initialized in one of the following three manners :

- strategy $\mathcal{I}_1$: the entries of C are generated randomly, which of course generically yields a bad initial guess way ;
- strategy $\mathcal{I}_2$: each block C_i consists of the lowest m_i (generalized) eigenvectors associated to the corresponding block matrices H_i and S_i in the matrices H and S, respectively. This provides with an initial guess, depending on the matrices H and S, thus of better quality than the random one provided by strategy $\mathcal{I}_1$;
- strategy $\mathcal{I}_3$: the initial guess provided by $\mathcal{I}_2$ is optimized with the local fine solver described in section 3.3.2.

3.4.1.3 Implementation details

Exact diagonalizations in the local steps are performed with the routine *dsbgv.f* from the LAPACK package [123]. In the global step, the resolution of the linear

system involving the Hessian matrix is performed iteratively, using SYMMLQ [143]. Diagonal preconditionning is used to speed up the resolution.

The calculations have been performed using only one processor of a bi-processor Intel Pentium IV-2.8 GHz.

3.4.1.4 Criteria for comparison of results

For assesment of the quality of the results, we have used two criteria, regarding the ground-state energy and the ground-state density matrix, respectively. For either quantity, the reference calculation is the calculation using the Gaussian package [45]. The quality of the energy is measured using the relative error $e_E = \dfrac{|E - E_0|}{|E_0|}$. For evaluation of the quality of the density matrix, we use the L^∞ matrix norm

$$e_\infty = \sup_{(i,j) \text{ s.t. } |H_{ij}| \leq \varepsilon} \left| D_{ij} - [D_\star]_{ij} \right|, \tag{3.30}$$

where we fix $\varepsilon = 10^{-10}$. The introduction of the norm (3.30) is consistent with the cut-off performed on the entries of H (thus the exact value of ε chosen). Indeed, in practice, the matrix D is only used for the calculations of various observables (for instance electronic energy and Hellman-Feynman forces), all of the form $\mathrm{Tr}(AD)$ where the symmetric matrix A shares the same pattern as the matrix H (see [1] for details). The result is therefore not sensitive to entries with indices (i, j) such that $|H_{ij}|$ is below the cut-off value.

3.4.2 Illustration of the role of the local and global steps

Our MDD method consists in three ingredients :
- the local optimization of each block performed in the local step ;
- the transfer of vectors from some blocks to other blocks, along with the modification of the block sizes m_i, again in the local step ;
- the optimization performed in the global step.

To highlight the necessity of each of the ingredients, and their impacts on the final result, we compare our MDD algorithm with three simplified variants. Let us denote by
- strategy $\mathcal{S}_1$: local optimization of the blocks, without allowing variations of the block sizes, and no global step ;
- strategy $\mathcal{S}_2$: full local step (as defined in Section 3.3.2), no global step ;
- strategy $\mathcal{S}_3$: local optimization of the blocks, without allowing for variations of the block sizes, and global step ;
- strategy $\mathcal{S}_4$: full algorithm.

We compare the rate of convergence for the above four strategies. Two categories of tests are performed, depending on the quality of the initial guess. The results displayed on Figures 3.6 to 3.7 concern polymer $\mathcal{P}_1$ with $n_m = 801$ monomers. This

corresponds to $N_b = 8050$ and $N = 5622$. Analogous tests were performed on $\mathcal{P}_2$ and $\mathcal{P}_3$, but we do not present them here, for brevity.

The energy of the ground state of this matrix (i.e. the minimum of (3.6)) is $E_0 = -27663.484$. The number of blocks considered is $p = 100$. For the global step, we have collected these 100 blocks in 99 overlapping groups of 2 blocks. Interestingly, such a partition provides with optimal results regarding CPU time and memory requirement. It is observed in Figures 3.6 to 3.7 that $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ are not satisfactory for they converge towards some local, non global, minima of (3.9) whatever the initial guess. The failure of the strategy S_3 performed on the initial guess $\mathcal{I}_2$ is surprising : this initial guess is not good enough. Indeed, if the initial guess is $\mathcal{I}_3$, we check numerically that the strategies $\mathcal{S}_3$ and $\mathcal{S}_4$ behave identically. Notice that the strategy $\mathcal{I}_3$ is identical to $\mathcal{S}_2$ applied to the initial guess $\mathcal{I}_2$.

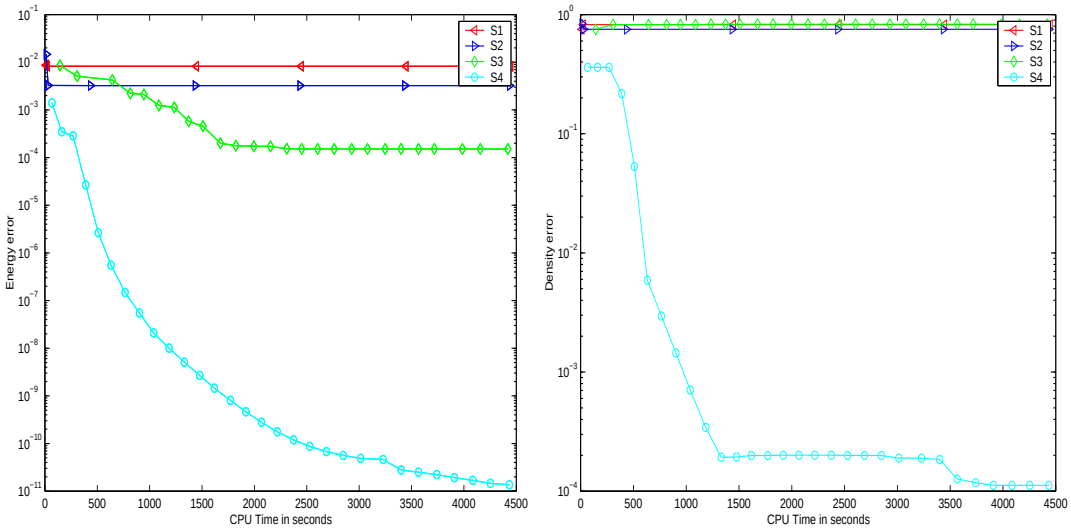


FIG. 3.6 – Density (right) and energy (left) errors versus CPU time obtained with a bad initial guess ($\mathcal{I}_1$).

We also remark that the strategy $\mathcal{S}_4$ performs very well whatever the initial guess (see Fig. 3.6 and Fig. 3.7). The same behavior is observed for the polymers $\mathcal{P}_2$ and $\mathcal{P}_3$. Finally, after orthonormalization, the Density Matrix Minimization (DMM) method [81] failed with the random initial guess and reveals very slow with the initial guess $\mathcal{I}_2$. That is the reason why we consider the initial guess $\mathcal{I}_3$ to compare these methods.

3.4.3 Comparison with two other methods

Having emphasized the usefulness of all the ingredients of our MDD algorithm, we now compare it to two other algorithms :

– the diagonalization routine *dsbgv.f* from the LAPACK library ;
– the Density Matrix Minimization (DMM) method [81].

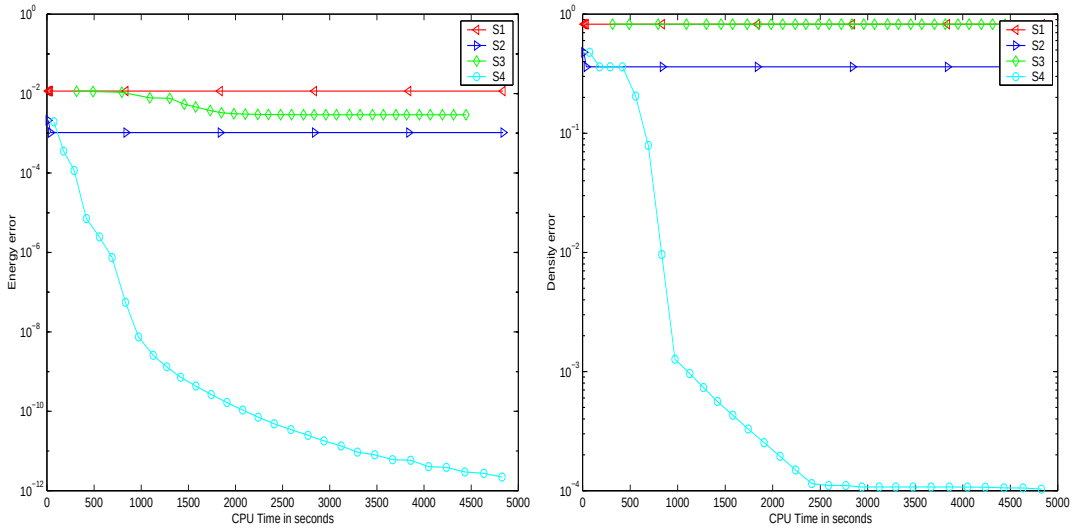


FIG. 3.7 – Density (right) and energy (left) errors versus CPU time obtained with a better initial guess ($\mathcal{I}_2$).

These two algorithms are seen as prototypical approaches for standard diagonalization algorithms and linear scaling techniques respectively. They are only used here for comparison purposes. Regarding linear scaling methods, two other popular approaches are the Fermi Operator method [58] and the McWeeny iteration method [90]. We have observed that, at least in our own implementation, based on the literature, they are outperformed by the DMM method for the actual chemical systems we have considered. We therefore take DMM as a reference method for our comparison.

Recall that the routine *dsbgv.f* consists in the three-step procedure
- transform the generalized eigenvalue problem into a standard eigenvalue problem by applying a Cholesky factorization to S ;
- reduce the new matrix to be diagonalized to a tridiagonal form ;
- compute its eigenelements by using the implicit QR method.

The algorithmic complexity of this approach is in N_b^3 and the required memory scales as N_b^2.

For the description of DMM method, we refer to [81]. Let us only mention here that this approach consists in a minimization procedure, applied to the energy expressed in terms of the density matrix. Both the algorithmic complexity and the memory needed for performing the DMM approach scale linearly with respect to the size N_b of the matrix. The DMM method is initialized with the density matrix $D = CC^t$ computed with the initial guess C of the domain decomposition method. Two important points for the tests shown below are the following.

First, we perform a cut-off on the coefficients on the various matrices manipulated throughout the calculation : only the terms of the density matrices within the frame defined in Figure 3.3 are taken into account. Such a cut-off has some impact on the

qualities of the results obtained with the DMM method. We are however not able to design a better comparison.

Second, the DMM method requires the knowledge of the Fermi level (as is the case for the linear scaling methods commonly used in practice to date). The determination of the Fermi level is the purpose of an outer optimization loop. In contrast, the MDD approach computes an approximation of the Fermi level at each iteration. Here, for the purpose of comparison, we *provide* DMM with the exact value of the Fermi level. Consequently, the CPU times for the DMM method displayed in the sequel are underestimated.

We emphasize that the routine *dsbgv.f* computes the entire spectrum of the matrix, both eigenvalues and eigenvectors. In contrast, the MDD approach only provides with the lowest N eigenvalues, among N_b, and the projector on the vector space spanned by the corresponding eigenvectors, not the eigenvectors themselves.

3.4.3.1 Comparison with Direct diagonalization and DMM

We have computed the ground states of the polymers $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$ with the three methods (direct diagonalization, DMM and MDD) and for various numbers n_m of monomers, corresponding to matrix sizes N_b in the range 10^3-10^5. For DMM and MDD, the initial guess is generated following the strategy $\mathcal{I}_3$.

The results regarding the CPU time at convergence and the memory requirement are displayed for the three polymers in Table 3.4.3.1. For small values of N_b, i.e. up to around 10^4, the results observed for the direct diagonalization, DMM and MDD agree. The CPU times for our MDD approach scale linearly with N_b. For larger values of N_b, the limited memory prevented us from either performing an exact diagonalization or from implementing DMM. So, we extrapolate the CPU time and memory requirement according to the scaling observed for smaller N_b.

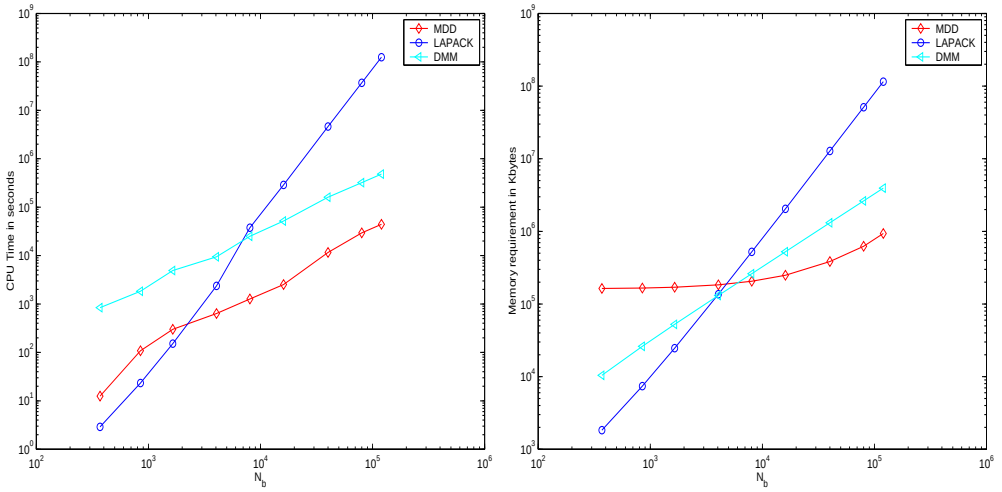
For the polymer $\mathcal{P}_3$, the data for the DMM method are not plotted as the DMM method does not converge when the number of monomers exceeds 10^3. From our point of view, it comes from the truncation errors which cause the divergence of the method (note that the truncation strategy we consider here is very simple).

3.4.3.2 Comparison with DMM and a hybrid strategy

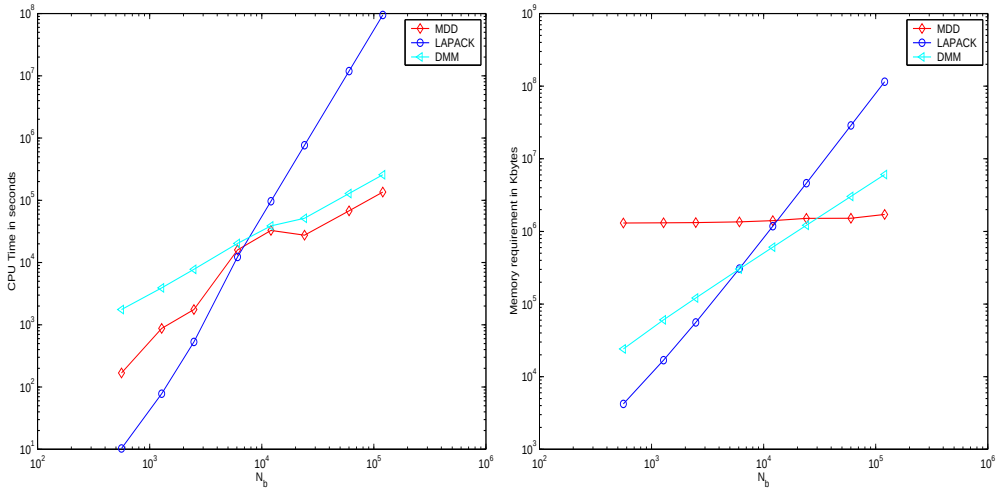
We now concentrate on the two approaches that scale linearly, namely DMM and MDD. We consider

- $\mathcal{P}_1$ with 4001 monomers, corresponding to $N_b = 40050$,
- $\mathcal{P}_2$ with 2404 monomers, corresponding to $N_b = 24080$,
- $\mathcal{P}_3$ with 208 monomers, corresponding to $N_b = 854$.

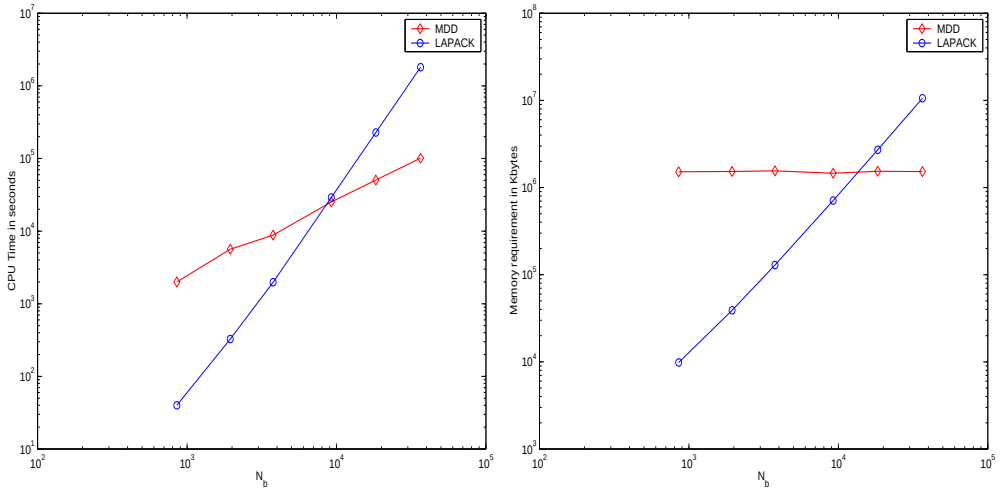
These particular values have been chosen for the purpose of having simple values for the numbers of blocks. For each of the three polymers, we compare the DMM and MDD methods initialized by the strategy $\mathcal{I}_3$ and a hybrid strategy. The hybrid strategy consists of a certain number of iterations performed with MDD, until



Polymer $\mathcal{P}_1$



Polymer $\mathcal{P}_2$



Polymer $\mathcal{P}_3$

TAB. 3.2 – Scaling of the CPU time (left) and memory requirement (right) for the three polymers.

convergence is reached for this method, followed by iterations with DMM. We use the following stopping criterion for MDD :

$$\|D_n - D_{n-1}\| \geq \|D_{n-1} - D_{n-2}\| \quad \text{and} \quad \|D_n - D_{n-1}\| \leq \epsilon_a \tag{3.31}$$

where ϵ_a is a threshold parameter. We take $\epsilon_a = 10^{-4}$, respectively $\epsilon_a = 10^{-3}$, for the polymer $\mathcal{P}_1$, respectively $\mathcal{P}_2$ and $\mathcal{P}_3$.

The Table 3.4.3.2 shows the evolution of the error in density versus CPU time. The hybrid version is demonstrated to be a very efficient combination of the two algorithms.

For completeness, let us highlight the temporary increase for the error in density when MDD is used on $\mathcal{P}_2$. Analogously, the energy of the current solution, which is actually below the reference energy, also increases. In fact, this is due to a loss of precision in the orthonormality constraints. In MDD, these constraints are not imposed exactly at each iteration, but only approximately (see equation 3.24).

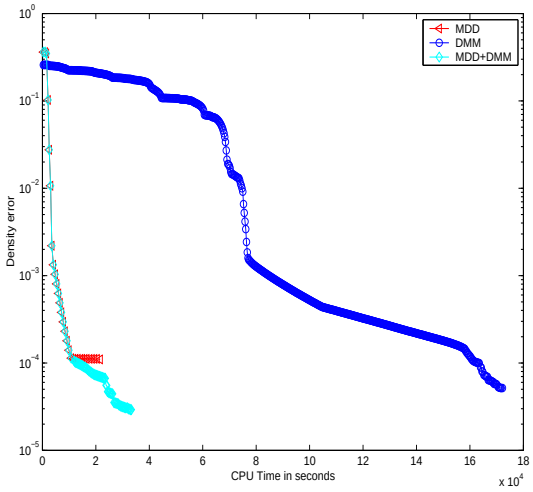
Finally, we report on Table 3.5 the results obtained with MDD for the largest possible case that can be perfomed on our platform, owing to memory limitation. We use the initial guesses obtained with the strategy $\mathcal{I}_3$. Notice that for the local step the memory requirement scales linearly with respect to the number n_m of monomers, while for the global step, the memory requirement is independent of n_m. Therefore, for large polymers, the memory needed by MDD is controled by the local step. In contrast, for small polymers, the most demanding step in terms of memory is the global step.

3.5 Conclusions and remarks

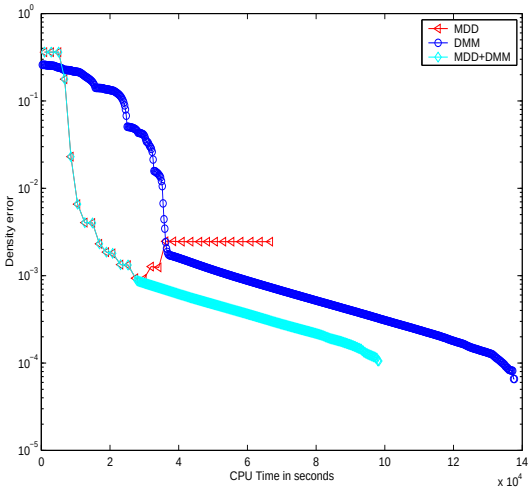
The domain decomposition algorithm introduced above performs well, in comparison to the two standard methods considered. More importantly, our approach is an effective *preconditionning technique* for DMM iterations. Indeed, MDD provides a rapid and accurate approximation, both in terms of energy and density matrix, regardless of the quality of the initial guess. In contrast, DMM outperforms MDD when the initial guess is good, but only performs poorly, or may even diverge, when this is not the case. The combination of the two methods seems to be optimal. More generally, our MDD algorithm could constitute a good preconditionner to all variational methods, such as the Orbital Minimization method [87].

Regarding the comparison with DMM, the following comments are in order.

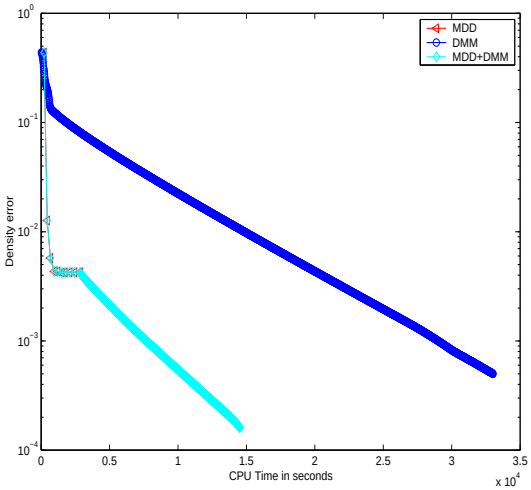
- All our calculations have been performed on a single processor machine. Potentially, both DMM and MDD should exhibit the same speed-up when parallelized. We therefore consider the comparison valid, at least qualitatively, for parallel implementations. The parallelization of the MDD is currently in progress, and hopefully will confirm the efficiency of the approach.
- We recall the Fermi level has to be provided to the DMM method. This is an additional argument in favor of the MDD approach.



Polymer $\mathcal{P}_1$



Polymer $\mathcal{P}_2$

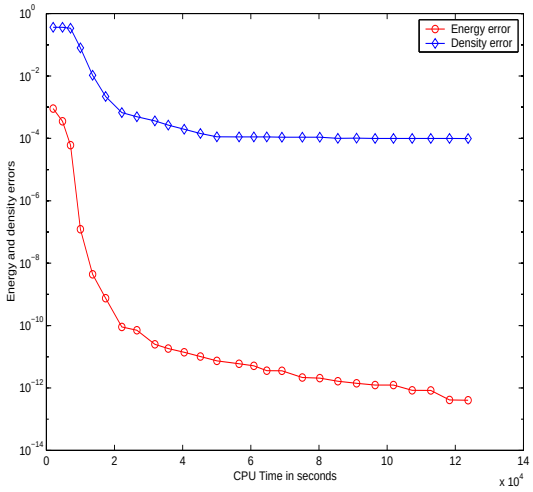


Polymer $\mathcal{P}_3$

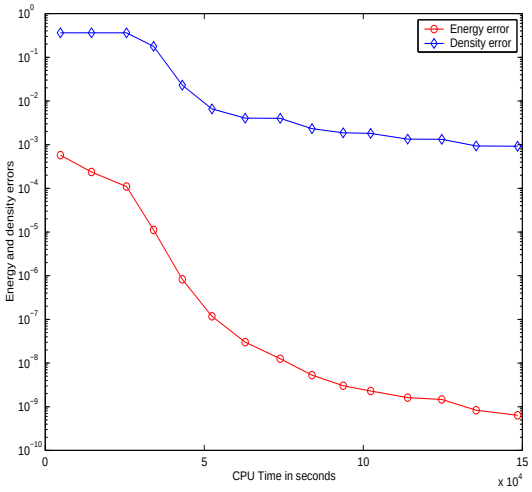
TAB. 3.3 – Evolution of the density error with the CPU time for the polymer $\mathcal{P}_1$ made of 4001 monomers (top), $\mathcal{P}_2$ made of 2404 monomers (middle), polymer $\mathcal{P}_3$ made of 208 monomers (bottom).

- The MDD method, in contrast to the other linear scaling methods, does not perform any truncation in the computations. So, once the profile of C is choosen, the method does not suffer of any instabilities, contrary to DMM (or OM) for which divergences have been observed for the polymer $\mathcal{P}_3$.
- The domain decomposition method makes use of several threshold parameters. For the three polymers we have considered, the optimal values of these parameters, except for the stopping criterion ϵ_a (equation (3.31)), are the same. We do not know yet if this interesting feature is a general rule.
- Recall our method solves problem (3.9), which is only an approximation of problem (3.8). Therefore, the relative error obtained in the limit is only a measure of the difference between (3.9) and (3.8). In principle, such a difference could be made arbitrarily small by an appropriate choice of the parameters of problem (3.9).
- Finally, let us emphasize that there is much room for improvement in both the local and the global steps. We have designed an overall multilevel strategy that performs well, but each subroutine may be significantly improved. Another interesting issue is the interplay between the nonlinear loop in the Hartree-Fock or Kohn-Sham problems (Self-Consistent Field - SCF - convergence [1,27,127]) and the linear subproblem considered in the present article. Future efforts will go in these directions.

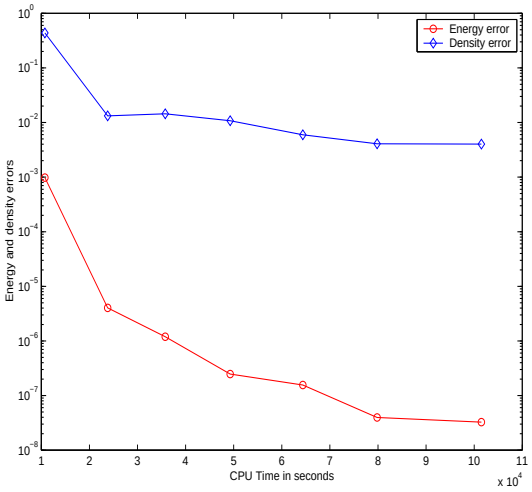
ACKNOWLEDGMENTS. We would like to thank Guy Bencteux (EDF) for valuable discussions and for his help in the implementation. C.L.B. and E.C. would like to acknowledge many stimulating discussions with Richard Lehoucq (Sandia National Laboratories).



Polymer $\mathcal{P}_1$



Polymer $\mathcal{P}_2$



Polymer $\mathcal{P}_3$

TAB. 3.4 – Evolution of the MDD energy and density errors versus CPU time for the polymer $\mathcal{P}_1$ (top) made of 20001 monomers ($N_b = 200050$), the polymer $\mathcal{P}_2$ (middle) made of 12004 monomers ($N_b = 120080$), the polymer $\mathcal{P}_3$ (bottom) made of 5214 monomers ($N_b = 36526$).

3.6 Une preuve de convergence dans un cadre simplifié

3.6.1 Présentation du problème

On considère le problème (3.8) pour $N = 2$. On désigne par X, respectivement Y, le bloc formé de la première colonne, respectivement de la seconde, de C. Le problème (3.8) s'écrit

$$\inf\big\{(HX,X)+(HY,Y),\ (X,X)=1,\ (Y,Y)=1,\ (X,Y)=0\big\}. \tag{3.32}$$

Soit (X^n, Y^n) l'itéré courant, le calcul de (X^{n+1}, Y^{n+1}) suivant la méthode présentée dans la première partie de ce chapitre s'effectue suivant

$$(I)\ \left|\begin{array}{rcl} \tilde{X} &=& \mathrm{arginf}\big\{(HX,X),\ (X,X)=1,\ (X,Y^n)=0\big\},\\ \tilde{Y} &=& \mathrm{arginf}\big\{(HY,Y),\ (Y,Y)=1,\ (\tilde{X},Y)=0\big\}. \end{array}\right. \tag{3.33}$$

$$(II)\ \left|\begin{array}{rcl} (X^{n+1},Y^{n+1}) &=& \big(X_{\alpha^*},Y_{\alpha^*}\big),\\ \alpha^* &=& \inf\limits_{\alpha\in\mathbb{R}}\big\{(HX_\alpha,X_\alpha)+(HY_\alpha,Y_\alpha)\big\},\\ (X_\alpha,Y_\alpha) &=& \left(\dfrac{\tilde{X}+\alpha\tilde{Y}}{\sqrt{1+\alpha^2}},\dfrac{\tilde{Y}-\alpha\tilde{X}}{\sqrt{1+\alpha^2}}\right). \end{array}\right.$$

On remarque que l"étape (II) est inutile suite à la simplicité du problème (3.32). En effet,

$$\forall\,\alpha\in\mathbb{R},\quad (HX_\alpha,X_\alpha)+(HY_\alpha,Y_\alpha)=(HX,X)+(HY,Y).$$

C'est pourquoi, on omet cette étape dans la suite. Soient (c_i,ϵ_i) les éléments propres de H, on suppose les réels ϵ_i distincts deux à deux. On introduit les matrices $D_{c_i} = c_i c_i^t$, et, à chaque itération les matrices $(D_X^n, D_Y^n, D^n)_n$ par

$$\forall\, n\in\mathbb{N},\quad (D_X^n,D_Y^n,D^n)=\Big(X^n(X^n)^t,Y^n(Y^n)^t,D_X^n+D_Y^n\Big).$$

3.6.2 Preuve de convergence

L'écriture des équations d'Euler-Lagrange correspondant à (3.33) conduit aux relations suivantes

$$\begin{array}{rcl} HX^{n+1} &=& \lambda^{n+1}X^{n+1}+\nu^{n+1}Y^n,\\ HY^{n+1} &=& \delta^{n+1}X^{n+1}+\mu^{n+1}Y^{n+1},\\ (X^{n+1},Y^n) &=& (X^{n+1},Y^{n+1}) \;=\; 0,\\ (X^{n+1},X^{n+1}) &=& (Y^{n+1},Y^{n+1}) \;=\; 1. \end{array} \tag{3.34}$$

D'où

$$\begin{array}{rcl} \lambda^{n+1} & = & (HX^{n+1}, X^{n+1}), \\ \nu^{n+1} & = & (HX^{n+1}, Y^{n}), \\ \delta^{n+1} & = & (HY^{n+1}, X^{n+1}), \\ \mu^{n+1} & = & (HY^{n+1}, Y^{n+1}). \end{array} \tag{3.35}$$

3.6.2.1 Etude des suites $(\lambda^n)_n$, $(\nu^n)_n$, $(\delta^n)_n$, $(\mu^n)_n$

Lemme 3.6.1 *Les suites $(\lambda^n)_n$ et $(\mu^n)_n$ sont des suites convergentes. On note λ et μ leur limite respectivement.*

Par (3.35), les suites $(\lambda^n)_n$ et $(\mu^n)_n$, sont minorées par ϵ_1. Par ailleurs, il vient par (3.33)

$$\left(HX^{n+1}, X^{n+1}\right) \leq \left(HX^{n}, X^{n}\right) \tag{3.36}$$

soit en utilisant (3.34)

$$\lambda^{n+1} \leq \lambda^{n}.$$

Par conséquent, $(\lambda^n)_n$ est une suite décroissante minorée. La suite converge donc vers un réel noté λ. De même, $(\mu^n)_n$ est une suite décroissante minorée donc convergente vers un réel noté μ. On peut toujours supposer

$$\lambda \leq \mu. \tag{3.37}$$

Dans le cas contraire, il suffit d'échanger X^0 et Y^0 pour obtenir (3.37).

Lemme 3.6.2 *Les suites $(\delta^n)_n$ et $(\nu^n)_n$ sont des suites convergentes vers la même limite positive notée ν.*

Dans un premier temps, on remarque qu'on peut toujours supposer $\nu^n \geq 0$, respectivement $\delta^n \geq 0$, pour tout n. En effet, dans le cas contraire, il suffit de changer le signe de X^n, respectivement Y^n, de façon à satisfaire (3.34). Les suites $(\delta^n)_n$ et $(\nu^n)_n$ sont donc minorées.
Par ailleurs, il vient en utilisant le caractère symétrique de H, (3.34) et (3.35)

$$\begin{array}{rcl} \delta^{n+1} & = & (HY^{n+1}, X^{n+1}), \\ & = & (Y^{n+1}, HX^{n+1}), \\ & = & \lambda^{n+1}(Y^{n+1}, X^{n+1}) + \nu^{n+1}(Y^{n+1}, Y^{n}), \\ & = & \nu^{n+1}(Y^{n+1}, Y^{n}). \end{array} \tag{3.38}$$

De manière similaire,

$$\nu^{n+1} = \delta^{n}(X^{n+1}, X^{n}). \tag{3.39}$$

X_n et Y_n étant normés, on obtient par l'inégalité de Cauchy-Schwartz

$$|(X^{n+1}, X^{n})| \leq 1 \quad \text{et} \quad |(Y^{n+1}, Y^{n})| \leq 1.$$

Il vient ensuite à l'aide de (3.38) et (3.39)

$$\ldots \leq \delta^{n+1} \leq \nu^{n+1} \leq \delta^n \leq \nu^n \leq \ldots \tag{3.40}$$

Par conséquent, $(\delta^n)_n$ et $(\nu^n)_n$ sont décroissantes et par suite convergent. En passant à la limite dans (3.40), elles ont la même limite qu'on note ν.

3.6.2.2 Cas $\nu = 0$

Lemme 3.6.3 *Les suites $(D_X^n)_n$ et $(D_Y^n)_n$ sont convergentes. Les suites $(D_X^n, \lambda^n)_n$ et $(D_Y^n, \mu^n)_n$ convergent vers $(D_{c_\alpha}, \epsilon_\alpha)$ et $(D_{c_\beta}, \epsilon_\beta)$ avec $\alpha < \beta$.*

Comme H est symétrique, on a

$$\begin{aligned} \forall\, 1 \leq k \leq N_b, \quad (c_k, HX^{n+1}) &= (Hc_k, X^{n+1}), \\ &= \epsilon_k (c_k, X^{n+1}). \end{aligned}$$

Il vient par (3.34)

$$\forall\, 1 \leq k \leq N_b, \quad (\epsilon_k - \lambda^{n+1})(c_k, X^{n+1}) = \nu^{n+1}(c_k, Y^n). \tag{3.41}$$

On obtient ensuite par Cauchy-Schwartz $|(c_k, Y^n)| \leq 1$. Par conséquent, en passant à la limite dans (3.41), on obtient en utilisant les lemmes 3.6.1-3.6.2 et l'hypothèse sur ν

$$\forall\, 1 \leq k \leq N_b, \quad \lim_{n\to\infty} (\epsilon_k - \lambda)(c_k, X^{n+1}) = 0. \tag{3.42}$$

λ est donc égal à une valeur propre de H. En effet dans le cas contraire, comme $(c_i)_i$ forment une base de $\mathbb{R}^{N_b}$, (3.42) implique que $(X^n)_n$ converge vers le vecteur nul de norme nulle. Comme

$$\forall\, k \neq \alpha, \quad \lim_{n\to\infty} (c_k, X^{n+1}) = 0, \tag{3.43}$$

la suite $(D_X^n)_n$ converge vers D_{c_α} suite à l'hypothèse sur les valeurs propres de H. De même μ est égal à une valeur propre de ϵ_β de H, et, la suite $(D_Y^n)_n$ converge vers D_{c_β}. Le passage à la limite dans (3.34) implique $\alpha \neq \beta$, soit par (3.37) $\alpha < \beta$.

Lemme 3.6.4 *Le couple (D_α, D_β) est le couple (D_{c_1}, D_{c_2}).*

Soit i_0 un entier de $\{1, \ldots, N_b\}$, on définit le vecteur e_{i_0} par

$$e_{i_0} = c_{i_0} - (c_{i_0}, X^{n+1}) X^{n+1}.$$

Par construction, $(X^{n+1}, e_{i_0}) = 0$. En utilisant (3.33) et (3.35), on a

$$\begin{aligned} \mu^{n+1} &\leq \left(H \frac{e_{i_0}}{\|e_{i_0}\|}, \frac{e_{i_0}}{\|e_{i_0}\|} \right), \\ &\leq \frac{1}{\|e_{i_0}\|^2} \Big(\epsilon_{i_0} - 2\epsilon_{i_0} (c_{i_0}, X^{n+1})^2 + \lambda^{n+1} (c_{i_0}, X^{n+1})^2 \Big). \end{aligned} \tag{3.44}$$

En passant à la limite dans (3.44), on obtient à l'aide de (3.43)

$$\forall\, i_0 \neq \alpha,\ \beta \quad \epsilon_\alpha < \epsilon_\beta \leq \epsilon_{i_0}. \tag{3.45}$$

On finit ensuite la démonstration par l'absurde. En effet, dans un premier temps on suppose que α, et donc β, sont différents de 1. Il suffit de choisir $i_0 = 1$ dans (3.45) pour aboutir à une contradiction. Par conséquent, α est égal à 1. Dans un second temps, on suppose $\beta \neq 2$. Il suffit de choisir $i_0 = 2$ dans (3.45) pour aboutir à une contradiction.

3.6.2.3 Cas $\nu \neq 0$

Par (3.34), la suite des couples $\Big(\big(X^n, Y^n\big)\Big)_n$ est bornée. Elle converge donc à extraction près. On se donne $\Big(X^{p(n)}, Y^{p(n)}\Big)_n$ une extraction convergente, soit

$$\lim_{n\to\infty} X^{p(n)} = X_p,\ \lim_{n\to\infty} Y^{p(n)}) = Y_p. \tag{3.46}$$

On définit D_{X_p}, D_{Y_p} et D_p par

$$(D_{X_p}, D_{Y_p}, D_p) = \Big(X_p(X_p)^t, Y_p(Y_p)^t, D_{X_p} + D_{Y_p}\Big).$$

Lemme 3.6.5 *La limite (X_p, Y_p) appartient à un sous espace vectoriel engendré par deux vecteurs propres $(c_{\alpha_p}, c_{\beta_p})$ de H avec $\epsilon_{\alpha_p} < \epsilon_{\beta_p}$.*

(3.38) pour $p(n)$ s'écrit

$$\delta^{p(n)+1} = \nu^{p(n)+1}\Big(Y^{p(n)+1}, Y^{p(n)}\Big). \tag{3.47}$$

Comme on se place dans le cas $\nu \neq 0$, le passage à la limite dans (3.47) implique

$$\lim_{n\to\infty} \Big(Y^{p(n)+1}, Y^{p(n)}\Big) = 1. \tag{3.48}$$

En utilisant (3.34), $\big|Y^{p(n)+1} - Y^{p(n)}\big| = 2\Big(1 - \Big(Y^{p(n)+1}, Y^{p(n)}\Big)\Big)$. Il vient ensuite par (3.48)

$$\lim_{n\to\infty} \big|Y^{p(n)+1}, Y^{p(n)}\big| = 0. \tag{3.49}$$

En procédant de même avec $X^{p(n)}$, on montre en associant (3.46) , (3.49) et l'inégalité de Cauchy-Schwartz

$$\lim_{n\to\infty} X^{p(n)+1} = X_p,\ \lim_{n\to\infty} Y^{p(n)+1} = Y_p.$$

Par conséquent, en passant à la limite dans (3.34) écrit en $p(n)$, (X_p, Y_p) vérifie donc

$$(\mathcal{S}) \left\{ \begin{array}{rcl} HX_p & = & \lambda X_p + \nu Y_p, \\ HY_p & = & \nu X_p + \mu Y_p. \end{array} \right.$$

L'espace vectoriel engendré par (X_p, Y_p) est stable par H. Il est donc engendré par deux éléments propres (c_{α_p}, α_p) et (c_{β_p}, β_p) de H. La diagonalisation de la matrice H restreinte à ce sous-espace vectoriel conduit à

$$\begin{array}{rcl} \epsilon_{\alpha_p} & = & \left(\lambda + \mu - \sqrt{(\mu-\lambda)^2 + \nu^2}\right)/2, \\ \epsilon_{\beta_p} & = & \left(\lambda + \mu + \sqrt{(\mu-\lambda)^2 + \nu^2}\right)/2, \\ \epsilon_{\alpha_p} + \epsilon_{\beta_p} & = & \lambda + \mu. \end{array} \tag{3.50}$$

Comme $\nu \neq 0$, (D_{X_p}, D_{Y_p}) ne coïncide pas avec $(D_{c_{\alpha_p}}, D_{c_{\beta_p}})$.

Lemme 3.6.6 *Les vecteurs $(c_{\alpha_p}, c_{\beta_p})$ et donc D_p ne dépendent pas de l'extraction p choisie. On note (α, β) les valeurs de (α_p, β_p) $(\alpha < \beta)$. Par ailleurs, les matrices (D_{X_p}, D_{Y_p}) ne dépendent pas de l'extraction p choisie.*

λ et μ sont indépendants de l'extraction choisie. En utilisant (3.50), $(\epsilon_{\alpha_p}, \epsilon_{\beta_p})$ sont indépendants de l'extraction p choisie. D_p étant le projecteur associé au sous-espace engendré par (X_p, Y_p), il est donc indépendant de l'extraction p choisie. Notamment, pour toute extraction p convergente

$$D_{X_p} + D_{y_p} = D_{c_\alpha} + D_{c_\beta}.$$

Ensuite par le lemme 3.6.5 associé aux contraintes d'orthonormalité satisfaites par X_p et Y_p, on a l'existence d'un réel $\theta_p \in]0, 2\pi[$ tel que

$$cos(\theta_p) sin(\theta_p) \neq 0$$

et

$$\left\{ \begin{array}{rcl} X_p & = & cos(\theta_p) c_\alpha + sin(\theta_p) c_\beta, \\ Y_p & = & -sin(\theta_p) c_\alpha + cos(\theta_p) c_\beta. \end{array} \right. \tag{3.51}$$

En multipliant (3.51) par H, on obtient par identification avec $(\mathcal{S})$ les égalités suivantes :

$$\left\{ \begin{array}{rcl} sin(\theta_p) cos(\theta_p) & = & \dfrac{\nu}{\epsilon_\beta - \epsilon_\alpha}, \\ sin^2(\theta_p) & = & \dfrac{\lambda - \epsilon_\alpha}{\epsilon_\beta - \epsilon_\alpha} = \dfrac{\epsilon_\beta - \mu}{\epsilon_\beta - \epsilon_\alpha}, \\ cos^2(\theta_p) & = & \dfrac{\epsilon_\beta - \lambda}{\epsilon_\beta - \epsilon_\alpha} = \dfrac{\mu - \epsilon_\alpha}{\epsilon_\beta - \epsilon_\alpha}. \end{array} \right. \tag{3.52}$$

Or, par définition de D_{X_p} et D_{Y_p}, il vient par (3.51)

$$\left| \begin{array}{rcl} D_{X_p} & = & cos^2(\theta_p) D_{c_\alpha} + cos(\theta_p) sin(\theta_p)(c_\beta c_\alpha^t + c_\alpha c_\beta^t) + sin^2(\theta_p) D_{c_\beta} \\ D_{Y_p} & = & sin^2(\theta_p) D_{c_\alpha} - cos(\theta_p) sin(\theta_p)(c_\beta c_\alpha^t + c_\alpha c_\beta^t) + cos^2(\theta_p) D_{c_\beta} \end{array} \right. .$$

On conclut par (3.52) que le couple (D_{X_p}, D_{Y_p}) est indépendant de l'extraction p choisie.

Lemme 3.6.7 *α est égal à 1. β est égal à 2 si et seulement si*

$$\frac{\epsilon_3 - \epsilon_2}{\epsilon_3 - \epsilon_1} > \frac{1}{2}.$$

On procède ici comme pour la démonstration du lemme 3.6.2. Soit i_0 un entier de $\{1, \ldots, N_b\}$, on définit le vecteur e_{i_0} par

$$e_{i_0} = c_{i_0} - \left(c_{i_0}, X^{p(n)+1}\right) X^{p(n)+1}.$$

Par construction, $\left(X^{p(n)+1}, e_{i_0}\right) = 0$. En utilisant (3.33) et (3.35), il vient

$$\begin{array}{rcl} \mu^{p(n)+1} & \leq & \left(H \dfrac{e_{i_0}}{\|e_{i_0}\|}, \dfrac{e_{i_0}}{\|e_{i_0}\|}\right), \\ & \leq & \dfrac{1}{\|e_{i_0}\|^2}\left(\epsilon_{i_0} - 2\epsilon_{i_0}(c_{i_0}, X^{p(n)+1})^2 + \lambda^{p(n)+1}(c_{i_0}, X^{p(n)+1})^2\right). \end{array} \tag{3.53}$$

Par ailleurs, on déduit du lemme 3.6.6 associé à (3.34), (3.46) et (3.49)

$$\forall\, i_0 \neq \alpha,\ \beta \quad \lim_{n\to\infty}\left(X^{p(n)}, c_{i_0}\right) = \lim_{n\to\infty}\left(X^{p(n)+1}, c_{i_0}\right) = \left(X^p, c_{i_0}\right) = 0. \tag{3.54}$$

Par conséquent en passant à la limite dans (3.53), on obtient à l'aide de (3.37) et (3.54)

$$\forall\, i_0 \neq \alpha,\ \beta \quad \lambda \leq \mu \leq \epsilon_{i_0}. \tag{3.55}$$

Par ailleurs, on déduit de (3.50)

$$\epsilon_\alpha < \lambda \leq \mu < \epsilon_\beta. \tag{3.56}$$

On obtient donc par (3.55) et (3.56) lorsque α n'est pas égal à 1

$$\epsilon_\alpha < \lambda < \epsilon_1,$$

et donc α est égal à 1. Supposons que β est différent de 2. Par (3.50) et (3.55), on a

$$\begin{array}{rcl} \lambda + \mu & = & \epsilon_1 + \epsilon_\beta, \\ & \leq & 2\epsilon_2. \end{array}$$

On aboutit à une contradiction lorsque

$$2\epsilon_2 < \epsilon_1 + \epsilon_\beta.$$

Par suite, par hypothèse sur les valeurs propres de H, β est égal à 2 dès que

$$2\epsilon_2 < \epsilon_1 + \epsilon_3 < \epsilon_1 + \epsilon_k \quad \text{pour tout } k \geq 3$$

soit dès que

$$\frac{\epsilon_3 - \epsilon_2}{\epsilon_3 - \epsilon_1} > \frac{1}{2}. \tag{3.57}$$

Il reste à démontrer que la condition (3.57) est optimale. Considèrons la matrice diagonale H de dimension 3 d'éléments diagonaux $\{\epsilon_1, \epsilon_2, \epsilon_3\}$ tels que $\epsilon_3 - \epsilon_2 < \epsilon_2 - \epsilon_1$. On définit (X_0, Y_0) par

$$(X_0, Y_0) = \Big((e_1 + e_3)/\sqrt{2}, (e_1 - e_3)/\sqrt{2}\Big),$$

d'où

$$\big(\lambda^0, \mu^0\big) = \Big(X_0^t H X_0, Y_0^t H Y_0\Big) = \Big((\epsilon_1 + \epsilon_3)/2, (\epsilon_1 + \epsilon_3)/2\Big).$$

Par hypothèse sur les ϵ_i, λ^0 et μ^0 sont strictement inférieurs à ϵ_2. Par suite, on a pour tout $n \in \mathbb{N}$, $(X_n, Y_n) = (\pm X_0, \pm Y_0)$ et l'espace vectoriel engendré par (X_n, Y_n) est celui engendré par (e_1, e_3).

3.6.3 Conclusion

Lorsque les valeurs propres de H sont deux à deux distinctes, la matrice limite de la suite $(D^n)_n$ est un projecteur sur l'espace vectoriel engendré par deux vecteurs propres (c_α, c_β) de H avec . $\alpha = 1$ et $\alpha < \beta$. β est égal à 2 dès que

$$\frac{\epsilon_3 - \epsilon_2}{\epsilon_3 - \epsilon_1} > \frac{1}{2}.$$

Chapitre 4

Vers le traitement des systèmes métalliques

Ce chapitre est consacré à la présentation d'une adaptation des méthodes de projection au traitement des systèmes métalliques. Il reprend un travail effectué avec V. Duwig, G. Bencteux (EDF) et E. Cancès (CERMICS) qui a fait l'objet d'un rapport interne EDF [128] et de quatre interventions orales à des congrès [I1] [I2] [I3] [I4].

On considère dans cette section une base de calcul localisée[1] pour la résolution du problème électronique telle que la matrice de recouvrement S soit égale à I_{N_b}. A chaque itération SCF, il s'agit de calculer le projecteur $D^\star$ solution de

$$\begin{cases} Fc_i = \epsilon_i c_i, & \epsilon_1 \leq \ldots \leq \epsilon_N \leq \epsilon_{N+1} \leq \ldots \leq \epsilon_{N_b}, \\ c_i^t c_j = \delta_{ij}, & \\ D^\star = \displaystyle\sum_{i=1}^{N} c_i c_i^t. & \end{cases} \tag{4.1}$$

où F, matrice de Fock construite à chaque itération SCF, est creuse quelle que soit la nature du système considéré suite au caractère local de la base de calcul. Afin de se placer dans le contexte métallique, on suppose sans perte de généralité

$$-1 = \epsilon_1 \leq \ldots \leq \epsilon_N < 0 < \epsilon_{N+1} \leq \ldots \leq \epsilon_{N_b} \leq 1. \tag{4.2}$$

avec $\epsilon = \epsilon_{N+1} = -\epsilon_N \simeq 0$. Ainsi, *le niveau de Fermi* ϵ_F est égal à 0 et le *gap* γ_F à

$$\frac{\epsilon_{N+1} - \epsilon_N}{\epsilon_{N_b} - \epsilon_1} = \frac{2\epsilon}{\epsilon_{N_b} - \epsilon_1} \simeq 0$$

4.1 Rappel sur les méthodes de projection

Les méthodes dites de projection [88] se basent sur la caractérisation de $D^\star$ suivante

$$D^\star = \mathcal{H}(F) \tag{4.3}$$

où $\mathcal{H}$ est la fonction définie sur $[-1, 1]$ par $\mathcal{H} = \mathbb{1}_{[-1,0]}$. Plus généralement,

$$D = q(F) \text{ pour tout polynôme } q \text{ tel que } \left| \begin{array}{ll} q(\epsilon_i) = 1 & \forall\ 1 \leq i \leq N \\ q(\epsilon_i) = 0 & \forall\ N+1 \leq i \leq N_b \end{array} \right. . \tag{4.4}$$

On est donc ramené à trouver le polynôme de plus bas degré possible approchant $\mathcal{H}$ au sens défini par (4.4). Comme on ne connaît pas *a priori* les réels ϵ_i, on cherche la meilleure approximation de $\mathcal{H}$ au sens de la norme infinie sur $[-1, -\epsilon] \cup [\epsilon, 1]$. Comme $\mathcal{H}$ s'approche d'un échelon d'autant plus que ϵ est faible, on doit recourir à un polynôme q de degré d'autant plus élevé que ϵ et donc γ est faible. Le caractère creux de $D^\star$ est en conséquence d'autant moins prononcé que γ est petit. Les systèmes métalliques se caractérisant par $\gamma \simeq 0$, la matrice $D^\star$ est pleine. L'évaluation de $q(F)$ mène par suite à des produits matriciels pleins et conduit à une méthode de complexité $\mathcal{O}(N^3)$ pour le traitement de tels systèmes.

[1] caractérisée par $N_b = \mathcal{O}(N)$.

4.2 Adaptation aux systèmes métalliques

L'idée naturelle pour traiter les systèmes métalliques consiste à adapter les méthodes de projection de façon à se ramener au traitement d'un isolant. Dans cette optique, on propose de créer artificiellement un *trou* dans le spectre de F autour du niveau de Fermi ici égal à 0. Cette technique est connue sous le terme de *déflation*.

La méthode de déflation se déroule en trois étapes :

- On calcule $2d$ éléments propres $(c_i, \epsilon_i)_{i=N-d+1,N+d}$ de F autour du niveau de Fermi (on a supposé par souci de simplicité qu'on calcule autant d'éléments au-dessus et en dessous du niveau de Fermi). Le calcul des éléments propres (c_i, ϵ_i) correspond à la recherche des $2d$ plus grands éléments propres de F^{-1} (comme $\epsilon_F = 0$) et est réalisé à l'aide de l'algorithme *Implicit Restart Arnoldi* (IRA) [121]. Les méthodes de sous-espace telle IRA utilisent des produits *matrice-vecteur*[2] faisant intervenir F^{-1} et des diagonalisations de matrices de taille proportionnelle à d [111].
- On applique une méthode de projection à la matrice $\bar{F}$ définie par

$$\bar{F} = F - \sum_{i=N-d+1}^{N+d} \epsilon_i c_i c_i^t, \tag{4.5}$$

  et de gap $\gamma_{\bar{F}} = \dfrac{\epsilon_{N+d+1} - \epsilon_{N-d}}{2\epsilon}\gamma_F$ plus grand que le gap de F.
- On construit l'approximation D_q de $D^\star$ par

$$D_q = q(\bar{F}) + \sum_{i=N-d+1}^{N+d} \mathbb{1}_{\epsilon_i<0}\, c_i c_i^t.$$

La calcul théorique de la complexité totale de la méthode n'est pas accessible dans le cas général. Il est réalisable dans les deux cas suivants :

- (i) F est une matrice bande et on ne tronque pas les produits matriciels,
- (ii) F est une matrice creuse générale et on tronque les calculs hors d'un *masque de calcul* prédéfini par un nombre de coefficients n_m constant par ligne et indépendant de N, en supposant que la troncature des calculs ne perturbe pas la convergence de la méthode.

4.2.1 Rappel sur le calcul d'éléments propres

L'algorithme IRA est décrit plus en détail dans cette section. Dans la suite, A désigne la matrice F^{-1}, I_d la matrice identité et $(.,.)$ représente le produit scalaire euclidien dans $\mathbb{R}^{N_b}$. Par ailleurs, on cherche à calculer les p plus grands éléments propres $(u_j, \lambda_j)_{i=j,p}$ de A,

[2] dont chacun est réalisable par une méthode itérative de complexité $\mathcal{O}(N_b)$ comme F est creuse [117].

4.2.1.1 La méthode des puissances itérées

Lorqu'on cherche le plus grand élément propre en valeur absolue (u_1, λ_1) d'une matrice A, la méthode des puissances itérées s'impose comme une méthode de choix. Elle suit le schéma suivant :

- *choisir* v_1 *tel que* $\|v_1\| = 1$,
- *pour* $i = 1$, *jusqu'à convergence de* (v_i, θ_i)
  - $\diamond$ $v = Av_i$,
  - $\diamond$ $\theta_i = v_i^t v$,
  - $\diamond$ $v_{i+1} = v/\|v\|$,
- *fin*

Si $(v_1, u_1) \neq 0$, (v_i, θ_i) converge vers (u_1, λ_1).

4.2.1.2 La méthode des itérations orthogonales

Lorsque $p > 1$, la méthode des puissances itérées se généralise à la méthode des itérations orthogonales :

- *choisir* $V_1 \in \mathcal{M}(N_b, p)$ *orthonormale,*
- *pour* $i = 1$, *jusqu'à convergence de* V_i
  - $\diamond$ $V = AV_i$,
  - $\diamond$ *orthonormalisation de* V *:* $V = QR$,
  - $\diamond$ $V_{i+1} = Q$,
- *fin*

Si $\forall\ 1 \leq j \leq p, \quad V_1^t u_j \neq 0_{\mathbf{R}^p}$, alors V_i converge vers une base orthonormale de l'espace vectoriel engendré par les vecteurs u_j. L'orthonormalisation de V à chaque itération est réalisée par une factorisation QR. Lorsque $p = N_b$, cet algorithme est sous certaines hypothèses équivalent à l'algorithme QR très utilisé pour la diagonalisation complète d'une matrice [119].

4.2.1.3 Les méthodes de sous-espace

A la convergence, la méthode des itérations orthogonales génère le sous-espace engendré par les vecteurs u_i cherchés à l'aide de A. Les méthodes de sous-espace généralisent cette méthode [111] :

- *choisir* $V_1 \in \mathcal{M}(N_b, p)$ *orthonormale, et* P_1 *un polynôme,*
- *pour* $i = 1$, *jusqu'à convergence*
  - $\diamond$ $V = P_i(A)V_i$,
  - $\diamond$ *orthonormalisation de* V *:* $V = V_{i+1}R$,
  - $\diamond$ *calcul de la matrice* $H_{i+1} = V_{i+1}^t A V_{i+1}$,
  - $\diamond$ *calcul des p plus grands éléments propres* $(z_i^j, \mu_i^j)_{j=1,p}$ *de* H_{i+1},
  - $\diamond$ *calcul de* P_{i+1} *à l'aide des* μ_i^j,
- *fin*

Sous la même condition sur V_1 que précédemment, les éléments $(z_i^j, \mu_i^j)_j$ convergent vers les éléments $(u_j, \lambda_j)_j$. Les polynômes P_i, le plus souvent construits à partir des polynômes de Chebyshev [116], permettent d'accélérer la convergence de la méthode vers les modes propres cherchés.

4.2.1.4 La méthode d'Arnoldi

Principe de la méthode

Dans les méthodes de sous-espace, la matrice H_i est reconstruite à chaque itération. La méthode d'Arnoldi remédie à ce problème en utilisant la propriété des espaces de Krylov et en relaxant la dimension de la matrice V_i. Cet algorithme suit le schéma suivant [125] :

- *choisir* v_1 *tel que* $\|v_1\| = 1$,
- *pour* $m = 1$, *jusqu'à stagnation de* $\mathcal{K}^m(A, v_1)$ *(au plus* n *itérations)*
  - ◇ *construction d'une base orthonormale* V_m *de* $\mathcal{K}^m(A, v_1)$,
  - ◇ *construction de* $H_m = V_m^t A V_m$,
  - ◇ *calcul des éléments propres* $(z_m^j, \mu_m^j)_{j=1,p}$ *de* H_m,
- *fin*

où $\mathcal{K}^m(A, v_1)$ désigne l'espace de Krylov relatif à A de dimension m associé à v_1 défini par

$$\mathcal{K}^m(A, v_1) = \text{Vect}(v_1, Av_1, ..., A^{m-1}v_1).$$

Construction de V_m et H_m

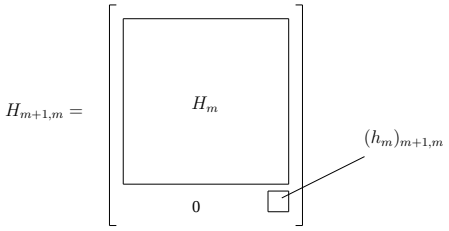
La construction des matrices V_m et H_m est effectuée colonne par colonne suivant le processus itératif d'Arnoldi :

- *choisir* v_1 *tel que* $\|v_1\| = 1$,
- *pour tout* $1 \leq j \leq m-1$ *faire*
  - ◇ $w = Av_j$,
  - ◇ *pour* $i = 1,\ j$
    - ○ $(h_m)_{ij} = (w, v_i)$,
    - ○ $w = w - (h_m)_{ij} v_i$,
  - ◇ *fin*
  - ◇ $(h_m)_{j+1,j} = \|w\|$,
  - ◇ $v_{j+1} = \dfrac{w}{(h_m)_{j+1,j}}$,
- *fin*

Ce procédé s'écrit sous la forme matricielle suivante

$$\forall\ m \text{ tel que } (h_m)_{m+1,m} \neq 0 \quad \left| \begin{array}{lcl} AV_m & = & V_{m+1} H_{m+1,m}, \\ & = & V_m H_m + (h_m)_{m+1,m} v_{m+1} e_m^t, \\ V_m^t A V_m & = & H_m. \end{array} \right. \qquad (4.6)$$

où les vecteurs (e_i) sont les vecteurs de la base canonique, H_m est la matrice des coefficients $(h_m)_{ij}$ et $H_{m+1,m}$ est la matrice de Hessenberg formée à l'aide de H_m et $(h_m)_{m+1,m}$. La matrice V_m vérifie par construction $V_m^t V_m = I_m$.



Propriétés de la méthode

Lorsque A est symétrique, la matrice H_m l'est aussi. On a aussi

$$(h_m)_{m+1,m} = 0 \Longleftrightarrow A\mathcal{K}^m(A, v_1) = \mathcal{K}^m(A, v_1). \tag{4.7}$$

Cette propriété permet de déceler la stagnation de $\mathcal{K}^m(A, v_1)$ et donc d'arrêter l'algorithme. Les éléments (z_m^j, μ_m^j) sont alors les éléments propres (u_j, λ_j) tels que $(v_1, u_j) \neq 0$. Par conséquent, l'algorithme peut converger en un nombre d'itérations très grand par rapport au nombre p de valeurs propres cherchées. Dans cette situation, l'algorithme devient inefficace puisqu'on doit diagonaliser une matrice de grande taille. Par ailleurs, une perte d'orthogonalité de la matrice V_m est constatée lorsque m devient grand.

4.2.1.5 L'algorithme *Implicit Restart Arnoldi* (IRA)

Il est nécessaire le plus souvent de redémarrer la méthode d'Arnoldi pour une valeur seuil de m. Soit v_1^{new} le nouveau vecteur de départ, un redémarrage pertinent se doit :

- d'utiliser l'information donnée par les éléments (z_j, μ_j) pour accélérer la convergence vers les éléments propres cherchés,
- de ne pas recalculer la matrice $H_{p+1,p}$ (on cherche p valeurs propres) par p itérations d'Arnoldi à partir de v_1^{new}.

Algorithme

Une solution élégante à ce problèmeest donnée par l'algorithme IRA [146] :

• *choisir v_1 normé, et m tel que $p+m$ soit la dimension maximale de travail autorisée. Soit $H_{p+1,p}^{(1)}$ la matrice obtenue après p itérations de l'algorithme d'Arnoldi à partir de v_1,*

• *pour $k = 1$, jusqu'à convergence*[3]

[3] c'est-à-dire jusqu'à l'obtention avec la précision souhaitée des p plus petits éléments (u_j, λ_j) tels que $v_1^t u_j \neq 0$.

◇ *obtention de* $H_{p+m}^{(k)}$ *suite à* m *itérations d'Arnoldi supplémentaires à partir de* $H_{p+1,p}^{(k)}$,

◇ *calcul des estimations d'éléments propres* $(z_k^j, \mu_k^j)_{j=1,\, p+m}$ *de* A *en diagonalisant* $H_{p+m}^{(k)}$,

◇ *calcul de* v_{k+1} *à l'aide des* (z_k^j, μ_k^j),

◇ *obtention de* $H_{p+1,p}^{(k+1)}$ *coïncidant avec la matrice qu'on obtiendrait en effectuant* p *itérations d'Arnoldi à partir de* v_{k+1},

• *fin*

Obtention de $H_{p+1,p}^{(k+1)}$

Soient $(\nu_k^i)_{i=1,m}$ les moins bonnes approximations des valeurs propres cherchées parmi les μ_k^j (soient les plus petites valeurs propres pour le cas considéré), v_{k+1} est défini par

$$v_{k+1} = \gamma \prod_{i=1}^{m} (A - \nu_k^i I_d) v_k \tag{4.8}$$

où γ est une constante de normalisation. (4.8) rend possible le calcul de $H_{p+1,p}^{k+1}$ de façon implicite à l'aide de m factorisations QR implicites sans recourir à p itérations d'Arnoldi à partir de v_{k+1}. On détaille ce calcul ci-dessous.

La factorisation QR de H consiste à calculer une matrice Q orthogonale et R triangulaire supérieure telle que $H = QR$. Lorsque H est une matrice de Hessenberg, Q l'est aussi. L'algorithme QR implicite [114] utilise cette propriété et le théorème Q-implicite pour effectuer plus rapidement qu'explicitement la factorisation QR.Pour plus de clarté, on décrit le fonctionnement de l'algorithme IRA en considérant des factorisations QR explicites[4]. Par ailleurs, on omet l'indice k. A la fin de l'itération k et pour tout j, $h_{j+1,j}$ désigne $(h_j)_{j+1,j}$. Il vient par (4.6)

$$AV_{p+m} = V_{p+m} H_{p+m} + h_{p+m+1,p+m} v_{p+m+1} e_{p+m}^t. \tag{4.9}$$

On considère le réel ν^1, la factorisation QR mène à $H_{p+m} - \nu^1 I_d = Q_1 R_1$. Par (4.9), on a

$$(A - \nu^1 I_d) V_{p+m} - V_{p+m}(H_{p+m} - \nu^1 I_d) = h_{p+m+1,p+m} v_{p+m+1} e_{p+m}^t \tag{4.10}$$

soit

$$(A - \nu^1 I_d) V_{p+m} - V_{p+m} Q_1 R_1 = h_{p+m+1,p+m} v_{p+m+1} e_{p+m}^t. \tag{4.11}$$

La multiplication de (4.11) à droite par Q_1 conduit à

$$AV_{p+m}^{(1)} - V_{p+m}^{(1)} H_{p+m}^{(1)} = h_{p+m+1,p+m} v_{p+m+1} e_{p+m}^t Q_1 \tag{4.12}$$

[4]Les deux algorithmes (implicite et explicite) sont équivalents sous des hypothèses peu restrictives qu'on suppose vérifiées.

avec $H^{(1)}_{p+m} = R_1 Q_1 + \nu^1 I_d = Q_1^t H_{p+m} Q_1$ et $V^{(1)}_{p+m} = V_{p+m} Q_1$. Comme Q_1 est une matrice orthogonale, $V^{(1)}_{p+m}$ constitue une famille de colonnes orthonormales. Par ailleurs comme Q_1 est une matrice de Hessenberg, $H^{(1)}_{p+m}$ l'est aussi et

$$h_{p+m+1,p+m} e^t_{p+m} Q_1 = \beta e^t_{p+m-1} + \sum_{i=p+m}^{p+m+1} \alpha_i e_i^t.$$

L'itération de ce processus pour les m réels ν^i définis après le calcul des μ^j est directe et s'écrit

$$A V^{(m)}_{p+m} - V^{(m)}_{p+m} H^{(m)}_{p+m} = h_{p+m+1,p+m} v_{p+m+1} e^t_{p+m} Q \tag{4.13}$$

avec $H^{(m)}_{p+m} = Q^t H_{p+m} Q$, $V^{(m)}_{p+m} = V_{p+m} Q$ et $Q = Q_1 Q_2 \dots Q_m$ où les matrices Q_i sont déterminées par les factorisations QR successives. Par ailleurs

$$h_{p+m+1,p+m} e^t_{p+m} Q = \beta e^t_p + \sum_{i=p+1}^{p+m+1} \alpha_i e_i^t. \tag{4.14}$$

On désigne par $\tilde{V}_p$ la matrice formée des p premières colonnes de $V^{(m)}_{p+m}$. Par orthonormalité des Q_i $\tilde{V}_p^t \tilde{V}_p = I_p$, et, il vient par (4.13) et (4.14)

$$\begin{aligned} A\tilde{V}_p &= \tilde{V}_p H^{(m)}_p + h^{(m)}_{p+1,p} v^{(m)}_{p+1} e^t_p + \beta v_{p+m+1} e^t_p, \\ &= \tilde{V}_p \tilde{H}_p + \tilde{h}_{p+1,p} \tilde{v}_{p+1} e^t_p, \end{aligned} \tag{4.15}$$

où $\tilde{H}_p = H^{(m)}_p$ et $\tilde{v}_{p+1} = \dfrac{h^{(m)}_{p+1,p} v^{(m)}_{p+1} + \beta v_{p+m+1}}{\| h^{(m)}_{p+1,p} v^{(m)}_{p+1} + \beta v_{p+m+1} \|} = \dfrac{h^{(m)}_{p+1,p} v^{(m)}_{p+1} + \beta v_{p+m+1}}{\tilde{h}_{p+1,p}}$.

$\tilde{v}_{p+1}$ est normé et orthogonal à $\tilde{V}_p$ par orthonormalité de la matrice Q et des colonnes de V_{p+m+1}. (4.15) est exactement le résultat de p itérations d'Arnoldi à partir du vecteur $\tilde{v}_1$ associé à la première colonne de $\tilde{V}_p$.

4.2.2 Calcul de la matrice densité

Comme la matrice $\bar{F}$ n'est pas creuse, on ne peut pas réaliser le calcul $q(\bar{F})$, et donc celui de D_q, directement. Ce calcul se ramène au calcul de $q(F)$. En effet, par définition

$$D_q = q(\bar{F}) + \sum_{i=N-d+1}^{N+d} \mathbb{1}_{\epsilon_i < 0}\ c_i c_i^t$$

soit, par définition de $\bar{F}$,

$$D_q = q\Big(F - \sum_{i=N-d+1}^{N+d} \epsilon_i c_i c_i^t\Big) + \sum_{i=N-d+1}^{N+d} \mathbb{1}_{\epsilon_i < 0}\ c_i c_i^t.$$

F étant symétrique, les éléments propres de F sont orthonormaux et donc pour tout $k \in \mathbb{N}$,

$$\left(F - \sum_{i=N-d+1}^{N+d} \epsilon_i c_i c_i^t\right)^k = F^k - \sum_{i=N-d+1}^{N+d} \epsilon_i^k c_i c_i^t. \tag{4.16}$$

Il vient alors

$$D_q = q(F) - \sum_{i=N-d+1}^{N+d} q(\epsilon_i) c_i c_i^t + \sum_{i=N-d+1}^{N+d} \mathbb{1}_{\epsilon_i<0}\, c_i c_i^t,$$

soit au final

$$D_q = q(F) + \sum_{i=N-d+1}^{N+d} \big(\mathbb{1}_{\epsilon_i<0} - q(\epsilon_i)\big) c_i c_i^t.$$

On approche donc la matrice $D^\star$ par la somme d'une matrice creuse $q(F)$ (si q est de suffisamment bas degré) et d'une matrice pleine représentée par une somme de corrections de rang 1 correspondant aux vecteurs c_i déflatés. D_q n'est jamais formée entièrement car seuls les coefficients $(D_q)_{ij}$ tels que $F_{ij} \neq 0$ suffisent pour accéder à l'énergie et aux forces électroniques.

4.2.3 Complexité théorique de la méthode

L'algorithme a une complexité théorique totale $\mathcal{C}(d, N)$ qui dépend du nombre d'éléments propres $2d$ calculés et du nombre total N d'éléments propres caractérisant $D^\star$ (on rappelle que $N_b = \mathcal{O}(N)$). La complexité optimale de l'algorithme s'obtient pour la valeur $d^\star(N)$ qui minimise $\mathcal{C}(d, N)$.

4.2.3.1 Calcul des éléments propres

Soit l la dimension maximale de travail autorisé dans l'algorithme IRA, une itération de cet algorithme demande

- l résolutions de système linéaire chacune de complexité $\mathcal{O}(N)$ comme F est creuse et $N_b = \mathcal{O}(N)$,
- la construction de la matrice $H_{l,l}$ tridiagonale (comme F, et donc F^{-1}, sont symétriques) sachant que chaque coefficient est un produit scalaire de complexité $N_b = \mathcal{O}(N)$,
- $l - 2d$ factorisations QR, chacune de complexité $\mathcal{O}(l^2)$ comme $H_{l,l}$ est tridiagonale.

L'algorithme IRA est utilisé le plus souvent avec $l = \mathcal{O}(p)$ où p est le nombre d'éléments propres cherchés. Suite à des observations numériques, on constate que le nombre d'itérations de l'algorithme IRA (quelques itérations en pratique) est indépendant de p. Sous cette hypothèse, la complexité totale du calcul des $2d$ éléments propres de F autour du niveau de Fermi est $\mathcal{O}(dN + d^3)$.

4.2.3.2 Calcul de la matrice densité

On peut estimer le calcul de $q(F)$ de façon théorique dans les cas (i-ii) introduits en début de section 4.2. Des calculs simples conduisent à

- une complexité $\mathcal{O}(NM^2)$ quand F est une matrice bande,
- une complexité $\mathcal{O}(n_m^2 NM)$ quand F est creuse et qu'on tronque les calculs.

4.2.3.3 Complexité totale

On considère dans cette section l'algorithme de projection *Fermi Operator Expansion* [58] afin d'utiliser les propriétés des polynômes de Chebyshev. Afin de déduire la valeur $d^\star$, il est indispensable d'obtenir une relation entre M, d et N. Pour cela, il est nécessaire de connaître

- la répartition des valeurs propres autour de l'énergie de Fermi, ce qui conduit à une relation entre d et le *gap* γ_d de la matrice $\bar{F}$,
- l'erreur $e_d = \sup_{|x|\geq\gamma_d/2} |\mathcal{H} - q|$ à atteindre pour satisfaire le critère d'erreur (sur l'énergie ou la matrice densité) qu'on se donne avec la précision ϵ à γ_d donné,
- le degré M du polynôme q à appliquer pour satisfaire $\sup_{|x|\geq\gamma_d/2} |\mathcal{H} - q_M| < e_d$ à γ_d donné.

En premier lieu, on fait l'hypothèse d'une répartition linéaire des valeurs propres autour du niveau de Fermi. On déduit par conséquent

$$\gamma_d = \mathcal{O}\Big(\frac{d}{N}\Big). \tag{4.17}$$

Dans un second temps, en notant $E_{ex} = \mathrm{Tr}(FD^\star)$ et $E_{ap} = \mathrm{Tr}(FD_q)$, on obtient par un calcul direct

$$\Big|\frac{E_{ex} - E_{ap}}{E_{ex}}\Big| \leq \epsilon \quad \text{si} \quad \frac{2e_d(N_b - 2d)}{\gamma_d(N-d)} \leq \epsilon. \tag{4.18}$$

Enfin, l'obtention d'une précision e_d satisfaisant (4.18) est donnée dans [16, 36] par

$$M = \mathcal{O}\Big(- log^2(e_d)\gamma_d^{-1}\Big). \tag{4.19}$$

L'association de (4.17), (4.18) et (4.19) conduit finalement dans le cas (i) à la complexité totale $\mathcal{C}(d, N)$

$$\mathcal{C}(d, N) = \mathcal{O}\left(\frac{N^3}{d^2} + Nd + d^3\right). \tag{4.20}$$

L'égalisation de chacun des termes dans (4.20) mène au choix optimal $d^\star = \mathcal{O}(N^{3/5})$ et par suite à une complexité théorique $\mathcal{C} = \mathcal{O}(N^{9/5})$ de la méthode de déflation. Un raisonnement similaire dans le cas (ii) mène au choix optimal $d^\star = \mathcal{O}(N^{1/3})$ et à la complexité théorique $\mathcal{C} = \mathcal{O}(N^{4/3})$. On rappelle que l'obtention de cette

complexité se base sur l'hypothèse que la troncature ne perturbe pas la convergence de la méthode.

On note enfin que l'utilisation de la méthode FOE comme méthode de projection n'est pas obligatoire car on peut adapter aussi la méthode de purification de la matrice densité [90] ainsi que la méthode *Density Matrix Minimization* (DMM) [81]. Toutefois, il est impossible de calculer une complexité théorique dans ces cas car le polynôme q associé ne provient pas d'un principe variationnel.

4.2.4 Description de l'algorithme pour la méthode DMM

Il est direct d'adapter l'algorithme pour des méthodes de projection basées sur l'application d'un polynôme prédéfini à une matrice bien choisie telle la méthode de purification de la matrice densité. Dans la méthode DMM, le polynôme appliqué est construit à chaque itération suite à la minimisation suivant la direction de descente de la fonctionnelle $\Omega(F,D) = \mathrm{Tr}\big(F(3D^2 - 2D^3)\big)$. On peut toutefois adapter cette méthode dans le cas où l'initial guess est une matrice qui commute avec F, typiquement un polynôme de F. L'algorithme se déroule alors comme suit.

Soit $D_0 = p_0(F)$ l'initial guess de l'algorithme, on introduit $\bar{D}_0$ et $\bar{F}$ par

$$\begin{aligned} \bar{F} &= F - \sum_{i=N-d+1}^{N+d} \epsilon_i c_i c_i^t, \\ \bar{D}_0 &= D_0 - \sum_{i=N-d+1}^{N+d} p_0(\epsilon_i) c_i c_i^t, \end{aligned} \tag{4.21}$$

on applique l'algorithme DMM ensuite au couple $(\bar{D}_0, \bar{F})$. A chaque itération, le calcul de $\bar{D}_n$ revient à calculer une matrice D_n et $2d$ coefficients $(p_n^i)_{i=N-d+1,N+d}$ tels que

$$\bar{D}_n = D_n - \sum_{i=N-d+1}^{N+d} p_n^i c_i c_i^t. \tag{4.22}$$

Le calcul de (D_n,p_n) s'effectue suivant les étapes suivantes :

– calcul du gradient $\bar{G}_n$ de $\Omega(\bar{F},\bar{D})$ en $\bar{D}_{n-1}$:

$$\begin{aligned} \bar{G}_n &= \Big(\nabla_{\bar{D}}\Omega(\bar{F},\bar{D})\Big)\big(\bar{D}_{n-1}\big), \\ &= 6\bar{F}\bar{D}_{n-1}\big(1-\bar{D}_{n-1}\big) \end{aligned}$$

par définition de $\Omega(F,D)$. Par (4.21) et (4.22), on obtient

$$\begin{aligned} \bar{G}_n &= 6FD_{n-1}\big(1-D_{n-1}\big) + \sum_{i=N-d+1}^{N+d} g_n^i c_i c_i^t, \\ &= \Big(\nabla_D\Omega(F,D)\Big)\big(D_{n-1}\big) + \sum_{i=N-d+1}^{N+d} g_n^i c_i c_i^t \end{aligned}$$

avec pour tout $N-d+1 \leq i \leq N+d$, $g_n^i = -6\epsilon_i p_{n-1}^i(1-p_{n-1}^i)$.

– calcul de la direction de descente $\bar{X}_n$:

$$\begin{aligned} \bar{X}_n &= -\bar{G}_n + \beta_n^{CG}\bar{X}_{n-1}, \\ &= X_n + \sum_{i=N-d+1}^{N+d} x_n^i c_i c_i^t \end{aligned}$$

avec pour tout $N-d+1 \leq i \leq N+d$, $x_n^i = -g_n^i + \beta_n^{CG} x_{n-1}^i$.

– calcul du pas de descente α_n :

$$\begin{aligned} \alpha_n &= \text{arginf}_{loc}\Big\{\Omega\big(\bar{F}, \bar{D}_{n-1} + \alpha\bar{X}_n\big),\ \alpha \in \mathbb{R}\Big\}, \\ &= \text{arginf}_{loc}\left\{\Omega\big(F, D_{n-1} + \alpha X_n\big) + \sum_{i=N-d+1}^{N+d} \epsilon_i\Big(3f_{n,i}^2(\alpha) - 2f_{n,i}^3(\alpha)\Big),\ \alpha \in \mathbb{R}\right\} \end{aligned}$$

avec pour tout $N-d+1 \leq i \leq N+d$, $f_{n,i}(\alpha) = -p_{n-1}^i + \alpha x_n^i$.

– calcul de $\bar{D}_n$:

$$\begin{aligned} \bar{D}_n &= \bar{D}_{n-1} + \alpha_n\bar{X}_n, \\ &= D_n - \sum_{i=N-d+1}^{N+d} p_n^i c_i c_i^t. \end{aligned}$$

avec $D_n = D_{n-1} + \alpha_n X_n$, et pour tout $N-d+1 \leq i \leq N+d$, $p_n^i = p_{n-1}^i - \alpha_n x_n^i$.

A chaque itération, on construit ensuite l'approximation D_q de $D^\star$ par

$$\begin{aligned} D_q &= \bar{D}_n + \sum_{i=N-d+1}^{N+d} \mathbb{1}_{\epsilon_i<0}\ c_i c_i^t, \\ D_q &= D_n + \sum_{i=N-d+1}^{N+d} \left(\mathbb{1}_{\epsilon_i<0} - p_n^i\right) c_i c_i^t. \end{aligned}$$

La prise en compte de la déflation des éléments c_i conduit à introduire la suite p_n. Sa mise à jour nécessite des modifications légères de chaque étape de DMM.

4.2.5 Résultats préliminaires

Des résultats préliminaires ont été obtenus sur un problème modèle, représentatif de problèmes obtenus à partir de modèles semi-empiriques très simples ($S = I_{N_b}$ et

F tridiagonale)). La matrice F est la suivante

$$F = \begin{bmatrix} \alpha & \beta & 0 & \cdots & 0 \\ \beta & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \beta \\ 0 & \cdots & 0 & \beta & \alpha \end{bmatrix}, \; N_b = 3N, \; (\alpha, \beta) = \Big(2, -1\Big). \tag{4.23}$$

Les hypothèses (4.2) et sur la répartition linéaire des valeurs propres autour du niveau de Fermi sont satisfaites. Par ailleurs $\gamma_{N_b} = 1/N_b$, ce qui assure les propriétés métalliques de ce modèle lorsque N_b est grand.
Pour chaque calcul, on introduit un paramètre ϵ_a qui conduit à l'arrêt de l'algorithme dès que $\epsilon_D < \epsilon_a$ avec $\epsilon_D = \max_{i,j} |\Delta D_{ij}|$ et

$$\begin{aligned} (\Delta D)_{ij} &= 2\frac{D^\star_{ij} - (D_q)_{ij}}{|D^\star_{ij}| + |(D_q)_{ij}|} \quad \text{si } F_{ij} \neq 0, \\ &= 0 \qquad\qquad\qquad\qquad \text{sinon.} \end{aligned}$$

Sur le plan numérique, l'étape de déflation est réalisée à l'aide de la routine *dsaupd.f* disponible dans la librairie ARPACK[5]. Les routines mises en œuvre pour la seconde étape utilisent une bibliothèque, implémentée par nos soins, qui met à profit le caractère creux et symétrique des matrices calculées tout au long du déroulement de l'algorithme. Au contraire des routines utilisées pour la déflation, ces routines n'utilisent pas les routines BLAS[6] optimisées pour chaque type de machines, et par conséquent sont moins optimisées.

Pour chaque valeur de N, la valeur optimale $d^\star(N)$ est obtenue en faisant tourner l'algorithme pour plusieurs valeurs de d. Suite au caractère optimisé des routines pour la déflation, on a considéré deux critères pour l'obtention de $d^\star$:

- critère C_1 : en minimisant le nombre total d'opérations n_{op},
- critère C_2 : en minimisant le temps de calcul global t_{CPU}.

Dans un premier temps, on a fait tourner l'algorithme pour des valeurs de N_b entre 10^3 et 10^4 avec $\epsilon_a = 10^{-3}$ sans réaliser de troncature des calculs. De plus grandes valeurs de N_b n'ont pas été considérées suite à une allocation mémoire trop importante. La Figure 4.1 représente en échelle log-log l'évolution du nombre total d'opérations n^*_{op} en milliard d'opérations, ou du temps de calcul t^*_{CPU} en seconde selon le critère considéré.
On obtient pour DMM des complexités numériques équivalentes pour les deux critères C_1 et C_2. C'est un peu moins net pour FOE. Les complexités numériques obtenues sont $\mathcal{O}(N^2)$, respectivement $\mathcal{O}(N^{2.4})$ pour FOE, respectivement DMM. La sur-restimation de la complexité pour FOE[7] s'explique par les faibles valeurs de N_b considérées et l'implantation de FOE mal adaptée pour une matrice pleine.

[5] http ://www.caam.rice.edu/software/ARPACK/index.html.
[6] http ://www.netlib.org/BLAS/index.html.
[7] en $\mathcal{O}(N^{9/5})$ d'un point du vue théorique.

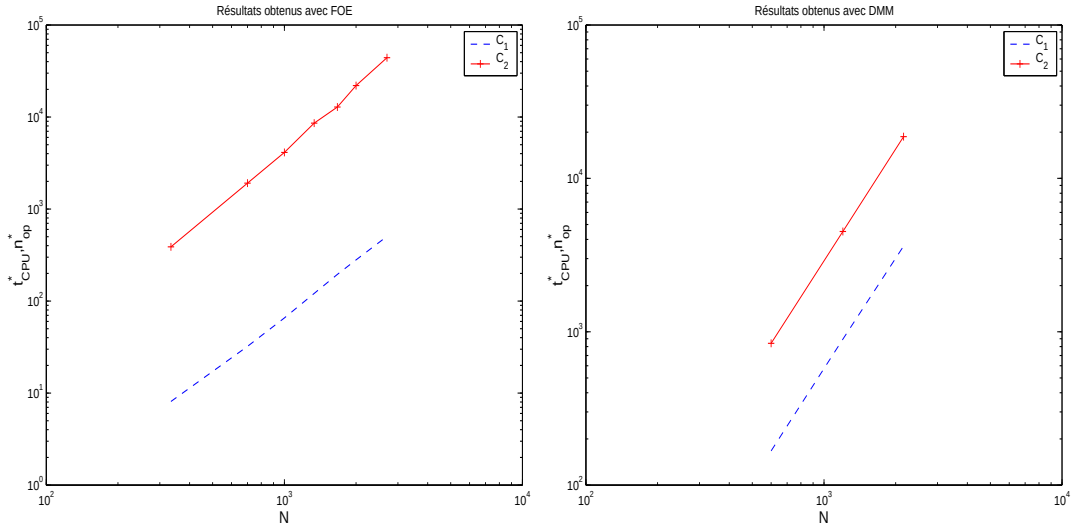


FIG. 4.1 – Complexité de la méthode sans troncature des calculs

Dans un second temps, on a fait tourner l'algorithme pour des valeurs de N_b entre 10^4 et 4.10^4 avec $\epsilon_a = 10^{-2}$ en tronquant les calculs hors du masque de calcul constitué par la diagonale et les 200 premières sur-diagonales et sous-diagonales. La Figure 4.2 représente en échelle log-log l'évolution du nombre total d'opérations n^*_{op} en milliard d'opérations, ou du temps de calcul t^*_{CPU} en seconde selon le critère considéré.

On obtient pour les deux méthodes de projection des complexités numériques équivalentes pour les deux critères C_1 et C_2. Les complexités numériques obtenues sont $\mathcal{O}(N^{1.3})$, respectivement $\mathcal{O}(N^{1.6})$ pour FOE, respectivement DMM. On retrouve la complexité théorique $\mathcal{O}(N^{4/3})$ pour FOE. Pour la précision considérée, la troncature des calculs ne perturbe pas la complexité de la méthode. On a observé toutefois qu'on ne peut pas exiger une très grande précision de l'algorithme suite à la troncature des calculs. En effet, une trop grande précision conduit à déflater un trop grand nombre de valeurs propres, et donc à dégrader la complexité de la méthode.

Par ailleurs, la complexité de DMM est plus importante que la complexité de FOE avec et sans troncature des calculs. Une explication peut résider dans l'optimalité des polynômes de Chebyshev qui conduit à construire des polynômes de degré plus élevé dans DMM pour une même précision. On a note par ailleurs que l'étape de déflation constitue entre 10% et 20% du temps de calcul, ou du nombre d'opérations, selon le critère choisi. Enfin 'a N fixé, on a constaté que l'optimum $d^\star$ évolue dans un intervalle assez large.

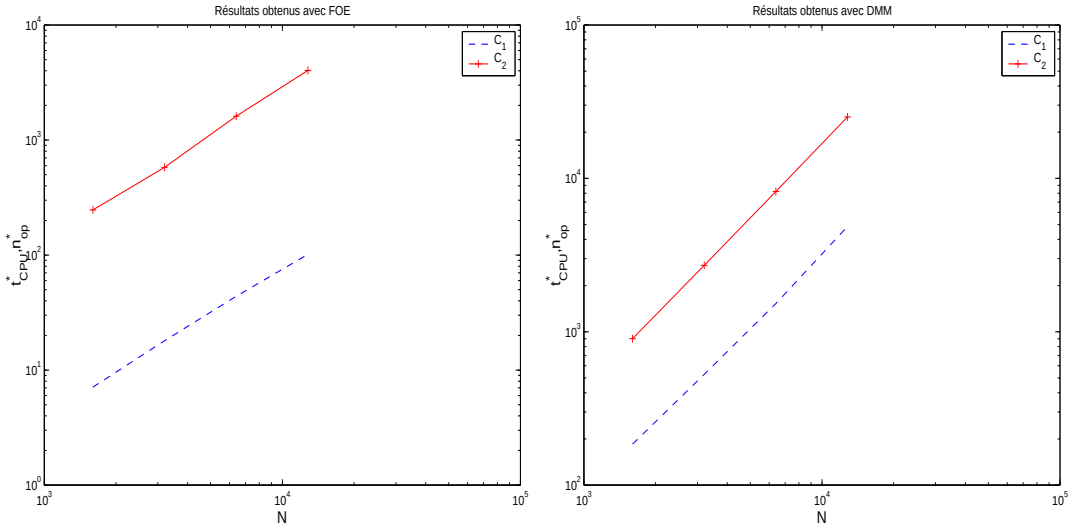


FIG. 4.2 – Complexité de la méthode avec troncature des calculs

4.3 Généralisation au cas $S \neq I_{N_b}$

On décrit dans cette section la généralisation de la méthode utilisant FOE pour une matrice creuse $S \neq I_{N_b}$. Le problème (4.1) devient

$$\begin{cases} Fc_i = \epsilon_i S c_i, & \epsilon_1 \leq \ldots \leq \epsilon_N \leq \epsilon_{N+1} \leq \ldots \leq \epsilon_{N_b}, \\ c_i^t S c_j = \delta_{ij}, \\ D^\star = \displaystyle\sum_{i=1}^{N} c_i c_i^t. \end{cases} \tag{4.24}$$

Le calcul de D_q ne peut plus s'effectuer comme à la section 4.2 car

$$F = \sum_{i=1}^{N_b} \epsilon_i S c_i c_i^t S,$$

et donc $q(F) \neq \displaystyle\sum_{i=1}^{N_b} p(\epsilon_i) c_i c_i^t$. En revanche, on a

$$q(S^{-1}F) = q\left(\sum_{i=1}^{N_b} \epsilon_i c_i c_i^t S \right) = \sum_{i=1}^{N_b} q(\epsilon_i) c_i c_i^t S,$$

d'où $D^\star = q\big(S^{-1}F\big)S^{-1}$ pour tout polynôme q tel que

$$\begin{cases} q(\epsilon_i) = 1 & \text{pour tout } 1 \leq i \leq N \\ q(\epsilon_i) = 0 & \text{pour tout } N+1 \leq i \leq N_b \end{cases}.$$

Soient $(c_i, \epsilon_i)_{i=N-d+1,N+d}$ les $2d$ éléments propres de F déflatés autour du niveau de Fermi à l'aide de l'algorithme IRA[8], on construit $\bar{F}$ par

$$\bar{F} = F - \sum_{i=N-d+1}^{N+d} \epsilon_i S c_i c_i^t S.$$

L'approximation D_q de $D^\star$ s'obtient ensuite par

$$D_q = q(S^{-1}\bar{F})S^{-1} + \sum_{i=N-d+1}^{N+d} \mathbb{1}_{\epsilon_i<0}\, c_i c_i^t.$$

De la même façon que dans le cas $S = I_{N_b}$, D_q s'exprime comme l'addition du terme $q(S^{-1}F)S^{-1}$ et de corrections de rang 1. Par S-orthogonalité des c_i, il vient

$$q(S^{-1}\bar{F}) = q(S^{-1}F) - \sum_{i=N-d+1}^{N+d} q(\epsilon_i) c_i c_i^t S.$$

Par définition de D_q, on obtient ensuite

$$D_q = q(S^{-1}F)S^{-1} - \sum_{i=N-d+1}^{N+d} q(\epsilon_i) c_i c_i^t + \sum_{i=N-d+1}^{N+d} \mathbb{1}_{\epsilon_i<0}\, c_i c_i^t,$$

soit au final

$$\begin{aligned} D_q &= q(S^{-1}F)S^{-1} + \sum_{i=N-d+1}^{N+d} \big(\mathbb{1}_{\epsilon_i<0} - q(\epsilon_i)\big) c_i c_i^t, \\ &= \tilde{D}_q + \sum_{i=N-d+1}^{N+d} \big(\mathbb{1}_{\epsilon_i<0} - q(\epsilon_i)\big) c_i c_i^t. \end{aligned}$$

avec $\tilde{D}_q$ tel que $\tilde{D}_q S = q(S^{-1}F)$. Le calcul de $S^{-1}F$ s'obtient par la résolution de $S(S^{-1}F) = F$. Il est affirmé dans la communauté des chimistes que $S^{-1}F$ est creuse [56]. Pour des polynômes q de faible degré, on peut supposer de la même façon que $\tilde{D}_q$ est creuse, et donc aboutir à une réduction de la complexité du calcul de $D^\star$. Ces deux hypothèses restent à vérifier en pratique. Soulignons que le calcul de $\tilde{D}_q$ n'est pas nécessaire pour le calcul de l'énergie électronique $E = \mathrm{Tr}(FD_q) = \mathrm{Tr}(S^{-1}FD_qS)$. Enfin, la méthode utilisant DMM se généralise en considérant la fonctionnelle généralisée $\Omega(F, D) = \mathrm{Tr}(3DSD - 2DSDSD)$ dès lors que l'initial guess D_0 vérifie $D_0SF = FSD_0$.

[8] l'algorithme IRA utilise des produits *matrice-vecteur* faisant intervenir $F^{-1}S$ dont chacun est réalisable par une méthode itérative de complexité $\mathcal{O}(N_b)$ comme F et S sont creuses.

4.3.1 Conclusions

Les résultats obtenus laisse entrevoir une réduction de la complexité des méthodes de projection pour les systèmes métalliques. Toutefois, quelques points restent à approfondir :

- il est nécessaire de coupler cette méthode avec une stratégie automatique pour le calcul de $d^\star$;
- il semble incoutournable de développer une stratégie de troncature des calculs plus adaptée, par exemple en faisant évoluer dynamiquement le masque de calcul ;
- l'utilisation de la méthode DMM n'est possible que pour des initial guess qui commutent avec F, ce qui n'est pas le cas de la matrice densité générée à l'itération SCF précédente ;
- la généralisation de cette méthode à $S \neq I_{N_b}$ demande le calcul de S^{-1} (approché tout du moins) lorsqu'on utilise la méthode FOE.

Chapitre 5

La méthode des bases réduites pour le problème électronique

La première partie de ce chapitre résume le travail effectué en collaboration avec E. Cancès (CERMICS), G. Turinici (CERMICS), C. Le Bris (CERMICS), Y. Maday (Université Paris VI), N.C. Nguyen (MIT) and T. Patera (MIT) sur l'application de la méthode des bases réduites au problème électronique dont on a relevé les difficultés dans le chapitre 1. Dans un premier temps, on introduit la méthode des bases réduites sur un exemple simple. Dans un second temps, on décrit la démarche et on présente les résultats obtenus avec MATLAB[1] sur des systèmes école.
Dans la seconde partie de ce chapitre, on reproduit une note CRAS publiée en 2004 [A3] dans laquelle une adaptation de cette méthode est proposée lorsque l'équation aux dérivées partielles à résoudre est non affine en les paramètres.

[1] http ://www.mathworks.com/.

5.1 Principe général

Pour certaines simulations, telles les simulations de dynamique moléculaire, on est amené à résoudre un problème mathématique $(\mathcal{P}_\mu)$ de dimension n_h très grande pour *un grand nombre* L de jeux de paramètres (géométriques, numériques) $(\mu_j)_{j=1,L}$. La dimension n_h est telle que la résolution de $(\mathcal{P}_\mu)$ pour tous les μ_j n'est pas réalisable.
Dans beaucoup de situations, l'espace vectoriel engendré par les solutions u_{μ_j} de $(\mathcal{P}_{\mu_j})$ peut être approché avec une précision suffisante par le sous-espace vectoriel engendré par m solutions u_{μ_k} correspondant à m configurations $(\mu_k)_{k=1,m}$ *bien choisies* avec $m \ll n_h$.

5.1.1 La méthode des bases réduites

La méthode des bases réduites se place dans un tel contexte. Afin de mettre à profit la propriété de faible variation des solutions u_{μ_j} à calculer, on définit un problème approché $(\tilde{\mathcal{P}}_\mu)$ de $(\mathcal{P}_\mu)$, qui consiste à chercher la meilleure approximation $\tilde{u}_\mu$ de u_μ dans l'espace vectoriel engendré par m solutions u_{μ_k}. La résolution de $(\tilde{\mathcal{P}}_\mu)$ revient donc à chercher m réels $(\alpha_{\mu,k})_{1\leq k\leq m}$ telle que $\tilde{u}_\mu$ définie par

$$\tilde{u}_\mu = \sum_{k=1}^{m} \alpha_{\mu,k} u_{\mu_k}$$

approche le mieux possible u_μ.

Une telle approche suppose de connaître *a priori* pour toute valeur de m, les configurations de référence μ_k conduisant à la meilleure approximation en un certain sens $(L^2,\ H^1)$ des solutions u_{μ_j}. Ceci est possible pour l'instant dans quelques cas académiques assez éloignés des problèmes rencontrés usuellement. La méthode des bases réduites doit donc s'accompagner d'*estimateurs a posteriori* de façon à estimer l'erreur commise par l'approximation $\tilde{u}_\mu$. Le calcul des solutions u_{μ_j} s'effectue ensuite de la façon suivante

- calcul de u_{μ_1} solution de $(\mathcal{P}_{\mu_1})$ et $\mathcal{B} = \{u_{\mu_1}\}$,
- pour tout $j = 2,\ L$
  1. calcul de $\tilde{u}_{\mu_j}$,
  2. calcul de l'erreur commise ε entre $\tilde{u}_{\mu_j}$ et u_{μ_j} à l'aide d'estimateurs *a posteriori* : si ε n'est pas acceptable, on calcule u_{μ_j} et $\mathcal{B} = \mathcal{B} \cup \{u_{\mu_j}\}$.

Le développement d'*estimateurs a posteriori* précis et peu coûteux est difficile. Bien que plus simple, la définition du problème $(\tilde{\mathcal{P}}_\mu)$ peut s'avérer difficile (comme on le verra pour le problème électronique). On propose de décrire ce passage sur un problème présentant aucune difficulté.

5.1.2 Illustration sur un exemple simple

5.1.2.1 Le problème initial

Soit μ un paramètre réel positif, on considère le problème continu

$$\left| \begin{array}{rcll} -\triangle u + \mu u & = & f & \text{dans } \Omega, \\ u & = & 0 & \text{sur } \partial\Omega. \end{array} \right. \tag{5.1}$$

où Ω désigne un ouvert suffisamment régulier de $\mathbb{R}^3$ et $f \in L^2(\Omega)$. On note $(\chi_i)_{i=1,n_h}$ la base de Galerkin de dimension n_h choisie qu'on suppose indépendante de μ et vérifiant les conditions de Dirichlet du problème. $\mathcal{P}_\mu$ s'écrit

$$\left| \begin{array}{rcl} \displaystyle\int_\Omega \nabla u_\mu . \nabla \chi_i \, d\Omega + \mu \int_\Omega u_\mu \chi_i \, d\Omega & = & \displaystyle\int_\Omega f \chi_i \, d\Omega \quad \forall\, 1 \leq i \leq n_h, \\ u_\mu & = & \displaystyle\sum_{i=1}^{n_h} U_{\mu,i} \chi_i. \end{array} \right. \tag{5.2}$$

Le vecteur U_μ s'obtient en résolvant

$$(A + \mu M) U_\mu = F \tag{5.3}$$

avec

$$\left\{ \begin{array}{rcl} A_{ij} & = & \displaystyle\int_\Omega \nabla \chi_i . \nabla \chi_j \, d\Omega, \\ M_{ij} & = & \displaystyle\int_\Omega \chi_i \chi_j \, d\Omega, \\ F_i & = & \displaystyle\int_\Omega f \chi_i \, d\Omega. \end{array} \right. \tag{5.4}$$

Le complexité de la résolution de (5.3), et donc de (5.2), est $\mathcal{O}(n_h^\gamma)$ où γ est égal à 1 ou 3 selon le caractère creux de A et M. Cette complexité est dépendante de n_h et par suite peut mener à un calcul irréalisable, ou pour le moins très long, en terme de temps de calcul.

5.1.2.2 Le problème base réduite

Donnons nous m solutions $(u_{\mu_k})_{k=1,m}$ et cherchons la solution de $(\mathcal{P}_\mu)$ pour un nouveau μ. Sur la base des u_{μ_k}, le problème $(\tilde{\mathcal{P}}_\mu)$ s'écrit

$$\left| \begin{array}{rcl} \displaystyle\int_\Omega \nabla \tilde{u}_\mu . \nabla u_{\mu_k} + \mu \int_\Omega \tilde{u}_\mu u_{\mu_k} & = & \displaystyle\int_\Omega f u_{\mu_k} \quad \forall\, 1 \leq k \leq m, \\ \tilde{u}_\mu & = & \displaystyle\sum_{k=1}^{m} \alpha_{\mu,k} u_{\mu_k}. \end{array} \right. \tag{5.5}$$

Le vecteur α_μ s'obtient en résolvant

$$(\tilde{A} + \mu \tilde{M}) \alpha_\mu = \tilde{F} \tag{5.6}$$

avec pour tout $1 \leq k, l \leq m$

$$\left\{ \begin{array}{rcl} \tilde{A}_{kl} & = & \displaystyle\int_\Omega \nabla u_{\mu_k}.\nabla u_{\mu_l}\, d\Omega, \\ \tilde{M}_{kl} & = & \displaystyle\int_\Omega u_{\mu_k} u_{\mu_l}\, d\Omega, \\ \tilde{F}_k & = & \displaystyle\int_\Omega f u_{\mu_k}\, d\Omega. \end{array} \right. \tag{5.7}$$

Soit V la matrice de $\mathcal{M}(n_h, m)$ formée par les vecteurs U_{μ_k}, on a $\tilde{A} = V^t A V$, $\tilde{M} = V^t M V$ et $\tilde{F} = V^t F$. Le précalcul de ces matrices, de complexité[2] $\mathcal{O}(m^\gamma n_h^\eta)$ est effectué à la suite du calcul des u_{μ_k}. Ensuite pour chaque μ, la complexité de la résolution de (5.6), et donc de (5.5), est indépendante de n_h. Comme les matrices $\tilde{A}$ et $\tilde{M}$ sont pleines indépendamment du caractère creux de A et M, cette complexité est $\mathcal{O}(m^3)$ ($\ll \mathcal{O}(n_h^\gamma)$ lorsque $m \ll n_h$).

5.1.2.3 Quelques difficultés

Afin d'introduire la seconde partie de cette section, on considère les problèmes non linéaires $(\mathcal{P}_\mu^1)$ et $(\mathcal{P}_\mu^2)$ définis par

$$(\mathcal{P}_\mu^1) \left| \begin{array}{rcll} -\triangle u + g(\mu, x)u & = & f & \text{dans } \Omega, \\ u & = & 0 & \text{sur } \partial\Omega. \end{array} \right.$$

avec $g(\mu, x) \neq f(\mu)h(x)$ strictement positive sur $\mathbb{R}$ pour tout μ, et,

$$(\mathcal{P}_\mu^2) \left| \begin{array}{rcll} -\triangle u + g(\mu, u) & = & f & \text{dans } \Omega, \\ u & = & 0 & \text{sur } \partial\Omega. \end{array} \right.$$

avec g infiniment dérivable et strictement positive sur $\mathbb{R}$.

Le problème $(\mathcal{P}_\mu^1)$ est non affine en le paramètre μ par hypothèse sur g. Si on suit la démarche décrite sur le problème (5.1), le problème base réduite requiert de calculer la matrice $\tilde{G}_\mu$ définie par

$$\forall\ 1 \leq k, l \leq m, \quad (\tilde{G}_\mu)_{kl} = \int_\Omega g(\mu, x) u_{\mu_k} u_{\mu_l}.$$

Par hypothèse sur g, $\tilde{G}_\mu \neq f(\mu)\tilde{M}$. Donc, pour tout nouveau μ, il est nécessaire de former la matrice $\tilde{G}_\mu$. La complexité de cette étape est $\mathcal{O}(m^2 n_h^\eta)$ où $\eta = 1, 2$ selon le caractère creux de M. Par conséquent, il n'est pas possible de définir avec une telle stratégie une méthode de complexité indépendante de n_h.

[2] $(\gamma, \eta) = (1, 2)$ si A est pleine et $(\gamma, \eta) = (2, 1)$ si A est creuse.

Le problème $(\mathcal{P}_\mu^2)$ est non linéaire en u. La résolution de ce problème par une méthode itérative de Newton conduit à poser à chaque itération $u_\mu^{k+1} = u_\mu^k + \delta_{u_\mu}^k$. $\delta_{u_\mu}^k$ est obtenu en résolvant

$$\left| \begin{array}{rcl} \displaystyle\int_\Omega \nabla\delta_{u_\mu}^k.\nabla\chi_i \, d\Omega + \int_\Omega g'(\mu,u_\mu^k)\delta_{u_\mu}^k\chi_i \, d\Omega & = & \displaystyle\int_\Omega f\chi_i - \int_\Omega \nabla u_\mu^k.\nabla\chi_i \, d\Omega \\ & & \displaystyle - \int_\Omega g(\mu,u_\mu^k)\chi_i \, d\Omega \quad \forall\, 1 \le i \le n_h, \\ \delta_{u_\mu}^k & = & \displaystyle\sum_{i=1}^{n_h} U_{\mu,i}\chi_i. \end{array} \right.$$

La démarche décrite sur le problème (5.1) s'applique de la même façon, le problème base réduite correspondant requiert à chaque itération de calculer le vecteur G_u et la matrice G'_u définis par

$$\begin{aligned} \forall\, 1 \le k \le m, \qquad & (G_u)_k = \int_\Omega g(\mu,u_\mu)u_{\mu_k}, \\ \forall\, 1 \le k,l \le m, \qquad & (G'_u)_{kl} = \int_\Omega g'(\mu,u_\mu)u_{\mu_k}u_{\mu_l}. \end{aligned}$$

Par conséquent pour tout nouveau μ, la complexité de la résolution du problème base réduite n'est pas indépendante de n_h.

5.2 Le problème électronique de la chimie quantique

On considère l'évolution d'un système moléculaire composé de N paires d'électrons et de M noyaux de charge $(z_k)_{k=1,M}$. On note $\mu = (\bar{x}_k)_{k=1,M}$ les positions des noyaux du système. L'évolution du système conduit à calculer les forces électroniques agissant sur les noyaux. Pour cela, il est nécessaire pour chaque géométrie par laquelle passe le système de résoudre le problème électronique (5.8) décrit dans la section 5.2.2. On note n_h la dimension de l'espace de Galerkin introduit pour la résolution de ce problème. Deux types de base existent :

- les bases spécifiques (bases de gaussiennes, bases d'orbitales atomiques numériques), qu'on note par la suite B_1, dont la taille est de l'ordre du nombre d'électrons du système ($n_h = \mathcal{O}(N)$). Ces bases sont très efficaces d'un point de vue du temps de calcul suite à leur taille minimale ;
- les bases génériques (bases d'Eléments Finis, base d'ondes planes) qu'on note par la suite B_2. Ces bases, plus simples d'utilisation, sont caractérisées par une très grande taille ($n_h >> N$).

Dans la suite, le domaine d'étude est noté Ω et d désigne sa dimension.

5.2.1 Motivation

Pour simuler l'évolution du système, une dynamique moléculaire *ab initio* exige la résolution du problème électronique pour un nombre L de pas de temps, et donc un nombre L de jeux de positions des noyaux. La complexité de la résolution du problème électronique est $\mathcal{O}(n_h^\beta)$ (avec β égal à 1 ou 3 selon le système moléculaire). Pour des systèmes d'intérêt réel, N est de l'ordre de 10^4 et L de l'ordre de 10^5. Il est donc impossible de simuler l'évolution d'un système par une dynamique moléculaire *ab initio*. Toutefois, l'évolution des paramètres, et par suite de la solution du problème électronique, est lente. En pratique, on met à profit cette propriété en construisant l'initial guess à un pas de temps à partir de la solution obtenue au pas de temps précédent. Une telle stratégie reste insuffisante car la taille du problème à résoudre reste inchangée.

La méthode des bases réduites a pour but de diminuer la taille du problème à résoudre à chaque itération en approchant la solution cherchée par une combinaison linéaire de m solutions du problème électronique calculées à des itérations précédentes, avec $m \ll n_h$.

5.2.2 Version continue du problème électronique pour le modèle de Kohn-Sham

Le problème électronique, posé en terme de paires d'électrons (systèmes à couche fermée), consiste à calculer N fonctions $\Psi = (\psi_i)_{i=1,N}$ solution du problème d'optimisation non linéaire suivant

$$E^{KS}(\Psi) = \sum_{i=1}^{N} \int_\Omega |\nabla \psi_i|^2 + \int_\Omega \rho V_\mu + \frac{1}{2} \int_\Omega \int_\Omega \frac{\rho(x)\rho(x')}{|x-x'|}\, dx dx' + E_{xc}(\rho) \qquad (5.8)$$

sous les contraintes

$$\forall\, 1 \leq i \leq j \leq N, \quad \int_\Omega \psi_i \psi_j = \delta_{ij}$$

avec

$$\rho(x) = 2 \sum_{i=1}^{N} |\psi_i(x)|^2,$$

$$V_\mu(x) = -\sum_{k=1}^{M} \frac{z_k}{|x - \bar{x}_k|}.$$

E_{xc} est la fonctionnelle d'échange corrélation associée au modèle de Kohn-Sham

considéré. Les équations d'Euler-Lagrange associées à ce problème sont

$$\left|\begin{array}{l} \forall\, 1 \le i \le N, \quad -\dfrac{1}{2}\triangle\psi_i + V_\mu\psi_i + \Big(\rho \star \dfrac{1}{|x|}\Big)\psi_i + v_{xc}(\rho)\psi_i = \displaystyle\sum_{j=1}^{N} \lambda_{ij}\psi_j, \\ \qquad\qquad \rho = 2\displaystyle\sum_{i=1}^{N} |\psi_i|^2, \\ \forall\, 1 \le i \le j \le N, \quad \displaystyle\int_\Omega \psi_i\psi_j = \delta_{ij}. \end{array}\right. \tag{5.9}$$

où $v_{xc}(\rho)$ est la dérivée de $E_{xc}(\rho)$ par rapport à ρ. Soit ϕ la fonction définie par

$$-\triangle\phi = 4\pi\rho \quad \text{dans } \mathbb{R}^d, \tag{5.10}$$

on a pour toute fonction u définie sur $\mathbb{R}^d$

$$\int_{\mathbf{R}^d} \Big(\rho \star \frac{1}{|x|}\Big)(x)u(x)\,dx = \int_{\mathbf{R}^d}\int_{\mathbf{R}^d} \frac{\rho(x')u(x)}{|x-x'|}\,dx'dx = \int_{\mathbf{R}^d} \phi(x)u(x)\,dx.$$

Comme on résout les équations (5.9) sur un domaine borné Ω, on remplace (5.10) par

$$\left\{\begin{array}{rcll} -\triangle\phi & = & 4\pi\rho & \text{dans } \Omega, \\ \phi & = & 0 & \text{sur } \partial\Omega. \end{array}\right.$$

En supposant Ω assez grand, on fait l'approximation pour toute fonction u définie sur Ω

$$\int_\Omega\int_\Omega \frac{\rho(x')u(x)}{|x-x'|}\,dx'dx \simeq \int_\Omega \phi(x)u(x)\,dx. \tag{5.11}$$

Cette approximation permet de diminuer considérablement le temps de calcul du terme de Coulomb. En effet, le calcul de l'intégrale double de (5.11) est de complexité $\mathcal{O}(n_h^2)$ du fait de la non localité du potentiel de Coulomb. En revanche, le calcul de l'intégrale simple de (5.11) est de complexité $\mathcal{O}(n_h)$. De même, le calcul de ϕ est de complexité $\mathcal{O}(n_h)$ du fait de la localité de la base d'éléments finis.
Les équations (5.9) deviennent

$$\left|\begin{array}{l} \forall\, 1 \le i \le N, \quad -\dfrac{1}{2}\triangle\psi_i + V_\mu\psi_i + \phi\psi_i + v_{xc}(\rho)\psi_i = \displaystyle\sum_{j=1}^{N} \lambda_{ij}\psi_j, \\ \rho = 2\displaystyle\sum_{i=1}^{N} |\psi_i|^2,\ -\triangle\phi = 4\pi\rho \text{ dans } \Omega, \quad \phi = 0 \text{ sur } \partial\Omega, \\ \forall\, 1 \le i \le j \le N, \quad \displaystyle\int_\Omega \psi_i\psi_j = \delta_{ij}. \end{array}\right. \tag{5.12}$$

5.2.3 Difficultés liées au problème électronique

Les équations (5.12) présentent un grand nombre de difficultés.

5.2.3.1 Non linéarité par rapport à ψ_i

Les équations d'Euler-Lagrange sont non linéaires par rapport à ψ_i. Par ailleurs, les contraintes sur les ψ_i sont non linéaires. Par conséquent, la forme cherchée de la solution base réduite (comme combinaison linéaire de solutions précédemment calculées) ne satisfait pas automatiquement les contraintes. Il est donc nécessaire de les imposer dans l'algorithme base réduite afin de converger vers une bonne approximation de la solution à approcher. Ceci n'est pas si simple car on n'est pas assuré que la solution base réduite la plus proche en norme L^2 de la solution à approcher vérifie les contraintes du problème. Par conséquent, on ne peut pas imposer trop fortement les contraintes dans l'algorithme base réduite sous peine de converger vers une mauvaise approximation.

5.2.3.2 Non linéarité par rapport à μ

Seul le potentiel V_μ dépend non linéairement de μ. Toutefois, les solutions ψ_i dépendent fortement de μ : les fonctions ψ_i présentent un *cusp* en les positions $\bar{x}_k$. Par conséquent, l'approche base réduite n'est pas pertinente si on considère les orbitales en temps que telles. En vue de proposer une méthode de base réduite efficace, on distingue deux voies.

- On peut soit considérer une base de type B_1 pour laquelle les fonctions η_k présentent une dépendance par rapport à μ. On applique ensuite la méthode de base réduite aux grandeurs algébriques représentant les orbitales dans la base considérée.
- On peut soit considérer une base de type B_2 pour laquelle les fonctions η_k ne présentent pas une dépendance par rapport à μ. Au préalable, on ramène le problème fin à un problème dépendant plus fortement de μ pour lequel les positions des singularités de la solution sont désormais indépendantes de μ.

Comme les bases de type B_1 sont complexes et difficiles d'accès, on a considéré une base d'Elements Finis P_1 qui partage toutefois avec les bases de type B_1 la propriété de localité des fonctions la formant. On ramène dans un premier temps le problème (5.8) à un problème dépendant plus fortement de μ dans un domaine de référence $\bar{\Omega}$ (associé à un maillage de référence) indépendants de μ dans lequel la position des noyaux est fixée. Ainsi, les singularités de la solution de ce nouveau problème sont fixes dans le domaine de référence. Ce point sera détaillé dans la Section 5.5.

5.2.3.3 N inconnues

Enfin, le problème électronique exige de calculer N fonctions. On ne peut donc pas espérer *a priori* un calcul base réduite de complexité moindre que $\mathcal{O}(N)$. Pour des isolants, une telle complexité est déjà atteinte par les méthodes de complexité linéaire (voir chapitre 1).

5.3 Extension de la méthode des bases réduites

On considère les problèmes $(\mathcal{P}_\mu^1)$ et $(\mathcal{P}_\mu^2)$ définis dans la Section 5.1.2. On note désormais $\mathcal{B}_u$ la base réduite $\mathcal{B}$ associée à u. Dans le cas où des estimateurs *a posteriori* ne sont pas disponibles, $\mathcal{B}_u$ est construite de façon à vérifier

$$\sup_{\mu\in S_\mu}\inf_{\alpha\in\mathbf{R}^m}\left\|u(\mu)-\sum_{k=1}^{m}\alpha_k u(\mu_k)\right\|_\infty \leq \epsilon \tag{5.13}$$

où S_μ désigne un ensemble de valeurs de μ pavant le domaine de variation de μ et ϵ une certaine tolérance.

5.3.1 EDP non linéaire en μ

Soit u_μ la solution de $(\mathcal{P}_\mu^1)$ défini par

$$\left|\begin{array}{rcll} -\triangle u + g(\mu,x)u & = & f & \text{dans } \Omega, \\ u & = & 0 & \text{sur } \partial\Omega. \end{array}\right.$$

5.3.1.1 Traitement de la non linéarité

Afin de traiter le terme non linéaire, on introduit une seconde base réduite $\mathcal{B}_g = \Big(g(\mu_k,x)\Big)_k$ où la famille des μ_k vérifie

$$\sup_{\mu\in S_\mu}\inf_{\gamma\in\mathbf{R}^K}\left\|g(\mu,x)-\sum_{k=1}^{K}\gamma_k g(\mu_k,x)\right\|_\infty \leq \epsilon. \tag{5.14}$$

On pose ensuite

$$g(\mu,x)=\sum_{k=1}^{K}\gamma_k(\mu)g(\mu_k,x).$$

Le coefficient $\gamma(\mu)\in\mathbb{R}^K$ s'obtient en résolvant le système

$$\forall\ 1\leq i\leq K,\quad g(\mu,\tilde{x}_i)=\sum_{k=1}^{K}\gamma_k(\mu)g(\mu_k,\tilde{x}_i)$$

où $(\tilde{x}_i)_{i=1,K}$, appelés *magic points*, sont des points de Ω bien choisis. Il sont générés par l'algorithme suivant proposé par Y. Maday [A3]

- (i) $r_1(x)=g(\mu_1,x)$,
- (ii) $\tilde{x}_1=argmax\big\{|r_1(x)|\big\}$, $B_{11}^1=g(\mu_1,\tilde{x}_1)$,
- (iii) pour tout $2\leq k\leq K$

1. $B^{k-1}\delta = [g(\mu_k, \tilde{x}_1), \ldots, g(\mu_k, \tilde{x}_{k-1})]^t$, $\delta \in \mathbb{R}^{k-1}$,
2. $r_k(x) = g(\mu_k, x) - \sum_{j=1}^{k-1} \delta_j g(\mu_j, x)$,
3. $\tilde{x}_k = argmax\{|r_k(x)|\}$,
4. $B^k_{ij} = g(\mu_i, \tilde{x}_j), \quad \forall\, 1 \leq i, j \leq k, \quad B^k \in \mathbb{R}^{k\times k}$.

Par conséquent, $\gamma(\mu)$ s'obtient en résolvant le système linéaire

$$B^K \gamma(\mu) = [g(\mu, \tilde{x}_1), \ldots, g(\mu, \tilde{x}_K)]^t.$$

5.3.1.2 Calcul base réduite

Avant de procéder au calcul d'une solution approchée pour différents μ, on génère la matrice B^K (et son inverse) et les *magic points* $\tilde{x}_k$. Ensuite on précalcule les éléments $\tilde{A}$ et $\tilde{F}$ par (5.7) ainsi que les matrices $(\tilde{G}^k)_{k=1,K}$ définies par

$$\tilde{G}^k_{ij} = \int_\Omega g(\mu_k, x) u_{\mu_i} u_{\mu_j}\, d\Omega. \tag{5.15}$$

Le coût des précalculs est $\mathcal{O}(Km^2 n_h)$. A μ donné, $\tilde{u}_\mu$ s'obtient en résolvant

$$\left(\tilde{A} + \sum_{k=1}^{K} \gamma_k(\mu)\tilde{G}^k\right)\alpha = \tilde{F}.$$

La complexité du calcul de α est $\mathcal{O}(m^3 + K^2 + Km^2)$. Pour ce type de problème, des estimateurs *a posteriori* adaptés du cas linéaire ont été testés avec succès. Par ailleurs, les valeurs de μ correspondant aux éléments de $\mathcal{B}_u$ et celles correspondant aux éléments de $\mathcal{B}_g$ se révèlent différentes.

5.3.2 EDP non linéaire en u et μ

Soit u_μ la solution de $(\mathcal{P}^2_\mu)$ défini par

$$\left|\begin{array}{rcll} -\triangle u + g(\mu, u) & = & f & \text{dans } \Omega, \\ u & = & 0 & \text{sur } \partial\Omega. \end{array}\right.$$

5.3.2.1 Traitement de la non linéarité

De la même façon que précédemment, on pose

$$g(\mu, u_\mu) = \sum_{k=1}^{K} \gamma_k(\alpha) g(\mu_k, u_{\mu_k}) \tag{5.16}$$

avec $\mathcal{B}_g = \Big(g(\mu_k, u_{\mu_k})\Big)_k$, la famille des μ_k vérifiant (5.14). On calcule ensuite γ en résolvant le système

$$\forall\, 1 \leq i \leq K, \quad g\Big(\mu, \sum_{j=1}^{m} \alpha_j u_{\mu_j}\Big)(\tilde{x}_i) = \sum_{k=1}^{K} \gamma_k(\alpha) g(\mu_k, u_{\mu_k})(\tilde{x}_i)$$

où $(\tilde{x}_i)_{i=1,K}$ sont générés à l'aide de l'algorithme présenté dans la section précédente. $\gamma(\alpha)$ s'obtient par la formule suivante

$$B^K \gamma(\alpha) = G(\alpha) \tag{5.17}$$

avec

$$\forall\, 1 \leq i \leq K, \quad G(\alpha)_i = g\Big(\mu, \sum_{j=1}^{m} \alpha_j u_{\mu_j}\Big)(\tilde{x}_i).$$

5.3.2.2 Calcul base réduite

Avant de procéder au calcul d'une solution approchée pour différents μ, on génère la matrice B^K (et son inverse) et les *magic points* $\tilde{x}_k$. Ensuite on précalcule les matrices $\tilde{A}$ et $\tilde{F}$ et la matrice $\tilde{G}$ définies par

$$\tilde{G}_{ik} = \int_\Omega g\big(\mu_k, u(\mu_k)\big) u_{\mu_i}\, d\Omega. \tag{5.18}$$

Le coût des précalculs est $\mathcal{O}(Kmn_h)$. A μ donné, la solution base réduite s'obtient en résolvant

$$\left| \begin{array}{rcl} \tilde{A}\alpha + \tilde{G}\gamma(\alpha) & = & \tilde{F}, \\ \gamma(\alpha) & = & \Big(B^K\Big)^{-1} G(\alpha). \end{array} \right. \tag{5.19}$$

Une méthode de Newton est utilisée pour résoudre (5.19). Soit α_n la solution à l'itération n, $\alpha_{n+1} = \alpha_n + \delta\alpha$ où

$$\Big(\tilde{A} + \tilde{G}\big(B^K\big)^{-1} G'(\alpha_n)\Big)\delta\alpha = \tilde{F} - \tilde{A}\alpha_n - \tilde{G}\gamma(\alpha_n)$$

avec

$$\forall\, 1 \leq k \leq K,\ \forall\, 1 \leq i \leq m, \quad G'_{ki}(\alpha_n) = \frac{\partial g}{\partial u}\Big(\mu, \sum_{j=1}^{m} \alpha_j u_{\mu_j}\Big)(\tilde{x}_k) u_{\mu_i}(\tilde{x}_i).$$

La complexité du calcul de α est $\mathcal{O}(m^3 + K^2 + Km^2)$ si on suppose que le nombre d'itérations de Newton est indépendant de m et K.

5.4 Un problème aux valeurs propres non linéaire

Le problème électronique est un problème d'optimisation non linéaire en ψ_i. Dans une première étape, on a considéré un problème aux valeurs propres non linéaire simplifié.

5.4.1 Problème fin

Soit Ω l'intervalle $[0, a]$, le problème est le suivant

$$\text{Inf}\left\{ \int_\Omega |\nabla u|^2 \, d\Omega - \mu \int_\Omega u^4 \, d\Omega, \ \int_\Omega u^2 \, d\Omega = 1, \ u \in H_0^1(\Omega) \right\}.$$

Le minimiseur (u_μ, λ_μ) satisfait les équations d'Euler-Lagrange suivante

$$\left\{ \begin{array}{rcl} -\triangle u_\mu - 2\mu u_\mu^3 & = & \lambda_\mu \, u_\mu \quad \text{dans } \Omega, \\ \displaystyle\int_\Omega u_\mu^2 \, d\Omega & = & 1, \\ u_\mu & \in & H_0^1(\Omega). \end{array} \right.$$

5.4.2 Stratégie base réduite

La stratégie base réduite proposée consiste à introduire la fonction g définie par $g(u) = u^3$. Ensuite

- on définit les bases $\mathcal{B}_u$ et $\mathcal{B}_g$ pour une tolérance ϵ donnée à l'aide de (5.13) et de (5.14),
- on calcule les *magic points* $(\tilde{x}_k)_{k=1,K}$ associés à $\mathcal{B}_g$,
- on précalcule la matrice B^K et son inverse,
- on précalcule les matrices $\tilde{A}$, $\tilde{M}$ et $\tilde{G}$ à l'aide de (5.7) et (5.18).

A μ donné, on calcule par une méthode de Newton la solution $(\alpha, \tilde{\lambda}_\mu)$ du système algébrique suivant

$$\left\{ \begin{array}{rcl} \tilde{A}\alpha - 2\mu\tilde{G}\gamma(\alpha) & = & \tilde{\lambda}_\mu \tilde{M}\alpha, \\ \alpha^t \tilde{M}\alpha & = & 1. \end{array} \right.$$

où $\gamma(\alpha)$ s'obtient par (5.17).

Remarque 5.4.1 *On peut définir une seconde stratégie base réduite qui consiste à résoudre le problème d'optimisation initial en tenant compte de* (5.16)*. Dans ce cas, on définit g par $g(u) = u^4$. Ceci permet d'utiliser une recherche linéaire lors de la résolution du problème en α, ce qui n'est pas possible pour la stratégie présentée ci-dessus (les équations vérifiées par α ne proviennent pas d'un problème d'optimisation).*

5.4.3 Résultats obtenus

On a considéré $a = 20$ et 200 valeurs de μ répartis uniformément sur l'intervalle $[0, a]$ pour définir l'espace de solutions S_μ. Chaque solution fine est définie sur un maillage constitué de n_h éléments.
Dans un premier temps, n_h est pris égal à 200. Pour $\epsilon = 10^{-5}$, (5.13) et (5.14) conduisent aux bases réduites $\mathcal{B}_u$ et $\mathcal{B}_g$ définies par

$$\begin{aligned} \mathcal{B}_u &= \{0;\ 3.21;\ 7.44;\ 14.86;\ 17.89;\ 19.90\},\ \mathrm{card}\big(\mathcal{B}_u\big) = 6, \\ \mathcal{B}_g &= \{0;\ 2.91;\ 7.09;\ 9.55;\ 11.76;\ 15.58;\ 18.39;\ 19.40;\ 20\},\ \mathrm{card}\big(\mathcal{B}_g\big) = 9. \end{aligned}$$

A la base réduite $\mathcal{B}_g$, on associe les *magic points* $(\tilde{x}_k)_{k=1,9}$ suivants

$$(\tilde{x}_k)_k = \big\{3.8;\ 6.8;\ 7.7;\ 9.2;\ 10;\ 11.3;\ 13.8;\ 14.6;\ 17.5\big\}.$$

La Figure 5.2 représente en fonction de μ les erreurs e_λ, e_L et e_H commises par l'algorithme base réduite avec

$$e_\lambda = |\tilde{\lambda}_\mu - \lambda_\mu|,\ e_L = \|\tilde{u}_\mu - u_\mu\|_{L^2(\Omega)},\ e_H = |\tilde{u}_\mu - u_\mu|_{H^1(\Omega)}. \tag{5.20}$$

Les précisions obtenues sont très satisfaisantes. On constate que la précision est d'autant meilleure qu'on considère dans cet ordre la semi-norme H^1 de u_μ, la norme L^2 de u_μ et λ_μ.
La Figure 5.3 représente en fonction de μ le temps de calcul de la solution fine (λ_μ, u_μ) (FEM) et de la solution base réduite $(\tilde{\lambda}_\mu, \tilde{u}_\mu)$ (BRD). Pour chaque valeur de μ, l'approche base réduite conduit à une réduction par 10 du temps de calcul de la solution associée.

On a ensuite fait varier la valeur de m entre 3 et 6 et K entre 3 et 9. La Figure 5.4 représente les erreurs e_λ, e_L et e_H pour plusieurs choix de couple (m, K) pour $\mu = 10$. On constate que les erreurs décroissent à m fixé quand K croît, et ce d'autant plus que m est grand. De même, les erreurs diminuent à K fixé quand m augmente, et ce d'autant plus que K est grand. On observe que le gain en terme de précision est plus important quand on augmente K à m fixé que quand on augmente m à K fixé. Pour ce problème, il est donc plus pertinent d'enrichir la base $\mathcal{B}_g$ que la base $\mathcal{B}_u$.

Dans un second temps, on définit pour $\mu = 10$ une solution de référence (u_{ref}, λ_{ref}) obtenue pour $n_h = 2560$. La Figure 5.5 représente en fonction de n_h les temps de calcul de (λ_μ, u_μ) et $(\tilde{\lambda}_\mu, \tilde{u}_m u)$ et les erreurs associées $e_\lambda^{ref}(.)$, $e_L^{ref}(.)$, $e_H^{ref}(.)$ définies par

$$e_\lambda^{ref}(.) = |\lambda_{ref} - .\ |,\ e_L^{ref} = \|u_{ref} - .\ \|_{L^2(\Omega)},\ e_H^{ref} = |u_{ref} - .\ |_{H^1(\Omega)}.$$

On observe, excepté pour le maillage de référence, que $(\tilde{\lambda}_\mu, \tilde{u}_\mu)$ est aussi précise que (λ_μ, u_μ). Par ailleurs, on note d'une part que le temps de calcul de $(\tilde{\lambda}_\mu, \tilde{u}_\mu)$ lorsque $n_h \in [80, 2560]$ est inférieur au temps de calcul de (λ_μ, u_μ) dès que $n_h = 80$. D'autre part, la précision de $(\tilde{\lambda}_\mu, \tilde{u}_\mu)$ pour $n_h \in [80, 2560]$ est meilleure que la précision de (λ_μ, u_μ) pour $n_h = 80$. Par conséquent dès que $n_h = 80$, la solution $(\tilde{\lambda}_\mu, \tilde{u}_\mu)$ est plus précise que la solution fine (λ_μ, u_μ) qu'on calculerait pour un même temps calcul.

5.5 La molécule H_2^+

Afin d'intégrer les difficultés liées au potentiel nucléaire V_μ et au *cusp* présentés par la solution en les positions des noyaux, on a considéré le système H_2^+ en géométrie 2D axisymétrique.

5.5.1 Problème fin

On introduit la distance internucléaire μ et (a, b, c) trois réels positifs tels que $c > \mu/2$ et $a >> \mu$. Soit $\Omega = [-a-\mu/2, a+\mu/2] \times [0, a]$, u_μ est solution du problème suivant

$$\mathrm{Inf}\left\{\frac{1}{2}\int_\Omega |\nabla u|^2 \, d\Omega - \int_\Omega V_\mu u^2 \, d\Omega, \ \int_\Omega u^2 \, d\Omega = 1, \ u \in H_0^1(\Omega)\right\}$$

avec

$$\forall\ (x,y) \in \Omega, \quad V_\mu(x,y) = \frac{1}{\sqrt{(x+\mu/2)^2+y^2}} + \frac{1}{\sqrt{(x-\mu/2)^2+y^2}}.$$

Le minimiseur u_μ présente deux *cusps* en $(-\mu/2, 0)$ et $(\mu/2, 0)$ (cf. la Figure 5.6 représentant quelques solutions obtenues pour la molécule H_2). On définit la transformation f_μ (voir Fig. 5.1) par

$$\left|\begin{array}{ll} f_{1,\mu}(x,y) = (x+\mu/2-b-c, y) & (x,y) \in [-a-\mu/2, -\mu/2+c] \times [0,a], \\ f_{2,\mu}(x,y) = \Big(\dfrac{2b}{\mu-2c}x, y\Big) & (x,y) \in [-\mu/2+c, \mu/2-c] \times [0,a], \\ f_{3,\mu}(x,y) = (x-\mu/2+b+c, y) & (x,y) \in [\mu/2-c, a+\mu/2] \times [0,a]. \end{array}\right. \tag{5.21}$$

$\bar{\Omega} = f_\mu(\Omega) = [-a-b-c, a+b+c]$ ne dépend plus de μ. On définit un *maillage de référence sur* $\bar{\Omega}$*, indépendant donc de* μ (le maillage sur Ω s'obtenant en appliquant f_μ^{-1} au maillage de référence). La solution $\bar{u}_\mu$ est solution du problème d'optimisation suivant

$$\bar{u}_\mu = \inf\left\{\mathcal{L}(\bar{u},\mu) - \mathcal{V}(\bar{u},\mu), \ \mathcal{C}(\bar{u},\mu) = 1, \ \bar{u} \in H_0^1(\bar{\Omega})\right\}$$

avec

$$\begin{aligned}\mathcal{L}(\bar{u},\mu) = \frac{1}{2}\int_{\bar{\Omega}_1} |\nabla \bar{u}|^2 \, d\bar{\Omega}_1 + \frac{1}{2}\int_{\bar{\Omega}_3} |\nabla \bar{u}|^2 \, d\bar{\Omega}_3 \\ + \frac{1}{2}\int_{\bar{\Omega}_2} \frac{2b}{\mu-2c}\Big(\frac{\partial \bar{u}}{\partial \tilde{x}_2}\Big)^2 + \frac{\mu-2c}{2b}\Big(\frac{\partial \bar{u}}{\partial \tilde{y}_2}\Big)^2 \, d\bar{\Omega}_2,\end{aligned}$$

$$\mathcal{C}(\bar{u},\mu) = \int_{\bar{\Omega}_1} \bar{u}^2 \, d\bar{\Omega}_1 + \frac{\mu-2c}{2b}\int_{\bar{\Omega}_2} \bar{u}^2 \, d\bar{\Omega}_2 + \int_{\bar{\Omega}_3} \bar{u}^2 \, d\bar{\Omega}_3$$

et

$$\mathcal{V}(\bar{u},\mu) = \int_{\bar{\Omega}_1} g_1(\mu,\bar{x}_1,\bar{y}_1)\bar{u}^2\, d\bar{\Omega}_1 + \int_{\bar{\Omega}_3} g_3(\mu,\bar{x}_3,\bar{y}_3)\bar{u}^2\, d\bar{\Omega}_3 + \frac{\mu-2c}{2b}\int_{\bar{\Omega}_2} g_2(\mu,\bar{x}_2,\bar{y}_2)\bar{u}^2\, d\bar{\Omega}_2$$

où

$$\left\{\begin{array}{rcl} g_1(\mu,\bar{x}_1,\bar{y}_1) & = & \dfrac{1}{\sqrt{(\bar{x}_1+b+c)^2+\bar{y}_1^2}} + \dfrac{1}{\sqrt{(\bar{x}_1-\mu+b+c)^2+\bar{y}_1^2}}, \\ g_2(\mu,\bar{x}_2,\bar{y}_2) & = & \dfrac{1}{\sqrt{\left(\frac{\mu-2c}{2b}\bar{x}_2+\frac{\mu}{2}\right)^2+\bar{y}_2^2}} + \dfrac{1}{\sqrt{\left(\frac{\mu-2c}{2b}\bar{x}_2-\frac{\mu}{2}\right)^2+\bar{y}_2^2}}, \\ g_3(\mu,\bar{x}_3,\bar{y}_3) & = & \dfrac{1}{\sqrt{(\bar{x}_3+\mu-b-c)^2+\bar{y}_3^2}} + \dfrac{1}{\sqrt{(\bar{x}_3-b-c)^2+\bar{y}_3^2}}. \end{array}\right.$$

Les équations d'Euler-Lagrange vérifiées par $(\bar{u}_\mu,\bar{\lambda}_\mu)$ sont

$$\left\{\begin{array}{rcl} \bar{A}_\mu\bar{u}_\mu & = & \bar{\lambda}_\mu\bar{M}_\mu\bar{u}_\mu \\ \bar{u}_\mu^t\bar{M}_\mu\bar{u}_\mu & = & 1 \end{array}\right.$$

avec

$$\begin{array}{rcl} \bar{A}_\mu & = & \bar{A}_1 + \dfrac{2b}{\mu-2c}\bar{A}_2^x + \dfrac{\mu-2c}{2b}\bar{A}_2^y + \bar{A}_3 + \bar{G}_1^\mu + \dfrac{\mu-2c}{2b}\bar{G}_2^\mu + \bar{G}_3^\mu, \\ \bar{M}_\mu & = & \bar{M}_1 + \dfrac{\mu-2c}{2b}\bar{M}_2 + \bar{M}_3. \end{array}$$

Les matrices $(\bar{A}_i)_i$ et $(\bar{M}_i)_i$ ne dépendent pas de μ et

$$\forall\ 1\le k\le 3,\quad \left(\bar{G}_k^\mu\right)_{ij} = \int_{\bar{\Omega}_k} g_k(\mu)\eta_i\eta_j\, d\bar{\Omega}_k,$$

les η_j désignant les fonctions eléments finis. Les forces électroniques $\bar{F}_\mu$ s'obtiennent ensuite par

$$\bar{F}_\mu = \frac{\partial\bar{\lambda}_\mu}{\partial\mu} = \frac{\bar{u}_\mu^t\bar{A}_\mu\bar{u}_\mu}{\partial\mu} = \bar{u}_\mu^t\bar{K}\bar{u}_\mu$$

où

$$\bar{K} = -\frac{2b}{(\mu-2c)^2}\bar{A}_2^x + \frac{1}{2b}\bar{A}_2^y - \frac{1}{2b}\bar{G}_2^\mu + \bar{G}'_{\mu,1} + \frac{\mu-2c}{2b}\bar{G}'_{\mu,2} + \bar{G}'_{\mu,3} - \frac{\bar{\lambda}_\mu}{2b}\bar{M}_2,$$

et

$$\forall\ 1\le k\le 3,\quad \left(\bar{G}'_{\mu,k}\right)_{ij} = \int_{\bar{\Omega}_k} g'_k(\mu)\eta_i\eta_j\, d\bar{\Omega}_k.$$

5.5.2 Stratégie base réduite

Afin d'approcher la non linéarité en μ des fonctions g_k, on génère par (5.14) trois jeux de paramètres $S_{g_k} = (\mu_l^k)_{l=1,\,K_k}$ et 3 jeux de *magic points* $(\tilde{x}_l^k, \tilde{y}_l^k)_{l=1,\,K_k}$. Puis, pour tout μ, on définit $(\gamma^k)_{k=1,\,3}$ par

$$\forall\, 1 \leq k \leq 3,\ \forall\, 1 \leq l \leq K_k, \quad g_k(\mu, \tilde{x}_l^k, \tilde{y}_l^k) = \sum_{p=1}^{K_k} \gamma_p^k g_k(\mu_p^k, \tilde{x}_l^k, \tilde{y}_l^k).$$

Remarque 5.5.1 *Les intégrales provenant de V_μ sont calculées à l'aide de formules de quadrature qui font intervenir les milieux des triangles élémentaires constituant le maillage de référence. Les deux noyaux coïncident avec deux sommets du maillage de référence.*

En posant $\tilde{u}_\mu = \sum_{i=1}^{m} \alpha_i \bar{u}_{\mu_i}$, α est solution de

$$\left\{ \begin{array}{rcl} \tilde{A}_\mu \alpha & = & \tilde{\lambda}_\mu \tilde{M}_\mu \alpha \\ \alpha^t \tilde{M}_\mu \alpha & = & 1 \end{array} \right.$$

avec

$$\begin{array}{rcl} \tilde{A}_\mu & = & \tilde{A}_1 + \dfrac{2b}{\mu - 2c}\tilde{A}_2^x + \dfrac{\mu - 2c}{2b}\tilde{A}_2^y + \tilde{A}_3 + \tilde{G}_1^\mu - \dfrac{\mu - 2c}{2b}\tilde{G}_2^\mu - \tilde{G}_3^\mu, \\ \tilde{M}_\mu & = & \tilde{M}_1 + \dfrac{\mu - 2c}{2b}\tilde{M}_2 + \tilde{M}_3. \end{array}$$

Les matrices $(\tilde{A}_i)_i$, $(\tilde{M}_i)_i$ et $(\tilde{G}_i^\mu)_i$ sont construites par

$$\forall\, 1 \leq k \leq 3,\ \forall\, 1 \leq i, j \leq m, \left| \begin{array}{rcl} \left(\tilde{A}_k\right)_{ij} & = & \bar{u}_{\mu_i}^t \bar{A}_k \bar{u}_{\mu_j} \\ \left(\tilde{M}_k\right)_{ij} & = & \bar{u}_{\mu_i}^t \bar{M}_k \bar{u}_{\mu_j} \end{array} \right. \tag{5.22}$$

et

$$\forall\, 1 \leq k \leq 3,\ \forall\, 1 \leq i, j \leq m, \quad \left(\tilde{G}_k^\mu\right)_{ij} = \sum_{l=1}^{K_k} \gamma_l^k \bar{u}_{\mu_i}^t \left(G_k^{\mu_l^k}\right) \bar{u}_{\mu_j}.$$

Les forces électroniques s'approchent finalement par

$$\tilde{F}_\mu = \frac{\partial \tilde{\lambda}_\mu}{\partial \mu} = \frac{\partial \alpha^t \tilde{A}_\mu \alpha}{\partial \mu} = \alpha^t \tilde{K} \alpha$$

où

$$\tilde{K} = -\frac{2b}{(\mu - 2c)^2}\tilde{A}_2^x + \frac{1}{2b}\tilde{A}_2^y - \frac{1}{2b}\tilde{G}_2^\mu - \sum_{k=1}^{3}\sum_{l=1}^{K_k} \frac{\partial \gamma_l^k}{\partial \mu} \tilde{G}_k^{\mu_l^k} - \frac{\tilde{\lambda}_\mu}{2b}\tilde{M}_2.$$

et

$$\begin{aligned}
\forall\, 1 \le k \le 3,\ \forall\, 1 \le l \le K_k, \quad \frac{\partial \gamma_l^k}{\partial \mu} &= \left((B^k)^{-1} d_{g_k}\right)_l, \\
\forall\, 1 \le k \le 3,\ \forall\, 1 \le l, p \le K_k, \quad (B^k)_{lp} &= g_k(\mu_p^k, \tilde{x}_l^k, \tilde{y}_l^k), \\
\forall\, 1 \le k \le 3,\ \forall\, 1 \le l \le K_k, \quad \left(d_{g_k}\right)_l &= \frac{\partial g_k}{\partial \mu}(\mu, \tilde{x}_l^k, \tilde{y}_l^k).
\end{aligned}$$

5.5.3 Résultats obtenus

On a considéré $a = 20$ et 100 valeurs de μ répartis uniformément sur l'intervalle $[1, 5]$ pour définir l'espace de solutions S_μ. On génère les bases réduites $\mathcal{B}_u$ et $\mathcal{B}_g^k$ (pour chaque terme non linéaire associé à g_k) pour $\epsilon = 10^{-5}$ à l'aide de (5.13) et (5.14).

Dans les graphiques présentés, K désigne la valeur commune des K_k, et les erreurs $|F(\mu) - F_N(\mu)|$, $|E(\mu) - E_N(\mu)|$ et $\|u(\mu) - u_N(\mu)\|_{H(\Omega)}$ correspondent respectivement aux erreurs $e_F = |\bar{F}_\mu - \tilde{F}_\mu|$, $e_E = e_\lambda$ et e_H définies en (5.20).

La Table 5.1 représente en fonction de μ, pour $m = 8$ et K entre 6 et 11, les erreurs e_E, e_F et e_u. Pour toutes les valeurs de K, les précisions obtenues sont très satisfaisantes. La précision sur λ_μ est meilleure que les précisions sur les forces électroniques et u_μ. On vérifie que les erreurs sur $(\tilde{\lambda}_\mu, \tilde{F}_\mu, \tilde{u}_\mu)$ décroissent en moyenne avec K.

La Figure 5.7 représente en fonction de K l'évolution des erreurs e_E, e_F et e_u pour $\mu = 2$ et $\mu = 2.6$. L'erreur sur u_μ, et globalement sur λ_μ, décroissent avec K. En revanche, l'erreur sur les forces électroniques ne décroît pas de façon monotone. Ce comportement numérique reste inexpliqué.

Afin de débrancher l'effet de l'interpolation des termes non linéaires, on a considéré l'algorithme de base réduite qui prend en compte le potentiel des noyaux de façon exacte. La Figure 5.8 représente en fonction de μ les erreurs e_E, e_F et e_u. On constate que les solutions obtenues sont plus précises d'un ordre 4 que les solutions obtenues lorsque le potentiel nucléaire est interpolé avec 11 fonctions. Par ailleurs, on note qu'une base réduite $\mathcal{B}_u$ de petite taille permet d'obtenir une très bonne précision pour des distances μ dans l'intervalle $[1, 5]$, ce qui représente un domaine de variation conséquent à l'échelle du noyau.

5.6 La molécule H_2

Afin d'intégrer les difficultés liées au terme de Coulomb, on a considéré le modèle *Restricted Hartree-Fock* pour décrire le système H_2 en géométrie 2D axisymétrique.

5.6.1 Problème fin

De la même façon que dans la Section 5.5, on introduit la distance internucléaire μ et (a_x, a_y, b, c) quatre réels positifs tels que $c > \mu/2$ et $a_x, a_y >> \mu$. Soit $\Omega = [-a_x - \mu/2, a_x + \mu/2] \times [0, a_y]$, u_μ est solution du problème suivant

$$\inf \left\{ \frac{1}{2} \int_\Omega |\nabla u|^2 \, d\Omega - \int_\Omega V_\mu u^2 + \int_\Omega \int_\Omega \frac{u^2(r) u^2(r')}{|r - r'|} \, d\Omega \, d\Omega, \ \int_\Omega u^2 \, d\Omega = 1, \ u \in H_0^1(\Omega) \right\}.$$

Or

$$\forall \, u \in L^3(\mathbb{R}^3), \quad \int_{\mathbf{R}^3} \int_{\mathbf{R}^3} \frac{u^2(r) u^2(r')}{|r - r'|} d\mathbb{R}^3 d\mathbb{R}^3 = \int_{\mathbf{R}^3} \phi u^2 \, d\mathbb{R}^3$$

avec $\phi \in L^2(\mathbb{R}^3)$ satisfaisant

$$-\triangle \phi = 4\pi u^2 \text{ dans } L^2(\mathbb{R}^3).$$

En tenant compte de l'axisymétrie du problème et considérant le domaine Ω assez grand, on fait l'approximation suivante

$$\int_\Omega \int_\Omega \frac{u^2(r) u^2(r')}{|r - r'|} \, d\Omega \, d\Omega \simeq \int_\Omega \phi u^2 \, d\Omega$$

avec

$$\left\{ \begin{array}{rcl} -\triangle \phi & = & 2u^2 \text{ dans } \Omega, \\ \phi & = & 0 \text{ sur } \partial\Omega. \end{array} \right.$$

Remarque 5.6.1 *Il est important de choisir un domaine Ω assez grand (soit a assez grand) afin d'approcher avec précision la fonctionnelle d'énergie et d'obtenir une approximation raisonnable de la solution. Notons que la connaissance du comportement asymptotique de ϕ loin des noyaux se traduit par une condition au limite sur $\partial\Omega$ de type Robin sur ϕ.*

En notant E_h l'espace d'approximation fin sur Ω, les équations d'Euler-Lagrange associées à ce problème d'optimisation satisfaites par (u_μ, ϕ_μ) s'écrivent sous leur forme variationnelle

$$\begin{aligned} \forall v \in E_h, \quad \frac{1}{2} \int_\Omega \nabla u_\mu . \nabla v \, d\Omega - \int_\Omega V_\mu u_\mu v \, d\Omega + 2 \int_\Omega \phi_\mu u_\mu v \, d\Omega &= \lambda \int_\Omega u_\mu v \, d\Omega, \\ \forall w \in E_h, \quad \int_\Omega \nabla \phi_\mu . \nabla w \, d\Omega &= 2 \int_\Omega u_\mu^2 w \, d\Omega, \\ \int_\Omega u_\mu^2 \, d\Omega &= 1. \end{aligned}$$

A l'aide de la transformation f_μ définie en (5.21), on définit $(\bar\lambda_\mu, \bar u_\mu, \bar\phi_\mu)$ la solution sur le maillage de référence. Soit $\bar E_h$ l'espace d'approximation fin sur $\bar\Omega$, les équations d'Euler-Lagrange vérifiées par $(\bar\lambda_\mu, \bar u_\mu, \bar\phi_\mu)$ s'écrivent sous leur forme variationnelle

$$(\bar{\mathcal{P}}_\mu) \quad \left\{ \begin{array}{rcl} \forall \bar v \in \bar E_h, \quad F(\bar u_\mu, \bar\phi_\mu, \bar v) & = & 0, \\ \forall \bar w \in \bar E_h, \quad G(\bar\phi_\mu, \bar u_\mu, \bar w) & = & 0, \\ \mathcal{C}(\bar u_\mu) & = & 1. \end{array} \right. \tag{5.23}$$

avec

$$\begin{aligned} F(\bar u, \bar\phi, \bar v) &= \frac{1}{2}\Big(\int_{\bar\Omega_1} \nabla \bar u.\nabla \bar v \; d\bar\Omega_1 + \int_{\bar\Omega_3} \nabla \bar u.\nabla \bar v \; d\bar\Omega_3 \Big) \\ &+ \frac{1}{2}\int_{\bar\Omega_2} \frac{2b}{\mu-2c}\frac{\partial \bar u}{\partial \bar x}.\frac{\partial \bar v}{\partial \bar x} + \frac{\mu-2c}{2b}\frac{\partial \bar u}{\partial \bar y}.\frac{\partial \bar v}{\partial \bar y} \; d\bar\Omega_2 \Big) \\ &- \int_{\bar\Omega_1} g_1(\mu, \bar x_1, \bar y_1)\bar u\bar v \; d\bar\Omega_1 - \frac{\mu-2c}{2b}\int_{\bar\Omega_2} g_2(\mu, \bar x_2, \bar y_2)\bar u\bar v \; d\bar\Omega_2 \\ &- \int_{\bar\Omega_3} g_3(\mu, \bar x_3, \bar y_3)\bar u\bar v \; d\bar\Omega_3 \\ &+ 2\Big(\int_{\bar\Omega_1} \bar\phi\bar u\bar v \; d\bar\Omega_1 + \frac{\mu-2c}{2b}\int_{\bar\Omega_2} \bar\phi\bar u\bar v \; d\bar\Omega_2 + \int_{\bar\Omega_3} \bar\phi\bar u\bar v \; d\bar\Omega_3 \Big) \\ &- \bar\lambda\Big(\int_{\bar\Omega_1} \bar u\bar v \; d\bar\Omega_1 + \frac{\mu-2c}{2b}\int_{\bar\Omega_2} \bar u\bar v \; d\bar\Omega_2 + \int_{\bar\Omega_3} \bar u\bar v \; d\bar\Omega_3 \Big), \end{aligned}$$

$$\begin{aligned} G(\bar\phi, \bar u, \bar w) &= \int_{\bar\Omega_1} \nabla \bar\phi.\nabla \bar w \; d\bar\Omega_1 + \int_{\bar\Omega_3} \nabla \bar\phi.\nabla \bar w \; d\bar\Omega_3 \\ &+ \int_{\bar\Omega_2} \frac{2b}{\mu-2c}\frac{\partial \bar\phi}{\partial \bar x}.\frac{\partial \bar w}{\partial \bar x} + \frac{\mu-2c}{2b}\frac{\partial \bar\phi}{\partial \bar y}.\frac{\partial \bar w}{\partial \bar y} \; d\bar\Omega_2 \\ &- 2\Big(\int_{\bar\Omega_1} \bar u^2\bar w \; d\bar\Omega_1 + \frac{\mu-2c}{2b}\int_{\bar\Omega_2} \bar u^2\bar w \; d\bar\Omega_2 + \int_{\bar\Omega_3} \bar u^2\bar w \; d\bar\Omega_3 \Big), \end{aligned}$$

$$\mathcal{C}(\bar u) = \int_{\bar\Omega_1} \bar u^2 \; d\bar\Omega_1 + \frac{\mu-2c}{2b}\int_{\bar\Omega_2} \bar u^2 \; d\bar\Omega_2 + \int_{\bar\Omega_3} \bar u^2 \; d\bar\Omega_3.$$

5.6.2 Stratégie base réduite et résultats obtenus

Les termes provenant de V_μ sont traités comme lors du traitement de H_2^+. De la même façon que pour $\bar u_\mu$, on définit une base réduite $\mathcal{B}_\phi$ pour approcher $\bar\phi_\mu$. On cherche ensuite, pour tout μ, $\tilde u_\mu$ et $\tilde\phi_\mu$ de la forme

$$\begin{aligned} \tilde u_\mu &= \sum_{i=1}^{m_u} \alpha_i \bar u_{\mu_i}, \\ \tilde\phi_\mu &= \sum_{j=1}^{m_\phi} \beta_j \bar\phi_{\mu_j}. \end{aligned}$$

α et β sont calculés en résolvant par une méthode de Newton les équations (5.23) écrites pour $(\bar{v}, \bar{w}) = (\bar{u}_{\mu_i}, \bar{\phi}_{\mu_j})$ pour tout $1 \leq i \leq m_u$ et pour tout $1 \leq j \leq m_\phi$.

Dans les tests numériques, on a considéré $a_x = 40$, $a_y = 8$ et 100 valeurs de μ répartis uniformément sur l'intervalle $[1, 10]$ pour définir l'espace de solutions S_μ. Pour $\epsilon = 10^{-5}$, (5.13) et (5.14) conduisent aux bases réduites $\mathcal{B}_u$, $\mathcal{B}_\phi$ et $\mathcal{B}_g^k$ (pour chaque terme non linéaire associé à g_k). La Figure 5.9, respectivement Fig. 5.10, représente les valeurs de μ associées aux solutions formant ces bases, respectivement les *magic points* associés à g_1 et g_2.

Dans les graphiques présentés, m, respectivement K, désigne la dimension commune de $\mathcal{B}_u$ et $\mathcal{B}_\phi$, respectivement des $\mathcal{B}_g^k$. Par ailleurs, les erreurs $e_E = |E(\mu) - E_N(\mu)|$, $e_F = |F(\mu) - F_N(\mu)|$, $e_\lambda = |\lambda(\mu) - \lambda_N(\mu)|$ et $e_u = \|u(\mu) - u_N(\mu)\|_{H(\Omega)}$ correspondent aux erreurs $|\bar{E}_\mu - \tilde{E}_\mu|$, $|\bar{F}_\mu - \tilde{F}_\mu|$, $|\bar{\lambda}_\mu - \tilde{\lambda}_\mu|$ et $\|\bar{u}_\mu - \tilde{u}_\mu\|_{H^1(\bar{\Omega})}$.

Les Tables 5.2 à 5.4 représentent pour plusieurs couples (m, K) les erreurs e_E, e_F, e_λ, et e_u en fonction de μ. De la même façon que dans la section 5.4, on observe à K fixé que la précision de $(\tilde{\lambda}_\mu, \tilde{u}_\mu, \tilde{\phi}_\mu)$ croît avec m, d'autant plus que K est grand. De même à m fixé, on constate que la précision de $(\tilde{\lambda}_\mu, \tilde{u}_m u, \tilde{\phi}_m u)$ augmente avec K, d'autant plus que m est grand. Cependant à l'inverse, quelle que soit les valeurs de K dans $[6, 9]$ et m dans $\{5, 7\}$, on remarque qu'il est plus judicieux d'enrichir de deux éléments les bases $\mathcal{B}_u$ et $\mathcal{B}_\phi$ que les bases $\mathcal{B}_g^k$. Ce comportement semble indiquer l'importance d'approcher correctement le terme dépendant de ϕ, soit le terme de Coulomb. Des tests plus exhaustifs sur (m_u, m_ϕ) à K fixé permettrait de confirmer cette affirmation.

Enfin, la Figure 5.11 représente l'évolution du temps de calcul de $(\bar{\lambda}_\mu, \bar{u}_\mu, \bar{\phi}_\mu)$ et de $(\tilde{\lambda}_\mu, \tilde{u}_\mu, \tilde{\phi}_\mu)$ en fonction de μ. On constate que calcul de $(\tilde{\lambda}_\mu, \tilde{u}_\mu, \tilde{\phi}_\mu)$ est 2000 fois plus rapide que le calcul de $(\bar{\lambda}_\mu, \bar{u}_\mu, \bar{\phi}_\mu)$ pour une précision satisfaisante sur l'énergie et les forces électroniques. On a constaté numériquement que l'erreur commise par l'algorithme de base réduite ne perturbait pas la simulation de l'évolution de la molécule à l'aide de l'algorithme de Verlet pour des temps longs.

Remarque 5.6.2 *L'extension de la méthodes des bases réduites au problème non linéaire peut s'appliquer directement pour le traitement du terme de Coulomb. Des tests numériques ont montré la pertinence d'une telle démarche. Cependant, le traitement du problème est plus efficace en terme de temps de calcul en introduisant la fonction ϕ.*

Remarque 5.6.3 *Le facteur* 2000 *obtenu entre le calcul fin et le calcul base réduite est une borne supérieure car on a considéré un maillage uniforme sur Ω alors que la solution varie significativement autour des noyaux seulement. Un maillage adapté conduirait très certainement à une diminution de ce facteur d'un ordre* $2,3$.

5.7 Conclusion

Les résultats encourageants obtenus lors des tests académiques sur des systèmes école laissent espérer un apport de la méthode des bases réduites pour le traitement de systèmes moléculaires. En vue de traiter des systèmes plus complexes et donc d'un intérêt plus grand, il reste encore trois difficultés à traiter. La première est la prise en compte du terme d'échange (pour *Hartree-Fock*) ou d'échange-corrélation (pour *Kohn-Sham*) lorsque le système considéré comporte plusieurs paires d'électrons.
La deuxième difficulté est liée à la définition de la transformation f_μ dans $\mathbb{R}^3$. Il devient alors indispensable d'utiliser des bases dépendantes de la position des noyaux. Les bases des chimistes, certes difficiles d'accés et peu maniables, sont les candidats idéaux dans cette optique.
La troisième difficulté est liée aux $\mathcal{O}(N^2)$ contraintes d'orthogonalité.

5.8 Présentation de la note CRAS

Dans cette note, une méthode rapide de base réduite pour la résolution d'équations partielles ayant une dépendance non affine en les paramètres du problème est présentée. On introduit d'une part un processus d'interpolation du terme non linéaire en les paramètres sur une base réduite adéquate via la construction d'ensembles emboîtés de points d'interpolation (correspondant aux "magic points" introduits dans les sections précédentes). D'autre part, on propose un estimateur *a posteriori* construit à partir des estimateurs développés pour le traitement de problèmes affines.

Dans un premier temps, le bon comportement du processus d'interpolation introduit est illustré sur une fonction quadratique en les paramètres. Dans un second temps, des résultats sur un système académique montrent la pertinence de l'association de la méthode des bases réduites avec ce processus d'interpolation, ainsi que de l'estimateur *a posteriori* proposé.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
C. R. Acad. Sci. Paris, Ser. I 339 (2004) 667–672
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Numerical Analysis

An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations

Maxime Barrault [a], Yvon Maday [b], Ngoc Cuong Nguyen [c], Anthony T. Patera [d]

[a] *CERMICS – ENPC, cité Descartes, Champs sur Marne, 77455 Marne la Vallée cedex 2, France*
[b] *Laboratoire J.-L. Lions, université Pierre et Marie Curie, B.C. 187, 75242 Paris cedex 05, France*
[c] *National University of Singapore, 10 Kent Ridge Crescent, Singapore 117576*
[d] *Massachusetts Institute of Technology, Department of Mechanical Engineering, Room 3-264, 77 Massachusetts Avenue, Cambridge, MA 02139-4307, USA*

Received 11 June 2004; accepted 28 August 2004

Available online 14 October 2004

Presented by Olivier Pironneau

Abstract

We present an efficient reduced-basis discretization procedure for partial differential equations with *nonaffine* parameter dependence. The method replaces nonaffine coefficient functions with a collateral reduced-basis expansion which then permits an (effectively affine) offline–online computational decomposition. The essential components of the approach are (i) a good collateral reduced-basis approximation space, (ii) a stable and inexpensive interpolation procedure, and (iii) an effective a posteriori estimator to quantify the newly introduced errors. Theoretical and numerical results respectively anticipate and confirm the good behavior of the technique. ***To cite this article: M. Barrault et al., C. R. Acad. Sci. Paris, Ser. I 339 (2004).***

Résumé

Une méthode d'« interpolation empirique » : application à la discrétisation efficace par base réduite d'équations aux dérivées partielles. Nous présentons dans cette Note une méthode rapide de base réduite pour la résolution d'équations aux dérivées partielles ayant une dépendance non affine en ses paramètres. L'approche propose de remplacer le calcul des fonctionelles non affines par un développement en base réduite annexe qui conduit à une évaluation en ligne effectivement affine. Les points essentiels de cette approche sont (i) un bon système de base réduite annexe, (ii) une méthode stable et peu coûteuse d'interpolation dans cette base, et (iii) un estimateur a posteriori pertinent pour quantifier les nouvelles erreurs introduites. Des résultats théoriques et numériques viennent anticiper puis confirmer le bon comportement de cette technique. ***Pour citer cet article : M. Barrault et al., C. R. Acad. Sci. Paris, Ser. I 339 (2004).***

E-mail address: patera@mit.edu (A.T. Patera).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2004.08.006

668 M. Barrault et al. / C. R. Acad. Sci. Paris, Ser. I 339 (2004) 667–672

Version française abrégée

Considérons une function $g(\cdot;\mu) \in L^\infty(\Omega)$ assez régulière. On propose tout d'abord une méthode constructive pour sélectionner une suite d'espaces emboités $W_M^g = \text{Vect}\{\xi_m = g(\cdot;\mu_m^g),\ 1 \leqslant m \leqslant M\}$ avec $M \leqslant M_{\max}$ de dimension exactement M. On construit ensuite des ensembles emboités de points d'interpolation $T_M = \{t_1,\dots,t_M\}$, $1 \leqslant M \leqslant M_{\max}$, en posant tout d'abord $t_1 = \arg\operatorname{ess\,sup}_{x\in\Omega} |\xi_1(x)|$, $q_1 = \xi_1(x)/\xi_1(t_1)$. Puis, pour $M = 2,\dots,M_{\max}$, on résout le système linéaire $\sum_{j=1}^{M-1} \sigma_j^{M-1} q_j(t_i) = \xi_M(t_i)$, $1 \leqslant i \leqslant M-1$, et on pose $r_M(x) = \xi_M(x) - \sum_{j=1}^{M-1} \sigma_j^{M-1} q_j(x)$, on définit alors le point suivant d'interpolation $t_M = \arg\operatorname{ess\,sup}_{x\in\Omega} |r_M(x)|$ et on pose $q_M(x) = r_M(x)/r_M(t_M)$. On approche enfin $g(x;\mu)$ par $g_M(x;\mu) = \sum_{m=1}^M \beta_m(\mu) q_m(x)$, où $\sum_{j=1}^M \beta_j(\mu) q_j(t_i) = g(t_i;\mu)$, $1 \leqslant i \leqslant M$.

Ce procédé d'interpolation peut être justifié. A priori tout d'abord, en introduisant la constante de type Lebesgue $\Lambda_M = \sup_{x\in\Omega} \sum_{m=1}^M |V_m^M(x)|$ où V_m^M est le seul élément de W_M^g tel que $V_m^M(t_i) = \delta_{im}$. On peut montrer que Λ_M est bornée par $2^M - 1$. L'erreur d'interpolation $\varepsilon_M(\mu) \equiv \|g(\cdot;\mu) - g_M(\cdot;\mu)\|_{L^\infty(\Omega)}$ vérifie alors $\varepsilon_M(\mu) \leqslant (1+\Lambda_M)\varepsilon_M^*(\mu)$ où $\varepsilon_M^*(\mu) \equiv \inf_{z\in W_M^g} \|g(\cdot;\mu) - z\|_{L^\infty(\Omega)}$, $\forall \mu \in \mathcal{D}$. Comme on le verra (et comme il est classique en approximation polynomiale) la borne sur la constante de Lebesgue bien que pessimiste est souvent compensée par la très rapide convergence de l'autre terme. On peut aussi proposer une approximation a posteriori en introduisant l'estimateur $\hat{\varepsilon}_M(\mu) \equiv |g(t_{M+1};\mu) - g_M(t_{M+1};\mu)|$, exact si $g(\cdot;\mu) \in W_{M+1}^g$ et asymptotique dans le cas contraire.

Le Tableau 1 synthétise les résultats numérique obtenus par la mise en oeuvre de cette interpolation. Il illustre le bon comportement de la méthode et des estimateurs sur le cas $g(x;\mu) \equiv \mathcal{V}((x_1,x_2);(\mu_1,\mu_2)) \equiv ((x_1-\mu_1)^2 + (x_2-\mu_2)^2)^{-1/2}$ pour $x \in \Omega \equiv]0,1[^2$ et $\mu \in \mathcal{D} \equiv [-1,-0.01]^2$. La convergence de la méthode est très rapide et la valeur moyenne $\bar{\eta}_M$ de $\hat{\varepsilon}_M(\mu)/\varepsilon_M(\mu)$ modérée.

Cette approche est ensuite couplée avec une méthode de discrétisation en base réduite pour l'approximation de la solution du problème : soit $\mu \in \mathcal{D}$, trouver $u(\mu) \in H_0^1(]0,1[^2)$ telle que $a(u,v;\mu) = f(v;\mu)$, $\forall v \in H_0^1(]0,1[^2$, où a est la forme introduite en (1), avec $g(x;\mu) \equiv \mathcal{V}(x;\mu)$; et $f(v;\mu) = \int_\Omega \mathcal{V}(x;\mu)v$.

La méthode en base réduite est alors : pour $\mu \in \mathcal{D}$, $u_{N,M}(\mu) \in W_N^u$ est la solution de $\int_\Omega \nabla u_{N,M}(\mu)\cdot\nabla v + \int_\Omega g_M(x;\mu)u_{N,M}(\mu)v = \int_\Omega \mathcal{V}(x;\mu)v$, $\forall v \in W_N^u$. L'espace en base réduite W_N^u est défini de façon classique par $W_N^u = \text{Vect}\{\zeta_n \equiv u(\mu_n^u),\ 1 \leqslant n \leqslant N\}$ où $\{\mu_n^u\}_{n=1,\dots,N_{\max}}$ est un jeu de paramètres bien choisis et $g_M(x;\mu) = \sum_{m=1}^M \beta_m(\mu) q_m(x)$ est l'interpolé « empirique » introduit ci-dessus. Le Tableau 2 présente les résultats de cette multiple approximation (base réduite + interpolation empirique) ainsi que la pertinence de l'estimateur a posteriori qui peut être construit en combinant l'estimateur précédent sur l'interpolation empirique et les estimateurs classiques en base réduite proposés par exemple dans [6,8].

1. Introduction

We consider a parametrized evaluation problem: Given a $\mu \in \mathcal{D} \subset \mathbb{R}^P$, evaluate $s(\mu) = \ell(u(\mu))$, where $u \in X$ is the solution of a second-order coercive elliptic partial differential equation $a(u,v;\mu) = f(v)$, $\forall v \in X$. Here μ and $\mathcal{D}$ are the parameter and parameter domain, respectively; X is a Hilbert space with associated inner product $(w,v)_X$ and norm $\|w\|_X$; $\Omega \subset \mathbb{R}^2$ is our spatial domain, a point in which shall be denoted (x_1,x_2); ℓ and f are linear bounded functionals; and, for any $\mu \in \mathcal{D}$, $a(\cdot,\cdot;\mu) : X \times X \to \mathbb{R}$ is a coercive continuous bilinear form.

In the reduced-basis approach [1–3,5,6] we first introduce nested parameter samples $S_N^u \equiv \{\mu_1^u,\dots,\mu_N^u\}$ and associated approximation spaces $W_N^u = \text{span}\{\zeta_n \equiv u(\mu_n^u),\ 1 \leqslant n \leqslant N\}$ for $N = 1,\dots,N_{\max}$; in actual practice, of course, $u(\mu_n^u)$ is replaced with a 'truth approximation' on (say) a suitably fine piecewise-linear finite element subspace of typically large dimension $\mathcal{N}$. The reduced-basis approximation is then: Given $\mu \in \mathcal{D}$, evaluate $s_N(\mu) = \ell(u_N(\mu))$, where $u_N(\mu) \in W_N^u$ is the solution of $a(u_N(\mu),v;\mu) = f(v)$, $\forall v \in W_N^u$. In general, $u_N(\mu) \to u(\mu)$ very rapidly as N increases [2,4].

We now expand $u_N(\mu) = \sum_{j=1}^N u_{Nj}(\mu)\zeta_j$. The u_{Nj}, $1 \leqslant j \leqslant N$, will then satisfy $\sum_{j=1}^N a(\zeta_j,\zeta_i;\mu)u_{Nj} = f(\zeta_i)$, $1 \leqslant i \leqslant N$; we may subsequently evaluate $s_N(\mu) = \sum_{j=1}^N u_{Nj}(\mu)\ell(\zeta_j)$. If $a(w,v;\mu)$ is affine in μ,

$a(w, v; \mu) = \sum_{k=1}^{K} \Theta^k(\mu) a^k(w, v)$, then an extremely efficient offline–online computational strategy (relevant in the many-query and real-time contexts) may be developed. In the offline stage we form $a^k(\zeta_j, \zeta_i)$, $1 \leqslant i, j \leqslant N_{\max}$, $1 \leqslant k \leqslant K$; in the online stage we need only assemble and invert $a(\zeta_j, \zeta_i; \mu) = \sum_{k=1}^{K} \Theta^k(\mu) a^k(\zeta_j, \zeta_i)$, $1 \leqslant i, j \leqslant N$. The online cost to evaluate $s_N(\mu)$ is thus $KN^2 + N^3 + N$ — independent of $\mathcal{N}$; since $N \ll \mathcal{N}$, large computational economies can be realized.

Unfortunately, if a is not affine in the parameter, the online complexity is no longer independent of $\mathcal{N}$. For example, for *general* $g(x; \mu)$, the bilinear form

$$a(w, v; \mu) \equiv \int_{\Omega} \nabla w \cdot \nabla v + \int_{\Omega} g(x; \mu) w v \tag{1}$$

will not admit an efficient online–offline decomposition. In this paper we describe a technique that recovers online $\mathcal{N}$ independence even in the presence of non-affine parameter dependence. Our approach (applied to (1), say) is simple: we develop a 'collateral' reduced-basis expansion $g_M(x; \mu)$ for $g(x; \mu)$; we then replace $g(x; \mu)$ in (1) with the (necessarily) affine approximation $g_M(x; \mu)$. The essential ingredients are (i) a 'good' collateral reduced-basis approximation space, (ii) a stable and inexpensive interpolation procedure, and (iii) an effective a posteriori estimator to quantify the newly introduced error terms.

In Section 2 we develop our coefficient-function approximation method; in Section 3 we present a priori and a posteriori error analyses; and in Section 4 we incorporate our coefficient-function approximation into the reduced-basis method. In both Sections 3 and 4 we present numerical results relevant to our model problem (1). In a future paper we provide further details, extend the method to (highly) nonlinear problems, and develop more realistic elliptic and parabolic examples.

2. Coefficient-function approximation: empirical interpolation

We are given a function $g(\cdot; \mu) \in L^\infty(\Omega)$ of sufficient regularity. To begin, we choose μ_1^g, and define $S_1^g = \{\mu_1^g\}$, $\xi_1 \equiv g(x; \mu_1^g)$, and $W_1^g = \text{span}\{\xi_1\}$; we assume that $\xi_1 \neq 0$. Then, for $M \geqslant 2$, we set $\mu_M^g = \arg\max_{\mu \in \Xi^g} \inf_{z \in W_{M-1}^g} \|g(\cdot; \mu) - z\|_{L^\infty(\Omega)}$, where Ξ^g is a suitably fine parameter sample over $\mathcal{D}$. We then set $S_M^g = S_{M-1}^g \cup \mu_M^g$, $\xi_M = g(x; \mu_M^g)$, and $W_M^g = \text{span}\{\xi_m,\ 1 \leqslant m \leqslant M\}$. Note that, thanks to our truth approximation, μ_M^g is the solution of a *standard linear program*.

We suppose that $M_{\max}$ is chosen such that the dimension of $\{g(\cdot; \mu) | \mu \in \mathcal{D}\}$ exceeds $M_{\max}$; we can then prove

Lemma 2.1. *For any $M \leqslant M_{\max}$, the space W_M^g is of dimension M.*

Proof. We first introduce some notation: $g_{M-1}^*(x; \mu) \equiv \arg\min_{z \in W_{M-1}^g} \|g(\cdot; \mu) - z\|_{L^\infty(\Omega)}$ and $\varepsilon_{M-1}^*(\mu) \equiv \|g(\cdot; \mu) - g_{M-1}^*(\cdot; \mu)\|_{L^\infty(\Omega)}$. It directly follows from our hypothesis on $M_{\max}$ that $\varepsilon_0 \equiv \varepsilon_{M_{\max}}^*(\mu_{M_{\max}+1}^g) > 0$; our 'arg max' construction then implies $\varepsilon_{M-1}^*(\mu_M^g) \geqslant \varepsilon_0$, $2 \leqslant M \leqslant M_{\max}$. We now prove Lemma 1 by induction. Clearly, $\dim(W_1^g) = 1$. Assume $\dim(W_{M-1}^g) = M - 1$; then if $\dim(W_M^g) \neq M$, $g(\cdot; \mu_M^g) \in W_{M-1}^g$; however, the latter contradicts $\varepsilon_{M-1}^*(\mu_M^g) \geqslant \varepsilon_0 > 0$. □

We now construct nested sets of interpolation points $T_M = \{t_1, \ldots, t_M\}$, $1 \leqslant M \leqslant M_{\max}$. We first set $t_1 = \arg\operatorname{ess\,sup}_{x \in \Omega} |\xi_1(x)|$, $q_1 = \xi_1(x)/\xi_1(t_1)$, $B_{11}^1 = 1$. Then for $M = 2, \ldots, M_{\max}$, we solve the linear system $\sum_{j=1}^{M-1} \sigma_j^{M-1} q_j(t_i) = \xi_M(t_i)$, $1 \leqslant i \leqslant M - 1$, and set $r_M(x) = \xi_M(x) - \sum_{j=1}^{M-1} \sigma_j^{M-1} q_j(x)$, $t_M = \arg\operatorname{ess\,sup}_{x \in \Omega} |r_M(x)|$, $q_M(x) = r_M(x)/r_M(t_M)$, and $B_{ij}^M = q_j(t_i)$, $1 \leqslant i, j \leqslant M$. It remains to demonstrate

Lemma 2.2. *The construction of the interpolation points is well-defined, and the functions $\{q_1, \ldots, q_M\}$ form a basis for W_M^g.*

Proof. We proceed by induction. Clearly, $W_1^g = \text{span}\{q_1\}$. Assume $W_{M-1}^g = \text{span}\{q_1, \ldots, q_{M-1}\}$; if (i) B^{M-1} is invertible, and (ii) $|r_M(t_M))| > 0$, then our construction may proceed and we may form $W_M^g = \text{span}\{q_1, \ldots, q_M\}$. To prove (i), we need only note that B^{M-1} is lower triangular with unity diagonal; to prove (ii), we observe that $|r_M(t_M)| \geqslant \varepsilon_{M-1}^*(\mu_M^g) \geqslant \varepsilon_0 > 0$. □

Lemma 2.3. *For any M-tuple $(\alpha_i)_{i=1,\ldots,M}$ of real numbers, there exists a unique element $w \in W_M^g$ such that $\forall i$, $1 \leqslant i \leqslant M$, $w(t_i) = \alpha_i$.*

Proof. It is a straightforward consequence of the invertibility of B^M. □

Finally, our coefficient function approximation is the interpolant of g over T_M as defined from Lemma 2.3: $g_M(x;\mu) = \sum_{m=1}^M \beta_m(\mu) q_m(x)$, where $\sum_{j=1}^M B_{ij}^M \beta_j(\mu) = g(t_i;\mu)$, $1 \leqslant i \leqslant M$. We define $\varepsilon_M(\mu) \equiv \|g(\cdot;\mu) - g_M(\cdot;\mu)\|_{L^\infty(\Omega)}$.

3. Error analyses for the empirical interpolation procedure

3.1. A priori framework

We define a 'Lebesgue constant' [7] $\Lambda_M = \sup_{x\in\Omega} \sum_{m=1}^M |V_m^M(x)|$, where V_m^M is the only element in W_M^g such that $V_m^M(t_n) = \delta_{mn}$ (the V_m^M are the characteristic functions as defined from Lemma 2.3). Note that Λ_M depends on W_M^g and T_M, but not on μ nor on our choice of basis for W_M^g. Observe also that $\sum_{j=1}^M B_{ji}^M V_j^M(x) = q_i(x)$, $1 \leqslant i \leqslant M$. We can then prove

Lemma 3.1. *The interpolation error $\varepsilon_M(\mu)$ satisfies $\varepsilon_M(\mu) \leqslant \varepsilon_M^*(\mu)(1 + \Lambda_M)$, $\forall \mu \in \mathcal{D}$.*

Proof. We define $e_M^*(x;\mu) = g(x;\mu) - g_M^*(x;\mu)$ and $g_M(x;\mu) - g_M^*(x;\mu) = \sum_{m=1}^M \delta_m^M(\mu) q_m(x)$. We then readily derive that $e_M^*(t_i;\mu) = \sum_{m=1}^M \delta_m^M(\mu) q_m(t_i) = \sum_{m=1}^M B_{im}^M \delta_m^M(\mu)$, $1 \leqslant i \leqslant M$. It thus follows that $|\varepsilon_M(\mu) - \varepsilon_M^*(\mu)| \leqslant \|\sum_{m=1}^M \delta_m^M(\mu) q_m(x)\|_{L^\infty(\Omega)} = \|\sum_{k=1}^M \sum_{m=1}^M B_{km}^M \delta_m^M(\mu) V_k^M(x)\|_{L^\infty(\Omega)} = \|\sum_{i=1}^M e_M^*(t_i;\mu) \times V_i^M(x)\|_{L^\infty(\Omega)} \leqslant \varepsilon_M^*(\mu)\Lambda_M$, since $|e_M^*(t_i;\mu)| \leqslant \varepsilon_M^*(\mu)$, $1 \leqslant i \leqslant M$. □

We can further show

Proposition 3.2. *The Lebesgue constant Λ_M satisfies $\Lambda_M \leqslant 2^M - 1$.*

Proof. We need only note that (i) B^M is lower triangular with unity diagonal — $q_m(t_m) = 1$, $1 \leqslant m \leqslant M$, and (ii) all entries of B^M are of modulus no greater than unity — $\|q_m\|_{L^\infty(\Omega)} \leqslant 1$, $1 \leqslant m \leqslant M$. Hence $|V_m^M(x)| \leqslant |q_m(x)| + \sum_{i=m+1}^M |V_i^M(x)| \leqslant 1 + \sum_{i=m+1}^M |V_i^M(x)|$. It follows, since $|V_M^M(x)| \leqslant 1$, that $|V_{M+1-m}^M(x)| \leqslant 2^{m-1}$, $1 \leqslant m \leqslant M$, and thus $\sum_{m=1}^M |V_m^M(x)| \leqslant 2^M - 1$. □

Proposition 3.2 is very pessimistic and of little practical value (though $\varepsilon_M^*(\mu)$ does often converge sufficiently rapidly that $\varepsilon_M^*(\mu)2^M \to 0$ as $M \to \infty$); this is not surprising given analogous results in the theory of polynomial interpolation [7]. However, Proposition 3.2 does provide some notion of stability.

3.2. A posteriori estimators

Given an approximation $g_M(x;\mu)$ for $M \leqslant M_{\max} - 1$, we define $\mathcal{E}_M(x;\mu) \equiv \hat{\varepsilon}_M(\mu) q_{M+1}(x)$, where $\hat{\varepsilon}_M(\mu) \equiv |g(t_{M+1};\mu) - g_M(t_{M+1};\mu)|$. We can then prove

Table 1
$\varepsilon^*_{M,\max}$, $\bar{\rho}_M$, Λ_M, and $\bar{\eta}_M$ as a function of M

M	$\varepsilon^*_{M,\max}$	$\bar{\rho}_M$	Λ_M	$\bar{\eta}_M$
4	2.65E–01	0.64	1.79	1.79
8	4.20E–02	0.65	2.07	2.01
12	8.66E–03	0.54	3.14	2.23
16	1.45E–03	0.85	2.09	2.62
20	1.85E–04	0.46	3.57	2.10

Proposition 3.3. *If* $g(\cdot;\mu) \in W^g_{M+1}$, *then* (i) $g(x;\mu) - g_M(x;\mu) = \pm\mathcal{E}_M(x;\mu)$ (*either* $\mathcal{E}_M(x;\mu)$ *or* $-\mathcal{E}_M(x;\mu)$), *and* (ii) $\|g(\cdot;\mu) - g_M(\cdot;\mu)\|_{L^\infty(\Omega)} \leqslant \hat{\varepsilon}_M(\mu)$.

Proof. Since by assumption $g(\cdot;\mu) \in W^g_{M+1}$, $g(x;\mu) - g_M(x;\mu) = \sum_{m=1}^{M+1} \kappa_m q_m(x)$. We may thus consider the linear system $\sum_{m=1}^{M+1} \kappa_m q_m(t_i) = g(t_i;\mu) - g_M(t_i;\mu)$, $1 \leqslant i \leqslant M+1$. However, $g(t_i;\mu) - g_M(t_i;\mu) = 0$, $1 \leqslant i \leqslant M$; thus, since the matrix $q_m(t_i)$ is lower triangular, $\kappa_m = 0$, $1 \leqslant m \leqslant M$, and since $q_{M+1}(t_{M+1}) = 1$, $\kappa_{M+1} = g(t_{M+1};\mu) - g_M(t_{M+1};\mu)$; this concludes the proof of (i). The proof of (ii) then directly follows from $\|q_{M+1}\|_{L^\infty(\Omega)} = 1$. □

Of course, in general $g(\cdot;\mu) \notin W^g_{M+1}$, and hence our estimator $\hat{\varepsilon}_M(\mu)$ is not quite a rigorous upper bound; however, if $\varepsilon_M(\mu) \to 0$ very fast, we expect that the effectivity, $\eta_M(\mu) \equiv \hat{\varepsilon}_M(\mu)/\varepsilon_M(\mu)$, shall be close to unity. Furthermore, the estimator is very inexpensive – *one additional evaluation* of $g(\cdot;\mu)$.

3.3. Numerical results

We consider $g(x;\mu) \equiv \mathcal{V}((x_1,x_2);(\mu_1,\mu_2)) \equiv ((x_1-\mu_1)^2 + (x_2-\mu_2)^2)^{-1/2}$ for $x \in \Omega \equiv]0,1[^2$ and $\mu \in \mathcal{D} \equiv [-1,-0.01]^2$; we choose for Ξ^g a random sample of 225 parameter points; and we take $\mu^g_1 = (-0.01,-0.01)$. We then construct S^g_M, W^g_M, T_M, and B^M, $1 \leqslant M \leqslant M_{\max}$, following the procedure of Section 2. We introduce a random parameter test sample Ξ^g_{Test} of size $Q_{\text{Test}} = 121$, and define $\varepsilon^*_{M,\max} = \max_{\mu\in\Xi^g_{\text{Test}}} \varepsilon^*_M(\mu)$, $\bar{\rho}_M = Q^{-1}_{\text{Test}} \sum_{\mu\in\Xi^g_{\text{Test}}} (\varepsilon_M(\mu)/(\varepsilon^*_M(\mu)(1+\Lambda_M)))$, $\bar{\eta}_M = Q^{-1}_{\text{Test}} \sum_{\mu\in\Xi^g_{\text{Test}}} \eta_M(\mu)$. We present in Table 1 $\varepsilon^*_{M,\max}$, $\bar{\rho}_M$, Λ_M, and $\bar{\eta}_M$ as a function of M ($M_{\max} = 20$). We observe that $\varepsilon^*_{M,\max}$ converges rapidly with M; that the Lebesgue constant provides a reasonably sharp measure of the interpolation-induced error; that the Lebesgue constant grows very slowly — $\varepsilon_M(\mu)$ is *only slightly larger that the min max result* $\varepsilon^*_M(\mu)$; and that the error estimator effectivity is reasonably close to unity. (Note also that B^M is quite well-conditioned given our choice of basis.)

4. Reduced-basis approximation

We consider the following model problem: Given $\mu \in \mathcal{D} \equiv [-1,-0.01]^2$, find $u(\mu) \in X$ such that $a(u,v;\mu) = f(v;\mu)$, $\forall v \in X$. Here $\Omega =]0,1[^2$; $X = H^1_0(\Omega)$; $(w,v)_X = \int_\Omega \nabla w \cdot \nabla v$; a is the bilinear form (1) for $g(x;\mu) \equiv \mathcal{V}(x;\mu)$; and $f(v;\mu) = \int_\Omega \mathcal{V}(x;\mu)v$. The solution develops a boundary layer in the vicinity of $x = (0,0)$ for μ near the 'corner' $(-0.01,-0.01)$. For our output, we consider $s(\mu) = \ell(u(\mu))$ for $\ell(v) = \int_\Omega v$.

Our reduced-basis approximation is thus: Given $\mu \in \mathcal{D}$, evaluate $s_{N,M}(\mu) = \ell(u_{N,M}(\mu))$, where $u_{N,M}(\mu) \in W^u_N$ is the solution of $\int_\Omega \nabla u_{N,M}(\mu) \cdot \nabla v + \int_\Omega g_M(x;\mu) u_{N,M}(\mu) v = \int_\Omega g_M(x;\mu) v$, $\forall v \in W^u_N$. Here W^u_N is defined in Section 1, and $g_M(x;\mu) = \sum_{m=1}^M \beta_m(\mu) q_m(x)$ is our coefficient-function approximation defined in Section 2 and analyzed in Section 3. Our discrete equations for $u_{N,Mj}(u_{N,M}(\mu) = \sum_{j=1}^N u_{N,Mj}(\mu)\zeta_j)$ are therefore $\sum_{j=1}^N (\int_\Omega \nabla\zeta_j \cdot \nabla\zeta_i + \sum_{m=1}^M \int_\Omega \beta_m(\mu) q_m(x) \zeta_j \zeta_i) u_{N,Mj} = \sum_{m=1}^M \int_\Omega \beta_m(\mu) q_m(x) \zeta_i$, $1 \leqslant i \leqslant N$. It is now a simple matter to develop an offline–online computational procedure: the online complexity is $O(N^2M) + O(N^3)$

Table 2
$\varepsilon^u_{N,M,\max}$ and $\bar{\eta}^u_{N,M}$ as a function of N (for $M = N$)

N	4	8	12	16	20
$\varepsilon^u_{N,M,\max}$	9.70E–02	5.53E–03	1.76E–03	4.53E–04	2.71E–05
$\bar{\eta}^u_{N,M}$	2.02	3.46	3.11	3.14	5.28

to respectively assemble and solve the requisite stiffness system and then $O(N)$ to evaluate $s_{N,M}(\mu)$; the essential point is that the online complexity is independent of $\mathcal{N}$.

It is readily demonstrated that the error $e_{N,M}(\mu) = u(\mu) - u_{N,M}(\mu)$ satisfies $\int_\Omega \nabla e_{N,M}(\mu)\cdot\nabla v + \int_\Omega g(x;\mu) \times e_{N,M}(\mu)v = R_{N,M}(v;\mu) + \int_\Omega (g(x;\mu) - g_M(x;\mu))v - \int_\Omega (g(x;\mu) - g_M(x;\mu))u_{N,M}(\mu)v$, $\forall v \in X$, where $R_{N,M}(v;\mu) \equiv \int_\Omega g_M(x;\mu)v - \int_\Omega \nabla u_{N,M}(\mu)\cdot\nabla v - \int_\Omega g_M(x;\mu)u_{N,M}(\mu)v$. It follows that, if we suppose $g(x;\mu) \in W^g_{M+1}$, then $\|e_{N,M}(\mu)\|_X \leqslant \Delta_{N,M}(\mu)$, where $\Delta_{N,M}(\mu) \equiv \hat{\varepsilon}_M(\mu)\sup_{v\in X} \frac{\int_\Omega q_{M+1}(x)(1-u_{N,M}(\mu))v}{\|v\|_X} + \sup_{v\in X} \frac{R_{N,M}(v;\mu)}{\|v\|_X}$. (Note an associated error bound on $s(\mu) - s_{N,M}(\mu)$ can be readily developed from standard duality considerations [6].) It is now possible [6] to develop an offline–online computational procedure for $\Delta_{N,M}(\mu)$: the online complexity to evaluate the requisite dual norms is $O(N^2M^2)$ – independent of $\mathcal{N}$. (We may invoke these inexpensive error estimators to develop good samples S^u_N: given S^u_{N-1}, we choose μ^u_N to be the arg max over (a fine sample in) $\mathcal{D}$ of $\Delta_{N,M_{\max}}(\mu)$ [8].)

We now introduce a random parameter test sample Ξ^u_{Test} of size $Q^u_{\text{Test}} = 289$, and define $\varepsilon^u_{N,M,\max} = \max_{\mu\in\Xi^u_{\text{Test}}} \|e_{N,M}(\mu)\|_X$ and $\bar{\eta}^u_{N,M} = (Q^u_{\text{Test}})^{-1}\sum_{\mu\in\Xi^u_{\text{Test}}} (\Delta_{N,M}(\mu)/\|e_{N,M}(\mu)\|_X)$. We present in Table 2 $\varepsilon^u_{N,M,\max}$ and $\bar{\eta}^u_{N,M}$ as a function of N for the particular choice $M = N$. We observe that the error decreases very rapidly, and that our error bound is quite sharp. Indeed, the results are largely indistinguishable from the standard Galerkin projection. However, the latter suffers from $O(\mathcal{N})$ online complexity, and is thus much more expensive than the coefficient-function approximation/empirical interpolation approach developed in this paper.

Acknowledgements

This work was supported by DARPA and AFOSR under Grant F49620-03-1-0356 and the Singapore-MIT Alliance. Our current project with Éric Cancès, Claude Le Bris, and Gabriel Turinici on the definition of reduced-basis strategies for the (highly nonlinear) Hartree Fock equations is certainly a natural candidate for the application of this 'empirical interpolation' method; we would like to thank this group for many stimulating and beneficial exchanges.

References

[1] B.O. Almroth, P. Stern, F.A. Brogan, Automatic choice of global shape functions in structural analysis, AIAA J. 16 (1978) 525–528.
[2] J.P. Fink, W.C. Rheinboldt, On the error behavior of the reduced basis technique for nonlinear finite element approximations, Z. Angew. Math. Mech. 63 (1983) 21–28.
[3] L. Machiels, Y. Maday, I.B. Oliveira, A.T. Patera, D.V. Rovas, Output bounds for reduced-basis approximations of symmetric positive definite eigenvalue problems, C. R. Acad. Sci. Paris, Ser. I 331 (2) (2000) 153–158.
[4] Y. Maday, A.T. Patera, G. Turinici, Global a priori convergence theory for reduced-basis approximation of single-parameter symmetric coercive elliptic partial differential equations, C. R. Acad. Sci. Paris, Ser. I 335 (3) (2002) 289–294.
[5] A.K. Noor, J.M. Peters, Reduced basis technique for nonlinear analysis of structures, AIAA J. 18 (4) (1980) 455–462.
[6] C. Prud'homme, D. Rovas, K. Veroy, Y. Maday, A.T. Patera, G. Turinici, Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods, J. Fluids Engrg. 124 (1) (2002) 70–80.
[7] A. Quarteroni, R. Sacco, F. Saleri, Numer. Math., Texts Appl. Math., vol. 37, Springer, New York, 1991.
[8] K. Veroy, C. Prud'homme, D.V. Rovas, A.T. Patera, A Posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations, in: Proceedings of the 16th AIAA Computational Fluid Dynamics Conference, June 2003, AIAA Paper 2003-3847.

5.9 Figures

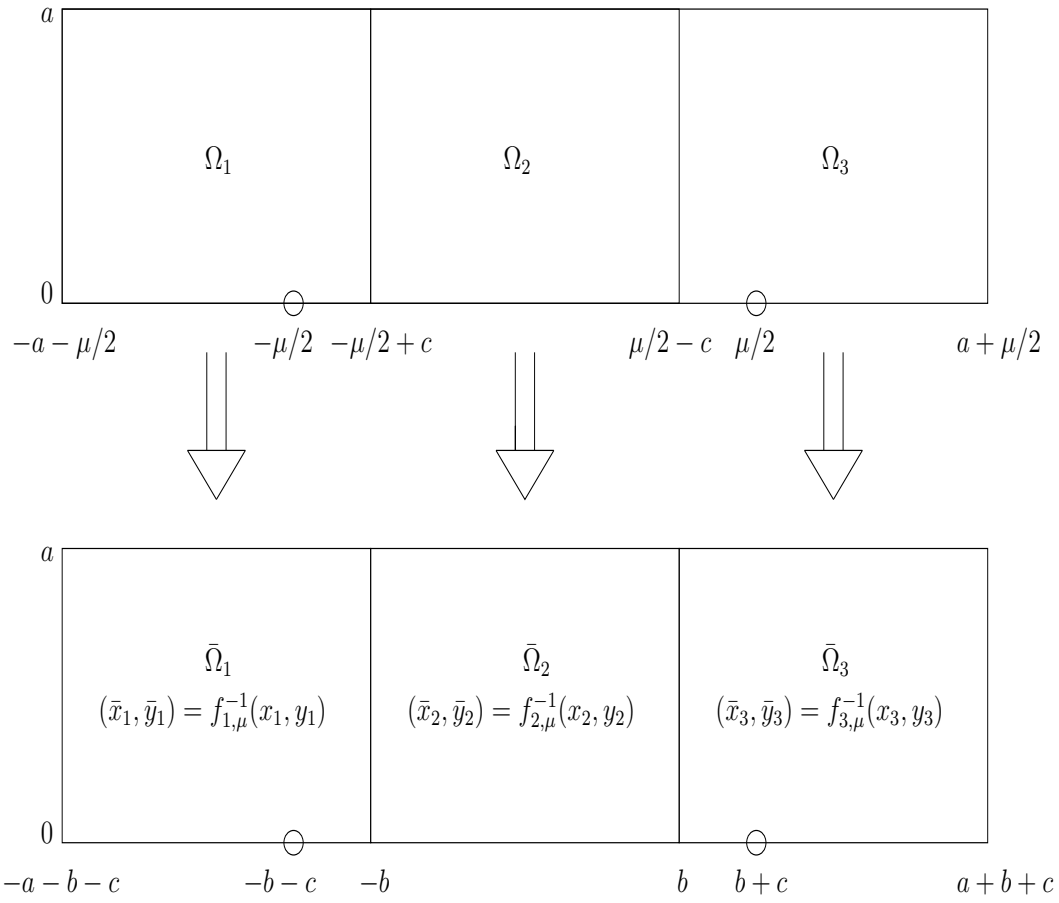
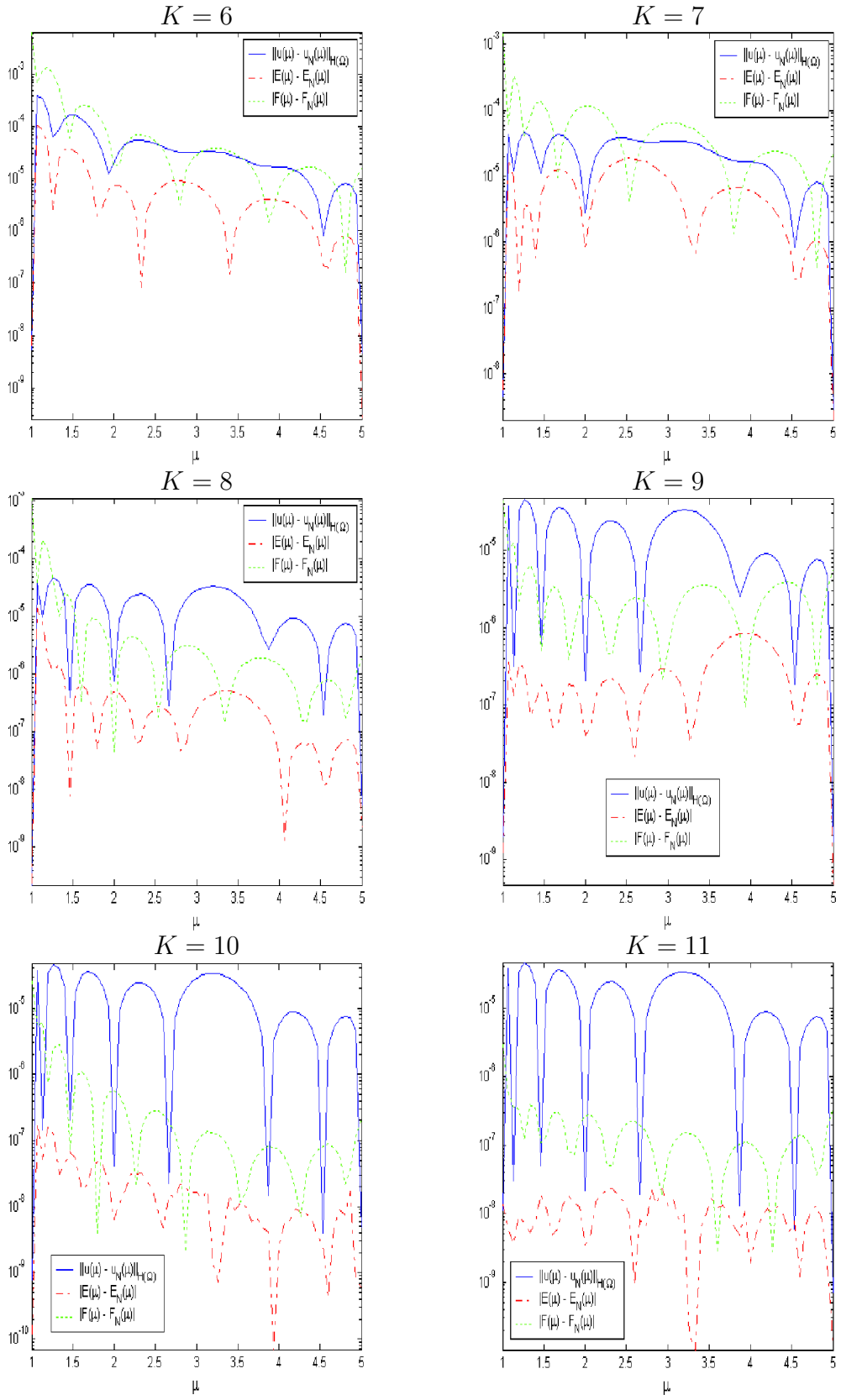
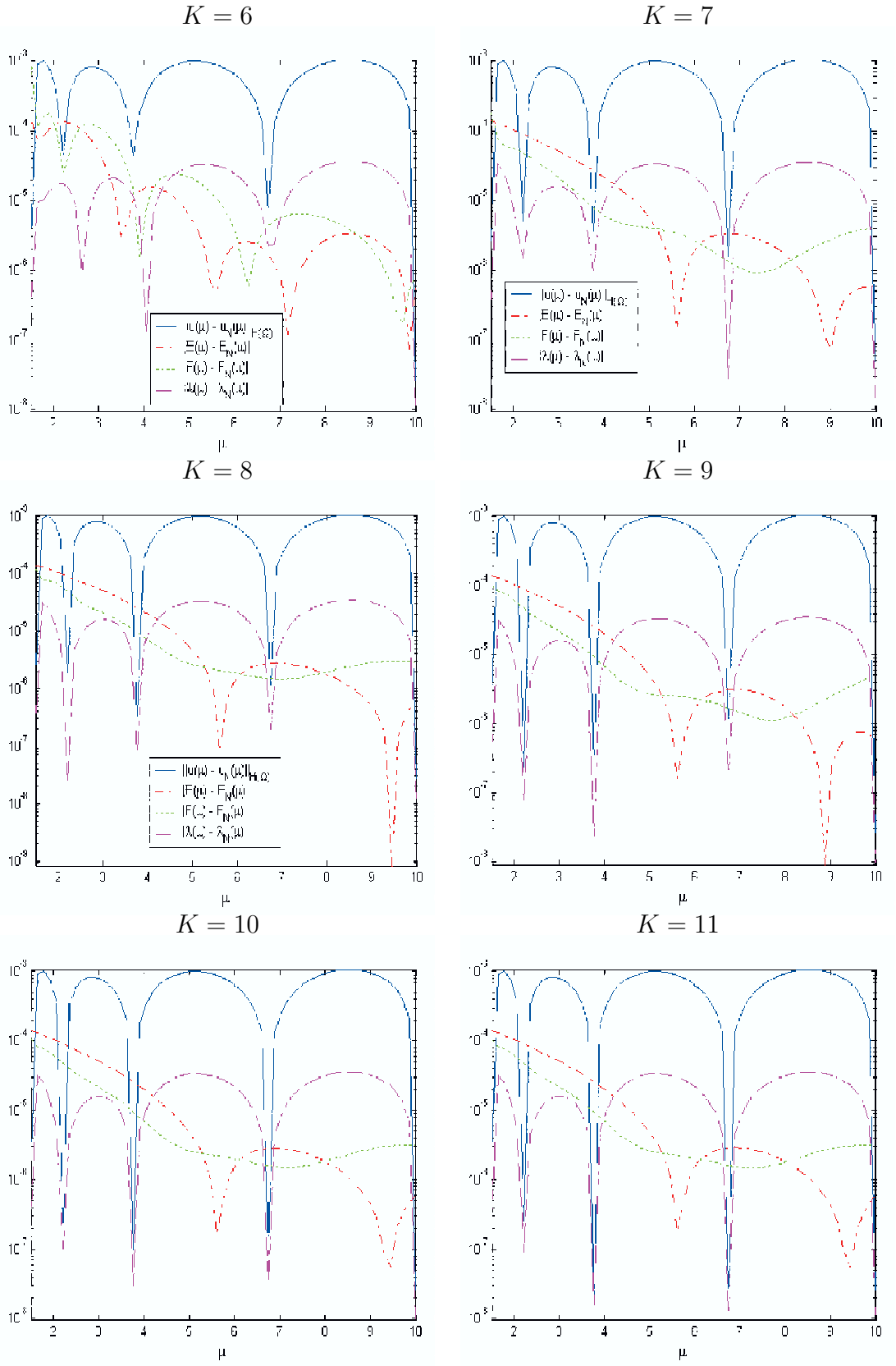


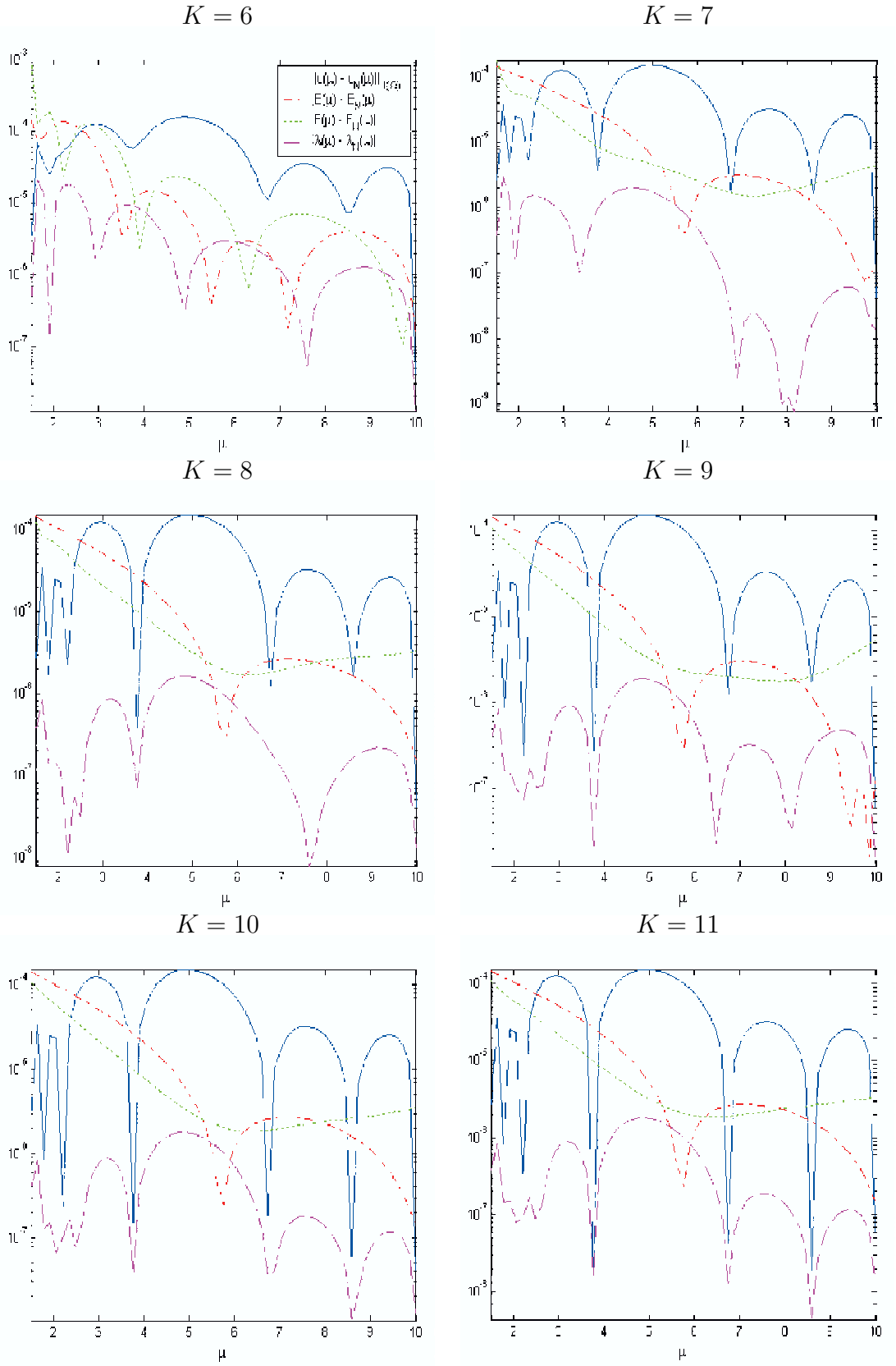
FIG. 5.1 – Transformations inverses $f_{i,\mu}^{-1}$ pour les systèmes H_2^+ et H_2



TAB. 5.1 – Evolution des erreurs e_E, e_F et e_u pour $m = 8$ et différentes valeurs de K.

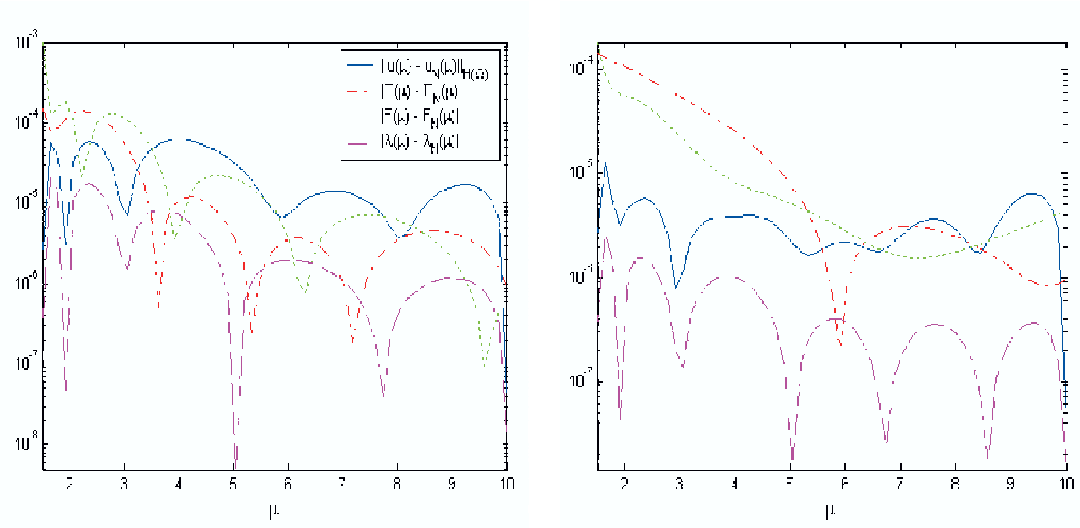


TAB. 5.2 – Evolution des erreurs e_E, e_F, e_λ, et e_u pour $m = 5$ et différentes valeurs de K.

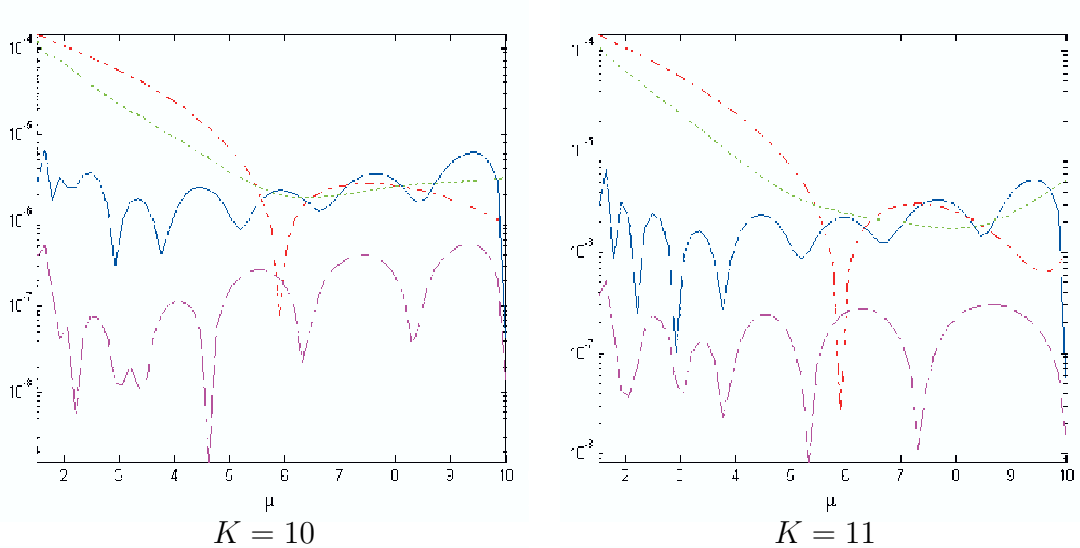


TAB. 5.3 – Evolution des erreurs e_E, e_F, e_λ, et e_u pour $m = 7$ et différentes valeurs de K.

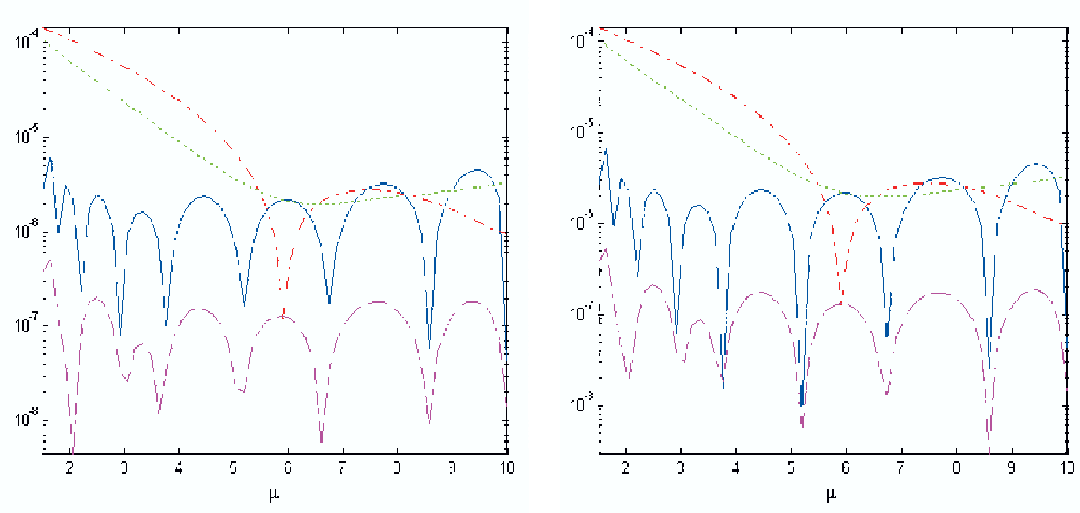
$K = 6$ $K = 7$



$K = 8$ $K = 9$



$K = 10$ $K = 11$



TAB. 5.4 – Evolution des erreurs e_E, e_F, e_λ, et e_u pour $m = 9$ et différentes valeurs de K.

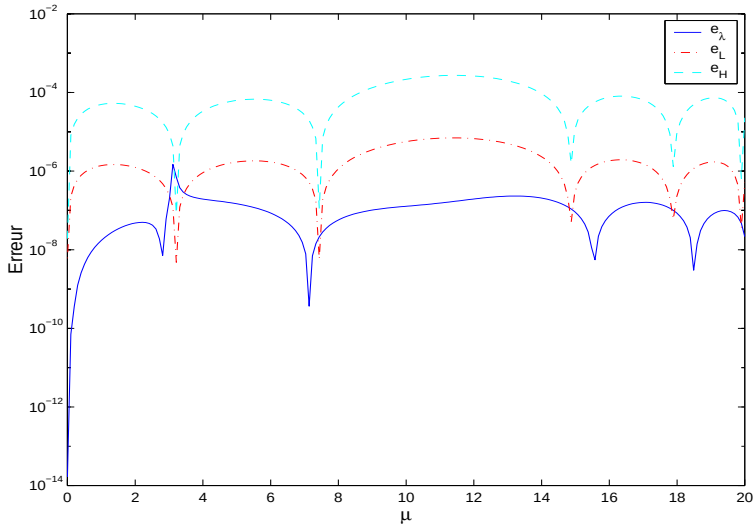


FIG. 5.2 – Erreurs commises par l'algorithme base réduite pour $(m, K) = (6, 9)$.

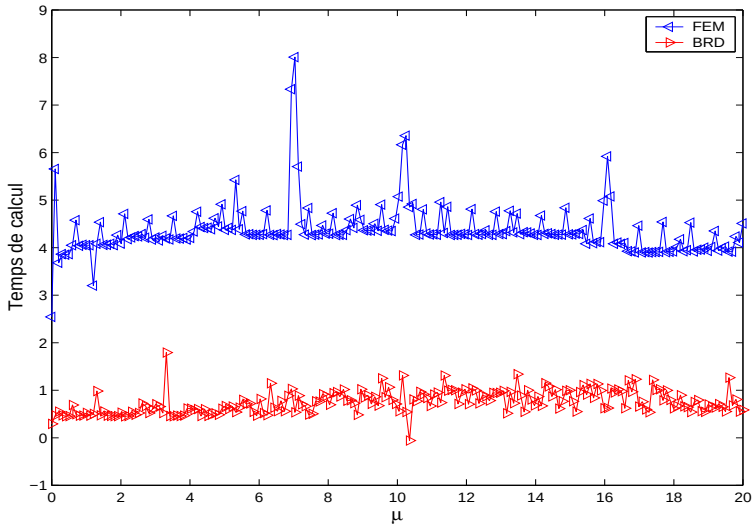


FIG. 5.3 – Temps de calcul des solutions fines (FEM) et base réduite (BRD).

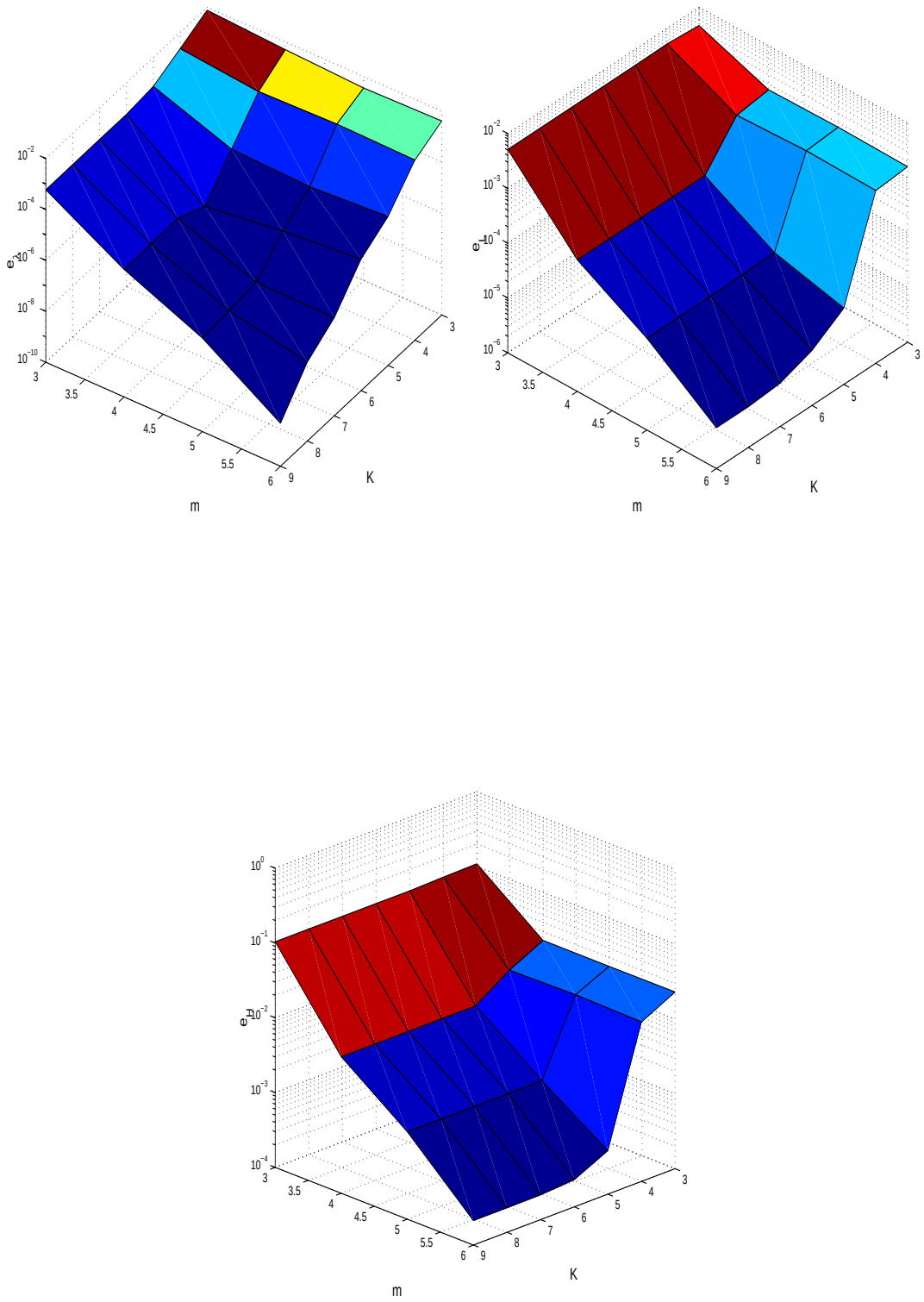


FIG. 5.4 – Erreur commises pour différentes valeurs de (m, K) pour $\mu = 10$.

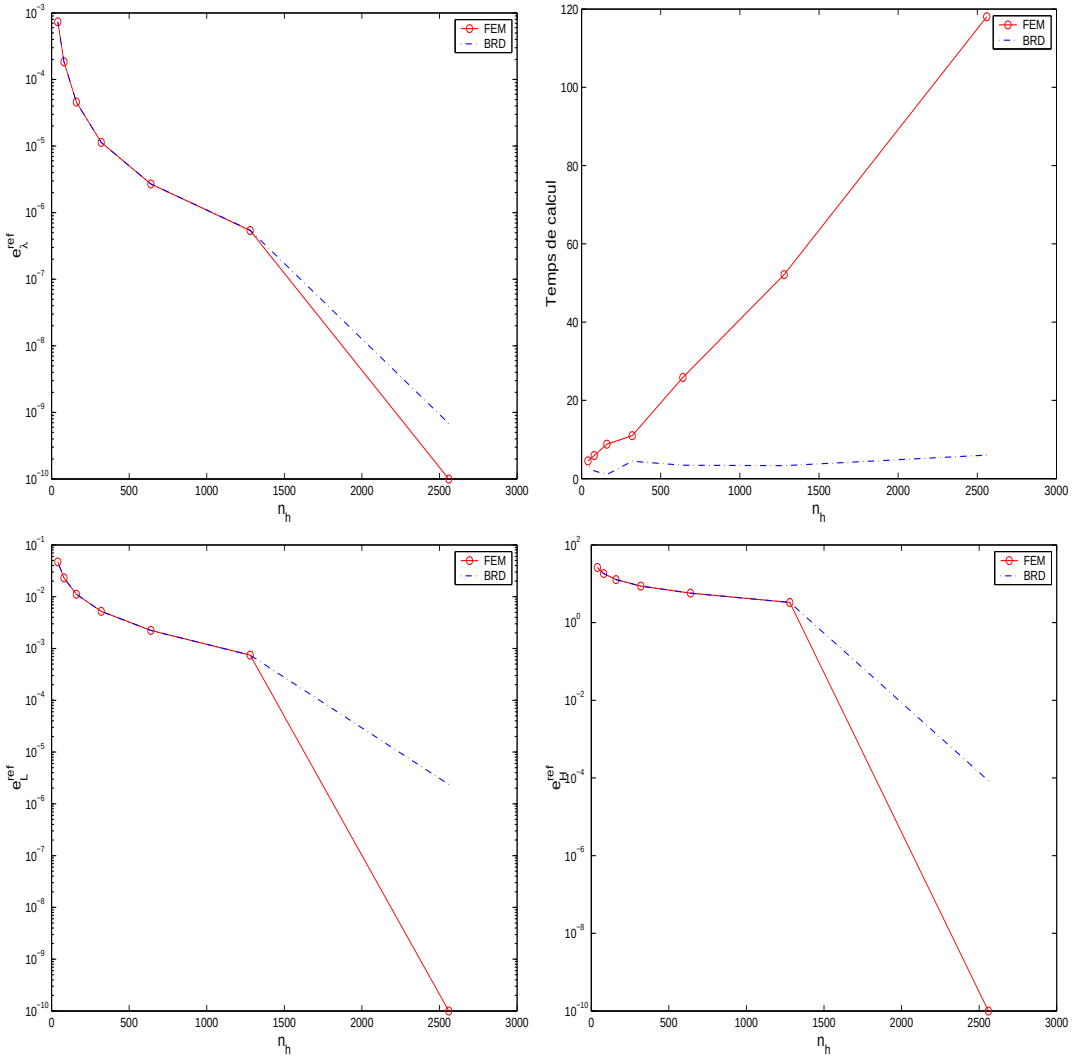


FIG. 5.5 – Erreurs et temps de calcul associés aux solutions fine (FEM) et base réduite (BRD) pour $\mu = 10$.

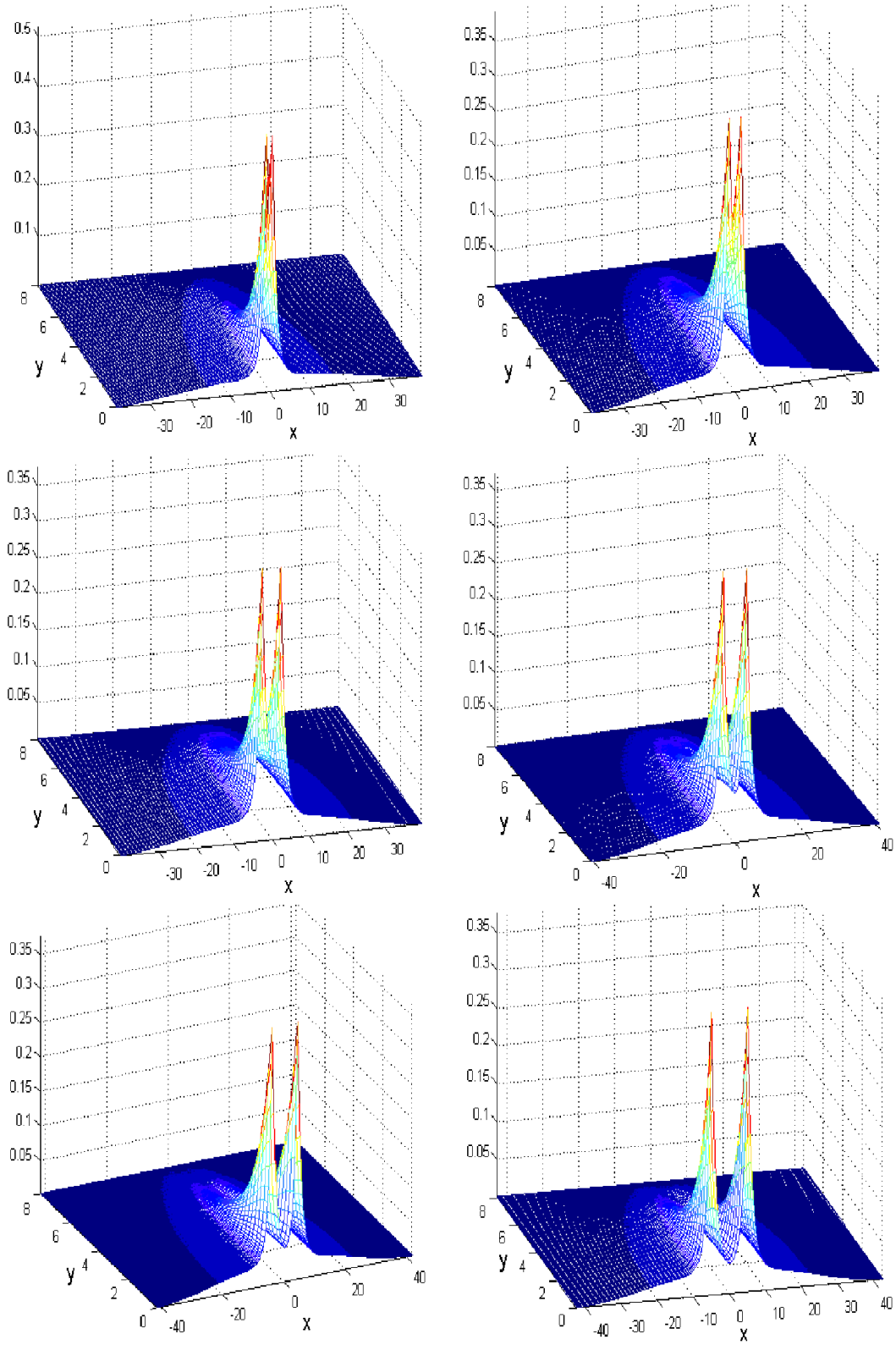


FIG. 5.6 – Solution u_μ associée au problème sur H_2 pour plusieurs valeurs de μ.

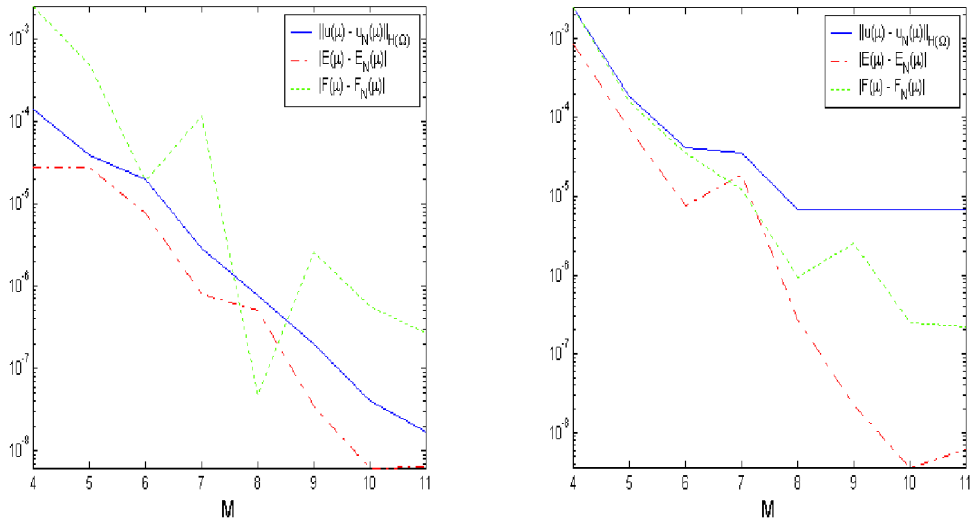


FIG. 5.7 – Evolution des erreurs e_E, e_F et e_u suivant K (M sur les figures) pour $\mu = 2$ (à gauche) et $\mu = 2.6$ (à droite).

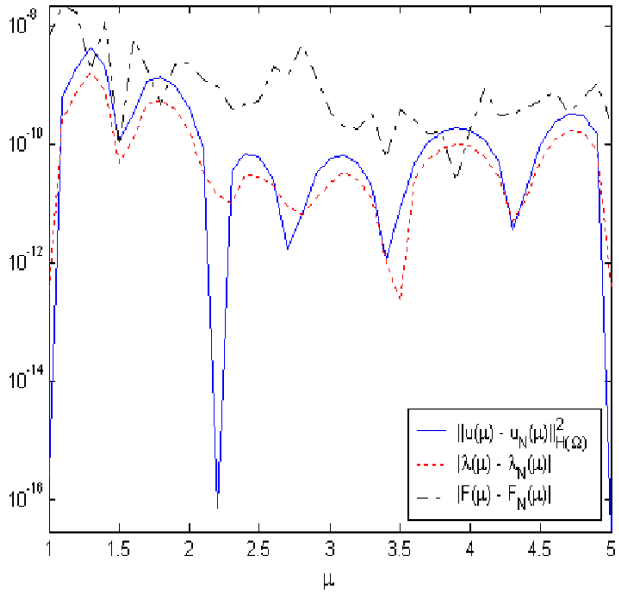


FIG. 5.8 – Evolution des erreurs e_E, e_F et e_u pour $(m, K) = (8, 8)$ en prenant en compte le potentiel des noyaux de façon exacte.

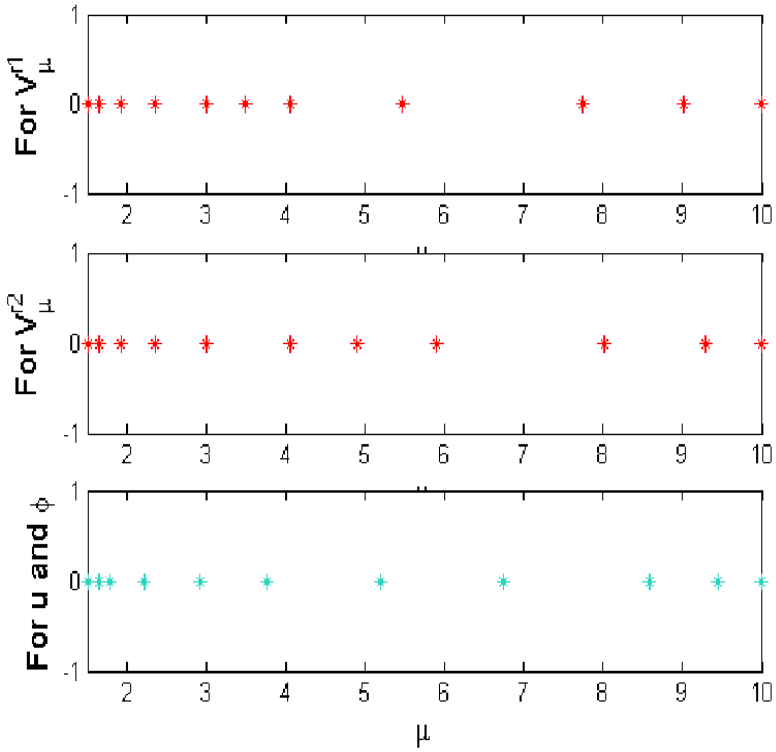


FIG. 5.9 – Jeux des paramètres définissant les bases réduites sur $\tilde{u}$ (u), $\tilde{\phi}$ (ϕ), g_1 (V_μ^{r1}) et g_2 (V_μ^{r2}).

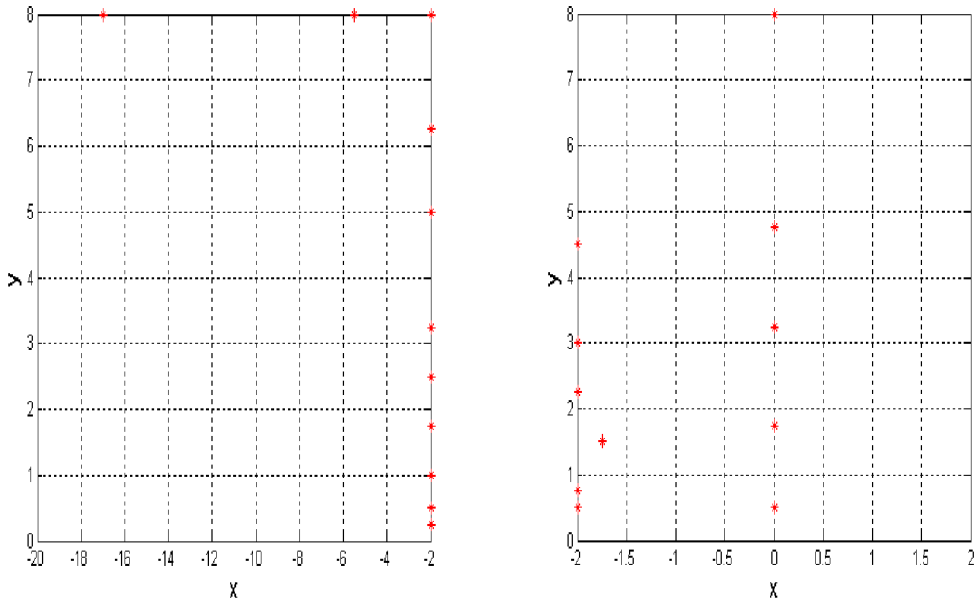


FIG. 5.10 – *Magic points* associés à g_1 (à gauche) et g_2 (à droite).

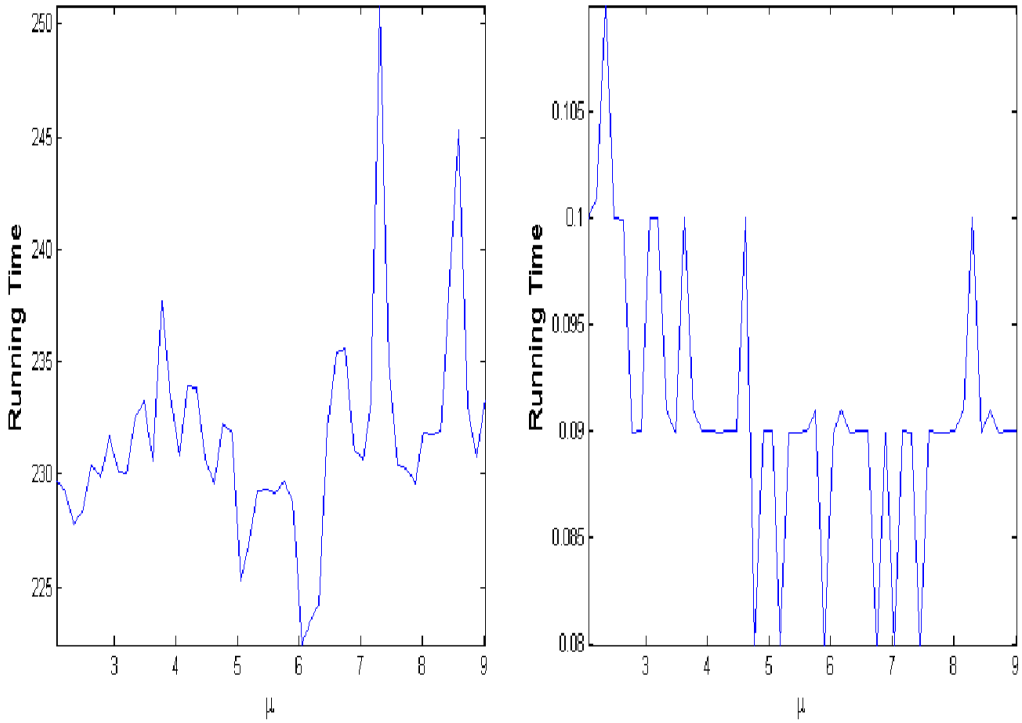


FIG. 5.11 – Evolution des temps de calcul de $\bar{u}_\mu$ (à gauche) et $\tilde{u}_\mu$ (à droite) suivant μ.

Chapitre 6

Conclusion générale

On résume dans ce dernier chapitre le travail réalisé dans la thèse. Après avoir tiré les conclusions de la thèse, on indique des pistes pour des travaux futurs.

6.1 Les pseudo-potentiels

La méthode des pseudo-potentiels repose sur des principes chimiques simples, mais d'une traduction mathématique peu claire. Même pour les systèmes les plus simples, la construction d'un pseudo-potentiel adapté ne peut pas être pour l'instant validée de façon rigoureuse. Pour des gros systèmes, le calcul tout électron n'est pas réalisable. La construction d'un potentiel valide pour ces systèmes nécessite donc la connaissance *a priori* des phénomènes physico-chimiques mis en jeu.
Ainsi, pour la description précise des systèmes dont on ne possède aucune connaissance *a priori*, la méthode des pseudo-potentiels n'est pas adaptée. Cependant, elle se révèle très précieuse pour isoler des tendances sur des systèmes d'un traitement *ab initio* précis hors de portée des moyens informatiques actuels. On peut citer l'exemple de la simulation de systèmes constitués d'atomes lourds.

La construction d'une approche systématique pour la génération de pseudo-potentiel semble, pour l'instant, vouée à l'échec. Un grand chemin reste à faire dans le domaine mathématique pour traduire la compréhension physico-chimique de façon précise. Dans cette optique, la méthode PAW offre un premier cadre formel pour l'étude de la construction des pseudo-potentiels. Des zones d'ombre restent toutefois, concernant la construction des opérateurs atomiques T_z et le passage aux systèmes moléculaires. Un premier travail pourrait s'articuler autour d'une étude de sensibilité des résultats au choix des rayons r_c^z, de la fonction $k(r)$ et du paramètre $\mathcal{V}_0$, et sur la prise en compte des termes supplémentaires liés à l'incomplétude de la base.
Des calculs sur des systèmes très simples comme les molécules H_2 ou Be_2 pourraient fournir des pistes de réflexion en vue d'une compréhension et d'une construction rigoureuses des pseudo-potentiels. L'introduction d'un pseudo-potentiel pour Be_2 (quatre orbitales *tout électron* à calculer) conduit à éliminer le calcul des deux orbi-

tales de plus basses énergie. En revanche, l'introduction d'un pseudo-potentiel pour H_2 permet seulement de régulariser les deux singularités présentées par l'orbitale *tout électron* à calculer. A l'autre extrême, l'implémentation de ces idées dans *Abinit* [109] pourrait permettre de les tester en vraie grandeur sur des systèmes réels.

6.2 La méthode de décomposition de domaine

Les résultats numériques obtenus sur des systèmes monodimensionnels réels (S et F bandes) avec la méthode présentée (MDD dans la suite) dans le chapitre 3 sont prometteurs. Il est important de noter que les tests ont été effectués sur un seul processeur. Toutefois, on a considéré l'algorithme sous sa forme parallélisée dans l'enchaînement interne des différentes étapes.

- On observe une complexité linéaire en terme de temps CPU des étapes locale et globale avec la dimension du système. Il s'ensuit une complexité linéaire de MDD dans sa globalité. Cette complexité est aussi observée pour notre implémentation de la méthode *D*ensity Matrix Minimization (DMM).
- Lorsque l'initial guess est mauvais, MDD converge très rapidement, pour les polymères considérés dans les tests, vers une solution précise (en terme d'énergie et de matrice densité), ce qui n'est pas le cas des méthodes variationnelles existantes comme DMM.
- Lorsque l'initial guess est de bonne qualité, DMM converge moins vite que MDD, mais vers une solution plus précise.
- MDD donne une bonne estimation du niveau de Fermi, donnée nécessaire aux algorithmes variationnels tel DMM.
- Les deux méthodes exhibent une complexité linéaire en terme de mémoire allouée. La mémoire requise par MDD augmente moins vite que celle requise par notre implémentation de DMM. Toutefois, au contraire de DMM, MDD nécessite une mémoire très importante pour des systèmes de petite taille.

La méthodologie développée semble donc apporter une solution au problème de l'obtention d'un initial guess de bonne qualité pour les méthodes variationnelles. Par ailleurs, MDD apporte une solution au traitement des systèmes de très grande taille qui requiert une mémoire trop importante lors de l'usage des méthodes variationnelles. Dans une perspective de dynamique moléculaire, cet apport devrait se vérifier d'autant plus que certaines zones du système étudié restent inchangées au cours de la simulation.

Toutefois, il est important de souligner que la comparaison de la méthode MDD avec la méthode DMM n'est pas équitable et en défaveur de MDD. Il n'est pas étonnant que la méthode DMM converge vers une solution plus précise puisque l'espace de minimisation est plus grand. Une comparaison avec une méthode de type OM serait plus pertinente. Notre implémentation de la méthode OM, implémentée dans SIESTA, a mené à des résultats sur des cas plus simples similaires qualitativement, mais moins bons quatitativement, à ceux obtenus avec DMM. Toutefois, on

a identifié l'extrême difficulté de générer un bon initial guess. Notamment, la solution obtenue par MDD ne s'avère pas toujours suffisante. Une comparaison avec la méthode originelle implémentée dans SIESTA permettrait d'apporter une réponse plus précise quant à la pertinence de MDD comme préconditionneur de OM. Par ailleurs, la connaissance du niveau de Fermi a été supposée pour DMM, ce qui n'est pas le cas en pratique. Cette méconnaissance du système conduit en pratique à faire tourner l'algorithme pour plusieurs estimations du niveau de Fermi. Dans les tests réalisés, on obtient avec la méthode MDD une bonne estimation du niveau de Fermi en parallèle d'un bon initial guess. Par conséquent, la méthode MDD joue le rôle de préconditionneur idéal pour les méthodes variationnelles.

Enfin, ces résultats encourageants motivent des investigations numériques supplémentaires sur le comportement de la méthode développée dans cette thèse (parallélisation, cas d'un traitement plus difficiles (S et F creuses) par les méthodes existantes). D'un point de vue algorithmique, les choix réalisés pour le traitement des étapes locales et globales de MDD sont loin d'être optimaux. Un traitement plus adapté de ces étapes conduirait à une performance accrue de MDD. Des investigations théoriques sont aussi nécessaires pour prouver la convergence de MDD, et obtenir une estimation de sa vitesse en fonction de γ, dans un cadre simplifié dans un premier temps.

6.3 Le passage aux systèmes métalliques

L'adaptation des méthodes de projection proposée dans le chapitre 4 constitue une étape intéressante vers des méthodes de complexité $\mathcal{O}\big(N^\alpha\big)$ avec $\alpha < 3$ pour le traitement des systèmes métalliques.

D'un point de vue plus général, il semble difficile d'atteindre un tel objectif si la base des χ_μ est exclusivement localisée. En effet, un système métallique comporte $\mathcal{O}(N)$ orbitales de Wannier ψ_i^l localisées et $\mathcal{O}(N)$ *orbitales de Wannier délocalisées* ψ_i^d. Par conséquent, il est difficile d'approcher les ψ_i^d par une combinaison linéaire de quelques fonctions χ_μ localisées. Une démarche possible pour le traitement d'un tel système consiste à traiter *de façon différente* chaque type d'orbitales et donc de repenser la modélisation du problème électronique. Dans cette logique, une voie non explorée jusqu'à maintenant consiste à utiliser des fonctions localisées χ_μ^l pour approcher les ψ_i^l et à remplacer le calcul des ψ_i^d par une densité délocalisée ρ^d.
Une autre voie, basée sur le calcul des orbitales, consiste à utiliser une base de fonctions délocalisées χ_μ^d telle que chaque ψ_i^d s'écrive comme une combinaison linéaire de quelques χ_μ^d. Par suite, la matrice C^d des coefficients des ψ_i^d dans la base χ_μ^d est creuse, et donc n'exclut pas *a priori* une complexité linéaire. Dans un tel contexte, le problème électronique se ramène à la résolution itérative de deux problèmes couplés (par la densité totale ρ) posés sur les ψ_i^l et les ψ_i^d. A chaque itération, on résout

- un problème aux valeurs propres généralisé *localisé*,
- un problème aux valeurs propres généralisé *délocalisé* caractérisé par une ma-

trice F^d, certes pleine, $S^d = I_d$ (comme pour la base d'ondes planes) et une matrice C^d creuse.

Dans un tel contexte, la méthode de décomposition de domaine présentée est pertinente comme elle nécessite l'existence d'une matrice C localisée (mais aussi la connaissance *a priori* de sa structure) et non le caractère creux de la matrice F à l'inverse des méthodes existantes.

6.4 La méthode des bases réduites

La stratégie générale présentée au chapitre 5 pour le traitement des termes non linéaires et non affines se révèle pertinente pour le traitement des systèmes H_2^+ et H_2. On constate cependant que la dimension de la base réduite obtenue est d'une taille similaire à celle des bases localisées utilisées en pratique (bases de gaussiennes) pour une même précision finale. Soulignons que les bases localisées caractérisées par des petites tailles $\big(N_b = \mathcal{O}(N)\big)$ peuvent s'interpréter comme des bases réduites en tant que combinaison linéaires d'orbitales atomiques (solutions de problèmes électroniques associés à des atomes isolés).

Dans le contexte d'une méthode de bases réduites basée sur l'approximation des ϕ_i, seule une amélioration du préfacteur des méthodes de complexité linéaire semble envisageable pour une base localisée. En revanche, l'application de la méthode des bases réduites suite à une discrétisation par ondes planes et à l'introduction d'un pseudo-potentiel pourrait certainement mener à une amélioration de la complexité cubique avec N des méthodes de résolution du problème électronique.

L'obtention d'une méthode de bases réduites de complexité indépendante de N semble passer par le traitement d'une grandeur scalaire. L'association de cette méthode avec les modèles *Orbital Free* posés sur ρ [29] (en vue du traitement de problèmes de grande taille) ou avec l'équation de Schrödinger (pour des problèmes de très petite taille) constitue en ce sens des directions prometteuses.

Bibliographie

Références en Chimie Computationnelle

Ouvrages généraux

[1] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday (2003), *Computational Quantum Chemistry : a Primer*, in : Handbook of Numerical Analysis, Special volume, Computational Chemistry, volume X, C. Le Bris guest editor, Ph. G. Ciarlet Editor, North-Holland.

[2] M. Defranceschi and C. Le Bris, eds. (2000), *Mathematical Models and Methods for Ab Initio Quantum Chemistry*, Lecture Notes in Chemistry, 74, Springer.

[3] W.J. Hehre, L. Radom, P.v.R. Schleyer, and J.A. Pople (1986), *Ab Initio Molecular Orbital Theory*, Wiley.

[4] J. Kohanoff (1998), *Electronic Structure Calculations and First-Principles Molecular Dynamics simulations*, Courses, TSLS Group, Queen's University Belfast.

[5] R. McWeeny, *Methods of molecular quantum mechanics*, 2nd edition, Academic Press 1992.

[6] R.G. Parr and W. Yang (1989), *Density Functional Theory of Atoms and Molecules*, Oxford Univeristy Press.

[7] A. Szabo and N. Ostlund (1982), *Modern Quantum Chemistry : An Introduction to Advanced Electronic Structure Theory*, MacMillan.

Thèses

[8] J.L Fattebert (1997), *Une méthode numérique pour la résolution des problèmes aux valeurs propres liés au calcul de structure électronique moléculaire*, Thèse de l'Ecole Polytechnique Fédérale de Lausanne.

[9] C.M. Goringe (1995), *D. Phil Thesis*, Oxford University.

Articles

[10] S. Adhikari, E. Baer (2001), *Augmented lagrangian method for order-N electronic structure*, J. Chem. Phys. *115*, 11-14.

[11] O.K. Andersen (1975), *Linear methods in band theory*, Phys. Rev. B *12*, 3060-3083.

[12] E. Anglada, E. Artacho,J.M. Junquera and J.M. Soler (2002), *Systematic generation of finite-range atomic basis sets for linear-scaling calculations*, Phys. Rev. B *66*, 205101-205104.

[13] D.A. Areshkin, O.A. Shenderova, J.D. Schall and D.W. Brenner (2003), *Convergence acceleration scheme for self consistent orthogonal basis set electronic structure methods*, Mol. Sim. *29*, 269-286.

[14] N.W. Ashcroft and N. D. Mermin (1976), *Solid-State Physics*, Saunders College Publishing.

[15] G.B. Bachelet, D.R. Hamann and M. Schluter (1982), *Pseudopotentials that work : from H to Pu*, Physical Review B *26*, 4199-4228.

[16] R. Baer and M. Head-Gordon (1997), *Sparsity of the density matrix in Kohn-Sham density functional theory and an assessment of linear system-size scaling methods*, Phys.Rev. Lett. *79*, 3962-3965.

[17] R. Baer, M. Head-Gordon (1997), *Chebyshev expansion methods in calculations of electronic structure of large molecules*, J. Chem. Phys. *107*, 10003-10013.

[18] R. Baer and M. Head-Gordon (1998), *Energy renormalization-group method for electronic structure of large systems*, Phys. Rev. B *58*, 15296-15299.

[19] R. Baer et al. (2003), *Improved Fermi operator expansion methods for fast electronic calculations*, J. Chem. Phys. *119*, 4117-4125.

[20] Z. Barandiarán and L. Seijo (1999), *The ab initio model potential method : A common strategy for effective core potential and embedded cluster calculations*, in Computational Chemistry : Reviews of Current Trends, *4*, edited by J. Leszczynski, (Wold Scientific, Singapur), 55-152.

[21] S. Baroni, E. Car, M. Parrinello and I. Stich (1989), Conjugate gradient minimization of the energy functional : A new method for electronic structure calculation, Phys. Rev. B *39*, 4997-5004.

[22] P.E. Blöchl (1994), *Projector augmented-wave method*, Phys. Rev. B *50*, 17953-17979.

[23] D.R. Bowler et al. (1997), *A comparison of linear scaling tight binding methods*, Modelling Simul. Mater. Sci. Eng. *5*, 199-222.

[24] D. Bowler, T. Miyazaki and M. Gillan (2002), *Recent progress in linear scaling ab initio electronic structure theories*, J. Phys. Condens. Matter *14*, 2781-2798.

[25] S.F. Boys (1950), *Electronic wavefunction I. A general method of calculation for the stationary states of any molecular system*, Proc. Roy. Soc. A 200, 542-554.

[26] I.J. Bush, M.J. Gillan, C.M. Goringe and E. Hernández (1997), *Linear scaling DFT-pseudopotential calculations on parallel computers*, Comp. Phys. Commun *102*, 1-3.

[27] E. Cancès, K. Kudin and G. Scuseria (2002), *A black-box self-consistent field convergence algorithm : one step closer*, J. Chem. Phys. *116*, 8255-8261.

[28] R. Car, M. Parrinello (1985), *Structural dynamical and electronic properties of amorphous silicon : an ab initio molecular-dynamics study*, Physical Review Letters *55*, 2471-2474.

[29] E.A. Carter and Y.A. Wang (2000), *Orbital-free kinetic-energy density functional theory*, in : Theoretical methods in condensed phase chemistry, S.D. Schwartz (ed.), Kluwer.

[30] M. Challacombe (1999), *A simplified density matrix minimization for linear scaling self-consistent field theory*, J. Chem. Phys. *110*, 2332-2342.

[31] M. Challacombe (2000), *Linear scaling computation of the Fock matrix, V. Hierarchical cubature for numerical integration of the exchange-correlation matrix*, J. Chem. Phys. *113*, 10037-10043.

[32] J.R. Chelikowsky, Y. Saad and N. Trouiller (1994), *Finite difference pseudopotential method : electronic structure calculations without a basis*, Phys. Rev. Lett. *72*, 1240-1243.

[33] J.R. Chelikowsky, L.O. Jay, H. Kim, Y. Saad (1999), *Electronic structure calculations for plane-wave codes without diagonalization*, Computer Physics Communications *118*, 21-30.

[34] P.A. Christiansen (1990), *Ab initio relativistic effective potentials with spin orbit operators. IV. Cs through Rn*, J. Chem. Phys. *93*, 6654-6670.

[35] M.L. Cohen, S. Froyen and S.G. Louie (1982), *Nonlinear ionic pseudopotentials in spin-density-functional calculations*, Physical Review B *26*, 1738-1742.

[36] L. Colombo and S. Goedecker (1994), *Efficient linear scaling algorithm for Tight-Binding molecular dynamics*, Phys.Rev. Lett. *73*, 122-125.

[37] A.D. Daniels, J.M. Millam and G.E. Scuseria (1997), *Semiempirical methods with conjugate gradient density matrix search to replace diagonalization for molecular systems containing thousands of atoms*, J. Chem. Phys. *107*, 425-431.

[38] A. Daniels and G. Scuseria (1999), *What is the best alternative to diagonalization of the Hamiltonian in large semiempirical calculations ?*, J. Chem. Phys. *110*, 1321-1328.

[39] E.R. Davidson and D. Feller (1986), *Basis set selection for molecular calculations*, Chem. Rev. *86*, 681-696.

[40] M.S. Daw (1993), *Model for energetics of solids based on the density matrix*, Phys. Rev. B *47*, 10895-10898.

[41] M. Dolg (1996), *On the accuracy of valence correlation energies in pseudopotential calculations*, J. Chem. Phys. *104*, 4061-4067.

[42] D.A. Drabold and U. Stephan (1998), *Order N projection method for first principles calculation of electronic quantities and Wannier functions*, Phys. Rev. B *57*, 6391-6408.

[43] J. Flesh, W. Meyer and W. Müller (1984), *Treatment of intershell correlation effects in ab initio calculations by use of core polarization potentials. Method and application to alkali and alkaline earth atoms*, J. Chem. Phys. *80*, 3297-3310.

[44] A.J. Freeman, A.J. Krakauer, H. Weinert and E. Winner (1981), *Full-potential self-consistent linearized-augmented-plane-wave method for calculating the electronic structure of molecules and surfaces : O_2 molecule*, Phys. Rev. B *24* 864-866.

[45] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, V.G. Zakrzewski, J.A. Montgomery, R.E. Stratmann, J.C. Burant, S. Dapprich, J.M. Millam, A.D. Daniels, K.N. Kudin, M.C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G.A. Petersson, P.Y. Ayala, Q. Cui, K. Morokuma, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J. Cioslowski, J.V. Ortiz, B.B. Stefanov, G. liu, A. Liashenko, P. Piskorz, I. Kpmaromi, G. Gomperts, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P.M.W. Gill, B.G. Johnson, W. Chen, M.W. Wong, J.L. Andres, M. Head-Gordon, E.S. Replogle and J.A. Pople, Gaussian 98 (Revision A.7), Gaussian Inc., Pittsburgh PA 1998.

[46] M. Fuchs and M. Scheffler (1999), *Ab initio pseudopotentials for electronic structure calculations of poly-atomic systems using density-functional theory*, Computer Physics Communications *119*, 67-98.

[47] G. Galli and M. Parrinello (1992), *Large scale electronic structure calculations*, Physical Review Letters *69*, 3547-3550.

[48] G. Galli and F. Mauri (1994), *Electronic structure calculations and molecular dynamics simulations with linear system size scaling*, Physical Review B *50*, 4316-4326.

[49] G. Galli, J. Kim, F. Mauri (1995), *Total energy global optimizations using nonorthogonal localized orbitals*, Phys. Rev. B *52*, 1640-1648.

[50] G. Galli (2000), *Large scale electronic structure calculations using linear scaling methods*, Phys. Stat. Sol. B *217*, 231-249.

[51] A. Gibson, R. Haydock and J.P. Lafemina (1993), *Ab initio electronic-structure computations with recursion methods*, Phys. Rev. B *47*, 9229-9237.

[52] P.M.W. Gill (1994), *Molecular integrals over gaussian basis functions*, Adv. Quantum Chem. *25*, 141-205.

[53] M.J. Gillan (1989), *Calculation of the vacancy formation energy in Al*, J. Phys. Condens. Matt. *1*, 689-671.

[54] M.J. Gillan, C.M. Goringe and E. Hernández (1996), *Linear-scaling density-functional-theory technique : The density-matrix approach*, Physical Review B *53*, 7147-7157.

[55] I. Grinberg, N.J. Ramer and A.M. Rappe (2001), *Quantitative criteria for transferable pseudopotentials in density functional theory*, Phys. Rev. B *63*, 201102(R).

[56] S. Goedecker (1995), *Low complexity algorithms for electronic structure calculations*, J. Comput. Phys. *118*, 261-268.

[57] S. Goedecker (1998), *The decay properties of the finite temperature density matrix in metals*, Physical Review B *58*, 3501-3502.

[58] S. Goedecker (1999), *Linear scaling electronic structure methods*, Rev. Mod. Phys. *71*, 1085-1123.

[59] X. Gonze, P. Kackell and M.Scheffler (1990), *Ghost states for separable, norm-conserving, ab initio pseudopotentials*, Physical Review B *41*, 12264-12267.

[60] X. Gonze, R. Stumpf and M. Scheffler (1991), *Analysis of separable potentials*, Physical Review B *44*, 8503-8513.

[61] D.R. Hamann, M. Schluter and C. Chiang (1979), *Norm-conserving pseudopotentials*, Physical Review Letters *43*, 1494-1497.

[62] C. Hartwigsen, J. Hutter and S. Goedecker (1998), *Relativistic separable dual-space Gaussian pseudopotentials from H to Rn*, Physical Review B *58*, 3641-3662.

[63] P.J. Hay, W.R. Wadt (1985), *Ab initio effective core-potentials for molecular calculations. Potentials for K to Au including the outermost core orbitals*, J. Chem. Phys. *82*, 299-310.

[64] P.D. Haynes and M. C. Payne (1999), *Corrected penalty-functional method for linear-scaling calculations within density-functional theory*, Phys. Rev. B *59*, 12173-12176.

[65] V. Heine (1970), *The Pseudopotential Concept*, in H. Ehrenreich, F. Seitz and D. Turnbull, eds., Solid State Physics Vol. 24 (Academic Press, New York) 1-36.

[66] V. Heine (1980), *Solid State physics : Advances in Research and Aplications*, in : F. Seitz, C. Turnbull, H. Ehrenreich (Eds), vol 35, Academic Press, New-York.

[67] C. Herring (1940), *A new method for calculating wave functons in crystals*, Physical Review *57*, 1169-1177.

[68] N.A.W. Holzwarth, A.R. Tackett and G.E. Matthews (2001), *A projector augmented wave (PAW) code for electronic structure calculations, Part I :* atompaw *for generating atom-centered functions*, Comput. Phys. Comm. *135*, 329-347.

[69] S. Ismail-Beigi and T.A. Arias (1999), *Locality of the density matrix in metals, semiconductors and insulators*, Physical Review Letters *82*, 2127-2130.

[70] S. Itoh, P. Ordejón, R.M. Martin (1995), *Order-N tight-binding molecular dynamics on parallel computers*, Comp. Phys. Commun. *88*, 173-185.

[71] D. Joubert and G. Kresse (1999), *From ultrasoft pseudopotentials to the projector augemnted-wave method*, Phys. Rev. B *59*, 1758-1775.

[72] G.P. Kerker (1980), *Non-singular atomic pseudopotentials for solid state applications*, J. Phys. C. : Solid ST. Phys. *13*, 189-94.

[73] L. Kleinman and J.C. Phillips (1959), *New method for calculating wave functions in crystals and molecules*, Physical Review *116*, 287-294.

[74] L. Kleinman (1980), *Relativistic norm-conserving pseudopotential*, Physical Review B *21*, 2630-2631.

[75] L. Kleinman and D.M. Bylander (1982), *Efficacious form for model pseudopotentials*, Physical Review Letters *48*, 1425-1428.

[76] W. Kohn (1959), *Analytic properties of Bloch waves and Wannier functions*, Phys. Rev. *115*, 809-821.

[77] W. Kohn (1993), *Density functional/Wannier function theory for systems of very many atoms*, Chem. Phys. Lett. *208*, 167-172.

[78] W. Kohn (1996), *Density functional and density method scaling linearly with the number of atoms*, Phys. rev. lett. *76*, 3168-3171.

[79] J. Kress et al. (1998), *Parallel O(N) tight-binding molecular dynamics of polyethylene and compressed methane*, J. Comput.-Aided Mater. Des. *5*, 295-316.

[80] S. Lewis, E.J. Mele, A.M. Rappe and C.Y. Wei (1998), *Efficient scaling of calculations involving separable nonlocal potentials*, Phys. Rev. B *58*, 3482-3485.

[81] X.-P. Li, R.W. Numes and D. Vanderbilt (1993), *Density-matrix electronic structure method with linear system size scaling*, Phys. Rev. B *47*, 10891-10894.

[82] A.Y.Liu, M.R. Pederson and D. Porezag (1999), *Importance of nonlinear core corrections for density-functional based pseudopotential calculations*, Physical Review B *60*, 14132-14139.

[83] N. Marzari, D. Vanderbilt (1997), *Maximally localized generalized Wannier functions for composite energy bands*, Physical review B *56*, 12847-12865.

[84] D.M.C. Nicholson et al. (1994), *Stationary nature of the density-functional free energy : application to accelerated multiple-scattering calculations*, Phys. Rev. B *50*, 14686-14689.

[85] A.M. Niklasson (2002), *Expansion algorithm for the density matrix*, Phys. Rev. B *66*, 155115-155120.

[86] R. Nunes, D. Vanderbilt (1994), *Generalization of the density-matrix method to a nonorthogonal basis*, Phys. Rev. B *50*, 17611-17614.

[87] P. Ordejón, D.A. Drabold, M.D. Grumbach and R.M. Martin (1993), *Unconstrained minimization approach for electronic computations that scales linearly with system size*, Phys. Rev. B *48*, 14646-14649.

[88] P. Ordejón (1998), *Order N tight binding methods for electronic structure and molecular dynamics*, Computational Materials Science *12*, 157-191.

[89] T.Ozaki (2001), *Efficient recursion method for inverting overlap matrix*, Phys. Rev. B *64*, 195110-195116.

[90] A. Palser and D. Manopoulos (1998), *Canonical purification of the density matrix in electronic structure theory*, Phys. Rev. B *58*, 12704-12711.

[91] W.E. Pickett (1989), *Pseudopotential methods in condensed matter applications*, Computer Physics Report *9*, 115-198.

[92] P. Pulay (1987), *Analytical derivative methods in quantum chemistry*, Adv. Chem. Phys. *69*, 241-286.

[93] P. Pyykkö (1988), *Relativistic effects in structural chemistry*, Chem. Rev. *88*, 563-594.

[94] S.Y. Qiu, C.Z. Wang, K.M. Ho, C.T. Chan (1994), *Tight-binding molecular dynamics with linear system size scaling*, J. Phys. Condens. Matter *6*, 9153-9172.

[95] A.M. Rappe, K.M. Rabe, E Kaxiras and J.D. Joannopoulos (1990), *Optimized pseudopotentials*, Physical Review B *41*, 1227-1230.

[96] C.C.J. Roothaan (1951), *New developments in molecular orbital theory*, Rev. Mod. Phys. *23*, 69-89.

[97] D. Sánchez-Portal, P. Ordejón, E. Artacho and J.M. Soler, *Density-functional method for very large systems with LCAO basis sets*, Int. J. Quantum Chem. 65 (1997) 453-461.

[98] E. Schwegler and M. Challacombe (2000), *Linear scaling computation of the Fock matrix*, J. Chem. Phys. *106*, 5526-5536.

[99] M.P. Teter and L.W. Wang (1992), *Simple quantum-mechanical model of covalent bonding using a tight-binding basis*, Phys. Rev. B *46*, 12798-12801.

[100] N. Troullier and J.L. Martins (1991), *Efficient pseudopotentials for plane-wave calculations*, Physical Review B *43*, 1993-2006.

[101] D. Vanderbilt (1990), *Soft Self-consistent pseudo-potentials in a generalized eigenvalue formalism*, Physical Review B *41*, 7892-7895.

[102] U. Von Barth and C.D. Gelatt (1980), *Validity of the frozen-core approximation and pseudopotential theory for cohesive energy calculations*, Physical Review B *21*, 2222-2228.

[103] C.A. White, P. Maslen, M.S. Lee and M. Head-Gordon (1997), *The tensor properties of energy gradients within a non-orthogonal basis*, Chem. Phys. Lett. *276*, 133-138.

[104] W. Yang and T. Lee (1995), *A density-matrix divide-and-conquer approach for electronic structure calculations of large molecules*, J. Chem. Phys. *163*, 5674-5678.

Liens

[105] *http ://www.gaussian.com/.*

[106] *http ://www.uam.es/departamentos/ciencias/fismateriac/siesta/.*

[107] *http ://www.accelrys.com/cerius2/dmol3.html.*

[108] *http ://cms.mpi.univie.ac.at/vasp/.*

[109] *http ://www.abinit.org/.*

Références en Mathématiques appliquées

Ouvrages généraux

[110] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal et C. Sagastizábal (2002), *Numerical Optimization : Theoretical and Practical Aspects*, Springer.

[111] F. Chatelin (1993), *Eigenvalues of matrices*, Wiley.

[112] J.W. Demmel (1997), *Applied Numerical Algebra*, SIAM Press, Philadelphia, PA.

[113] R. Fletcher (1987), *Practical Methods of Optimization*, Wiley, Chichester, 2nd edition.

[114] G. Golub and C.F. Van Loan (1988), *Matrix Computations*, The John Hopkins Univ. Press.

[115] A. Quarteroni and A. Valli (1999), *Domain Decomposition Methods for Partial Differential Equations*, Oxford Science Publications, Clarendon Press, Oxford.

[116] Y. Saad (1992), *Numerical Methods for Large Eigenvalue Problem : Theory and Algorithms*, Manchester University Press.

[117] Y. Saad (1996), *Iterative Methods for Sparse Linear Systems*, PWS Publishing Co., Boston.

[118] R. Verfurth (1996), *A Review of a Posteriori Error Estimation and Adaptative Refinement Techniques*, Chichester, England.

[119] J.H. Wilkinson (1965), *The Algebraic Eigenvalue Problem*, Oxford University Press.

Thèses

[120] I. Dhillon (May 1997), *A new algorithm for the symmetric tridiagonal eigenvalue/eigenvector problem*, PhD Thesis, University of California, Berkeley, California.

[121] R.B. Lehoucq (May 1995), *Analysis and implementation of an implicitly restarted Arnoldi iteration*, PhD Thesis, Rice University - Houston Texas.

[122] N.C. Nguyen (2005), *PhD Thesis*, Singapore-MIT Alliance, National University of Singapore. In progress.

Articles

[123] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney and D. Sorensen, *LAPACK users' guide, 3rd edition*, SIAM 1999.

[124] P. Arbenz, U.L. Hetmaniuk, R.B. Lehoucq and R.S. Tuminaro, *A comparison of eigensolvers for large-scale 3D modal analysis using AMG-preconditioned iterative methods*, Int. J. Numer. Meth. Engng 64 (2005) 204-236.

[125] W.E. Arnoldi (1951), *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math. *9*, 17-19.

[126] E. Balmes (1996), *Parametric families of reduced finite element models. theory and applications*, Mechanical Systems and Signal Processing *4*, 381-394.

[127] E. Cancès and C. Le Bris (2000), *Can we outperform the DIIS approach for electronic structure calculations*, Int. J. Quantum Chem. *79* 82-90.

[128] V. Duwig (2002), note HI-23/2002/024 interne EDF, *Vers des méthodes d'ordre $\mathcal{O}(N)$, présentation de la technique de la déflation.*

[129] R. Greengard and V. Rokhlin (1997), *A new version of the fast multipole for the Laplace equation in three dimensions*, Acta Numerica *6*, 229-269.

[130] U.L. Hetmaniuk and R.B. Lehoucq, *Multilevel methods for eigenspace computations in structural dynamics*, Proceedings of the 16th International Conference on Domain Decomposition Methods, Courant Institute, New-York, January 12-15, 2005.

[131] K. Ito and S.S. Ravindran (1998), *A reduced basis method for control problems governed by PDEs*, International Series of Numerical Mathematics *126* 153-168, Birkhauser Verlag, Base.

[132] C. Le Bris (2005), *Computational chemistry from the perspective of numerical analysis*, Acta Numerica, *14*, 363-444.

[133] P.L. Lions (1987), *Solutions of Hartree-Fock equations for Coulomb systems*, Commun. Math. Phys. *53*, 22-116.

[134] L. Machiels, Y. Maday, A.T. Patera and D.V. Rovas (2000), *Blackbox reduced-basis output bound methods for shape optimization* In proceedings 12^{th} International Domain Decomposition Conference, 429-436, Chiba, Japan.

[135] L. Machiels, A.T. Patera and D.V. Rovas (2001), *Reduced basis output bound methods for parabolic problems*, Computer Methods in Applied Mechanics and Engineering, submitted.

[136] L. Machiels et al. (2002), *Reliable real-time solution of parametrized partial differential equations : reduced-basis outpout bound methods*, Journal of Fluids Engineering *124*, 70-80.

[137] Y. Maday, A.T. Patera and J. Peraire (1999), *A general formulation for a posteriori bounds for output functionals of partial differential equations ; application to the eigenvalue problem*, C. R. Math. Acad. Sci. Paris, Série I. *328*, 9, 823-828.

[138] Y. Maday and A.T. Patera (2000), *Numerical analysis of a posteriori finite element bounds for linear functional outputs*, Math. Models Methods Appl. Sci. *10*, 5, 785-799.

[139] Y. Maday, A.T. Patera and G. Turinici (2002), *A priori convergence theory for approximation of single-parameter symmetric coercive elliptic partial differential equations*, C.R. Acad. Sci. Paris Série I *335*, 289-294.

[140] A.K. Noor and J.M. Peters (1980), *Reduced basis technique for nonlinear analysis of structures*, AIAA Journal *18*, 455-462.

[141] A.T. Patera, C. Prud'homme, K. Veroy and D.V. Rovas (2003), *A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations*, Proceedings 16th AIAA Computational Fluid Dynamics Conference.

[142] A.T. Patera, C. Prud'homme and K. Veroy (2003), *Reduced basis approximation of the viscous burgers equation : rigorous a posteriori error bounds*, C.R. Acad. Sci. Paris 337, *9*, 619-624.

[143] C. Paige and M. Saunders (1975), Solution of sparse indefinite systems of linear equations, SIAM J. Nimer. Anal., *12*, 617-629.

[144] J.S. Peterson (1989), *The reduced basis method for incompressible viscous flow calculations*, SIAM J. Sci. Stat. Comput. *10*, 777-786.

[145] T.A. Porsching (1985), *Estimation of the error in the reduced basis method solution of nonlinear equations*, Mathematics of Computation *45*, 487-496.

[146] D.C. Sorensen (1992), *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Ana. Appl. *13*, 357-385.