École des Ponts ParisTech

Inria

THÈSE DE DOCTORAT

de l'École des Ponts ParisTech

Reduction methods for physics and quantum chemistry models

École doctorale MSTIC

Discipline: Mathématiques Appliquées

Thèse préparée au CERMICS, au sein de l'équipe Inria MATHERIALS

Thèse soutenue le 20 juin 2024, par

Laurent Vidal

Composition du jury

Xavier Blanc Professeur, Université Paris Cité	Rapporteur
Phanish Suryanarayana Professeur, Georgia Institute of Technology	Rapporteur
Geneviève Dusson Chargée de recherche CNRS, Université de Franche-Comté	Examinatrice
Eleonora Luppi Maîtresse de Conférences, Sorbonne Université	Examinatrice
Andreas Savin Directeur de recherche CNRS émérite, Sorbonne Université	Examinateur
Benjamin Stamm Professeur, Université de Stuttgart	Examinateur
Eric Cancès Professeur, École des Ponts & Inria	Directeur de thèse
Antoine Levitt Professeur junior, Université Paris-Saclay	Directeur de thèse

EMC² Extreme-scale Mathematically-based Computational Chemistry

erc European Research Council
Established by the European Commission

Résumé

Cette thèse de doctorat porte sur l'analyse numérique et le test de nouvelles méthodes pour le calcul de l'état fondamental en théorie de la structure électronique. Nous nous concentrons sur des méthodes variationnelles basées sur la théorie des fonction d'ondes (WFT) et la théorie de la fonctionnelle de la densité (DFT), pour lesquelles le problème de l'état fondamental consiste en la minimisation d'une fonctionnelle d'énergie sur des familles orthonormées de fonctions carrées intégrables, dites *orbitales moléculaires.*

La première partie de notre travail est consacrée aux systèmes moléculaires. Dans le chapitre 1, nous étudions des algorithmes de minimisation directe pour les modèles de Hartree-Fock restreint à couche ouverte (ROHF) et de champ auto-cohérent de l'espace actif complet (CASSCF), deux modèles standard de fonction d'onde pour les systèmes à espace actif non vide (l'espace fonctionel lié à la présence d'électrons non appairés dans le système). En utilisant les outils de l'optimisation Riemannienne sur des variétés quotient matricielles, nous exprimons dans un cadre commun les algorithmes de minimisation directe pour ROHF et CASSCF existants dans la littérature chimique, et proposons de nouveaux algorithmes dont nous testons les performances. Dans le deuxième chapitre, ce même formalisme nous permet d'expliquer l'instabilité inhérente aux algorithmes de champ auto-cohérent pour ROHF, qui sont les plus couramment utilisés pour ce modèle. Nous proposons ensuite un nouveau type de méthode de point fixe qui est plus robuste et sans paramètre, et qui rivalise avec les méthodes classiques les plus performantes. Le troisième chapitre aborde l'optimisation des bases d'orbitales atomiques pour les systèmes moléculaires. Nous introduisons un cadre mathématique général pour ce problème et comparons la précision des bases optimisées pour certains critères d'optimisation utilisés dans la littérature sur un modèle jouet unidimensionnel.

La deuxième partie de cette thèse se concentre sur les matériaux cristallins. Le chapitre 4 est consacré à la correction de certaines erreurs de discrétisation liées à l'utilisation de bases d'ondes planes tronquées dans les calculs de DFT. Dans ce chapitre, nous introduisons une méthode de Galerkin basée sur l'introduction d'un opérateur cinétique modifié. Nous dérivons une estimation de l'erreur d'approximation pour les énergies obtenues et prouvons des estimées optimales sur la régulatité des diagrammes de bandes associés, en fonction d'un paramètre défini par l'utilisateur. Enfin, le chapitre 5 détaille deux contributions numériques liées à la structure électronique du graphène bicouche twisté (TBG). Tout d'abord, nous posons les bases d'un code de simulation du TBG en langage Julia, conçu pour être facilement utilisable et modifiable. En second lieu, nous appliquons une méthode de compression, présente dans la littérature, pour développer les fonctions de Wannier correspondant aux deux bandes de valence les plus basses du graphène, dans une base d'orbitales atomiques adaptée à la symétrie. Ces fonctions compressées devraient permettre d'effectuer de grands calculs de type liaisons fortes pour le TBG, l'un des outils principaux pour décrire la physique nouvelle observée sur ce matériau moiré.

Abstract

This PhD thesis is concerned with the numerical analysis and testing of novel methods for ground state calculation in electronic structure theory. We focus on variational methods based on Wave-Function Theory (WFT) and Density Functional Theory (DFT), for which the ground state problem consists in the minimization of an energy functional over orthonormal families of square integrable functions, the so-called *molecular orbitals*.

The first part of our work is dedicated to molecular systems. In chapter 1, we investigate direct minimization algorithms for the restricted open-shell Hartree-Fock (ROHF) and complete active space self-consistent field (CASSCF) models, two standard WFT models for systems with non empty active space (which designates the space of molecular orbitals related to the presence of unpaired electrons in the system). By using the tools of Riemannian optimization on matrix quotient manifolds, we express in a common framework the direct minimization algorithms for ROHF and CASSCF already existing in the chemistry literature, and propose and test the performance of new algorithms. In the second chapter, the same formalism allows us to explain the inherent instability of self-consistent field algorithms for ROHF, which are the most commonly used for this model. We then propose a new type of fixed-point method that is more robust and parameter free, while competing with the best performing standard methods. The third chapter discusses the optimization of atomic orbital basis sets for molecular systems. We introduce a general mathematical framework for that problem and compare the accuracy of optimized basis sets for some optimization criteria used in the literature on a one-dimensional toy model.

The second part of this PhD focuses on crystalline materials. Chapter 4 is devoted to the correction of discretization errors related to the use of truncated Fourier basis in DFT calculations. Within this chapter, we introduce a plane-wave Galerkin discretization method based on a modified kinetic operator. We derive an error estimate for approximate energies and prove optimal regularity results on band diagrams, which depend on a user-defined tunable parameter. Lastly, chapter 5 details two numerical contributions related to the electronic structure of twisted-bilayer graphene (TBG). First we establish the groundwork for a code in Julia language, designed as a user-friendly plateform for the simulation of TBG. Second, we use an existing compression method to expand the Wannier functions corresponding to the two lowest valence bands of graphene on a symmetry-adapted atomic-orbital basis. These compressed functions should allow to perform large tight binding calculation for TBG, a useful tool to describe the novel physics observed on this moiré materials.

Remerciements

Ces quatre dernières années m'ont appris que la pratique des mathématiques est avant tout collective. Les pages de ce manuscrit portent en filigrane les visages et les noms de celles et ceux qui m'ont permis de cheminer et d'arriver à bon port sans trop d'encombres. Tout du moins, d'arriver quelque part. Je souhaiterais maintenant temps de les remercier.

Mon attention se porte d'abord sur mes directeurs de thèse Éric Cancès et Antoine Levitt. Avant toute chose, je leur suis reconnaissant de m'avoir laissé du temps. Le temps de faire mes preuves, en prolongeant de quatre mois mon stage de master qui m'avait paru trop court pour quitter les bancs de la fac et entrer dans le monde de la recherche. Le temps d'apprendre, lorsque lisant dans mes yeux une n-ième fois que je n'avais rien compris, ils ont accepté de me réexpliquer. En bref, le temps de me transformer, et cela m'a été très précieux.

Je les remercie également de m'avoir inclus dès le premier jour dans les groupes de travail, les conférences, les écoles d'été et les projets qu'ils jugeaient utiles à mon apprentissage. Les rencontres que j'y ai faites ont été déterminantes dans la bonne conduite de ma thèse. Je les remercie d'avoir su me motiver et orienter mon travail de thèse en fonction de mes aspirations et de mes capacités. Enfin je les remercie de m'avoir appris la dure précision nécessaire à la pratique scientifique, tout en restant accessibles et agréables à côtoyer, effaçant partout où ce n'était pas nécessaire les arguments d'autorité.

Je souhaite ensuite remercier grandement Xavier Blanc et Phanish Suryanarayana d'avoir accepté de rapporter ma thèse. Je suis également reconnaissant à Geneviève Dusson, Eleonora Luppi, Andreas Savin et Benjamin Stamm d'avoir accepté de faire partie du jury. Merci à eux de rendre possible cet évènement important de ma vie scientifique et personnelle. En particulier je remercie pour leurs précieux conseils et recommandations Geneviève Dusson, avec qui j'ai eu le plaisir de collaborer, et Andreas Savin dont j'ai croisé la route régulièrement lors de ma thèse.

Plus largement, je souhaite remercier tous les chercheurs ayant contribué directement à ce travail de thèse. Je pense à Michael Herbst, pour m'avoir encouragé pendant mon prédoctorat à prendre confiance dans mes capacités, pour son franc-parler et ses conseils. Et bien sûr, pour avoir créé avec Antoine Levitt le code DFTK qui a servi de support à toute la deuxième partie de ma thèse. Je pense aussi à Muhammad Hassan avec qui j'ai eu grand plaisir à travailler sur la régularisation de diagrammes de bandes, mais aussi à philosopher et plus largement partager des moments conviviaux. Son esprit alerte, son aptitude à toujours se questionner et sa gentillesse m'ont marqué durablement. Je pense à Gaspard Kemlin avec qui j'ai collaboré sur l'optimisation de bases d'orbitales atomiques. Un "grand frère" de thèse, devenu maître de conférence, qui fut souvent d'un grand secours. Un mot d'ordre : toujours précis mais jamais sérieux ! Je pense à Robert Benda, un autre "grand frère", dont l'enthousiasme indéfectible fut un moteur important de mes premiers pas en thèse. Notre collaboration avec Hubert fut très inspirante. Merci à Susi Lehtola pour son aide précieuse dans l'exploration de la littérature de chimie, pour avoir adapté à la volée son code HelFEM par deux fois et pour la grande patience dont il a fait preuve pour m'aider à appliquer notre étude des bases gaussiennes à des systèmes réels. Son regard sur l'état actuel de la chimie quantique a été très éclairant. Merci également à Emmanuel Giner d'avoir accompagné mes premiers pas dans la rédaction d'un code scientifique à travers notre collaboration sur le modèle ROHF, en utilisant le code quantum package 2. Je le remercie pour nos discussions très instructives et rassurantes sur la multiplicité des approches pour faire de la recherche. Enfin je pense à Filippo Lipparini et Tommaso Nottoli, sans qui je n'aurais jamais pu publier en si peu de temps nos recherches sur l'optimisation Riemannienne en chimie quantique, qui me tenait particulièrement à cœur. Leur assurance et leur capacité à faire vite et bien fut un grand renfort à la fin de ma thèse.

J'aimerais également remercier les chercheurs de passage au CERMICS ou rencontrés dans le cadre très riche des collaborations liées à l'ERC EMC2 et au GDR NBODY. Je pense en particulier à Mi-Song Dupuy, Emmanuel Fromager, Louis Garrigue et Julien Toulouse avec qui j'ai particulièrement apprécié discuter.

Je pense maintenant à mes compagnons d'équipage que je n'ai pas cités :

- ceux de **l'équipe quantique**, Alicia, Alfred, Andrea, Eloïse, Etienne, Long, Solal ;
- du **CERMICS**, Alberic, Amandine, Antonin, Camila, Clément, Coco, Edoardo, Emanuele, Fabian, Giulia, Guillaume, Hadrien, Hervé, Jean, Julien, Kacem, Louis, Louis-Pierre, Luca, Léo, Mathias, Nerea, Noé, Raphaël, Regis, Renato, Roberta, Rutger, Seta, Shiva, Simon, Zoé ;
- ceux de **Jussieu**, Agustin, Anatole, Diata, Thomas ;
- du **CEMRACS**, Adrien, Beatrice, Ludovica ;
- de **l'Inria** comme Édouard, et tous ceux rencontrés à **Roscoff** et à **Aussois** ;
- enfin tous ceux rencontrés et enfin tout ceux que ma pauvre mémoire des noms et me fait oublier au moment d'écrire ces remerciements.

Avec eux j'ai partagé repas, banquets, voyages en train, en bus et à vélo, expérimentations mathématiques, musicales, littéraires, gâteaux (en grande quantité), pétanques, tennis de table, chasse au trésor, bains de plage, de lac et de rivière, randonnées en villes, sur une île ou en montagne, verres de vin et autres spiritueux (en grande quantité), jeux de société, inquiétudes (parfois), joies (souvent), pauses (en quantité modérée bien sûr), jeux de mots et d'esprits, improvisations, mots croisés, escalade.

Tous sont de précieux collègues. Ceux que j'ai eu le temps de connaître sincèrement sont mes amis.

Enfin je souhaiterais remercier les gens qui m'ont aidé en dehors du cercle professionnel. D'abord, je remercie affectueusement mes parents pour m'avoir fait confiance et pour m'avoir laissé une grande liberté d'expérimenter, en m'aventurant parfois dans des voies bien étranges. Je suis aussi reconnaissant envers mon frère et ma sœur pour leur soutien inaltérable. Plus largement merci à mes amis de toujours. Je conclurai en remerciant toutes les personnes qui m'ont permis de développer, en parallèle de ma thèse, une activité musicale et théâtrale riche en émotions.

Et merci à tous les lecteurs éventuels de ce manuscrit !

List of contributions

Published papers (chapter 3 and chapter 4)

[LV1] Eric Cancès, Geneviève Dusson, Gaspard Kemlin, and Laurent Vidal. "On basis set optimisation in quantum chemistry". In: *ESAIM: Proceedings and Surveys* 73 (2023), pp. 107–129.

[LV2] Eric Cancès, Muhammad Hassan, and Laurent Vidal. "Modified-operator method for the calculation of band diagrams of crystalline materials". In: *Mathematics of Computation* (2023).

Preprints (chapter 1 and chapter 2)

[LVp1] Laurent Vidal, Tommaso Nottoli, Filippo Lipparini, and Eric Cancès. "Geometric optimization of Restricted-Open and Complete Active Space Self-Consistent Field wavefunctions". *Submitted*.

[LVp2] Robert Benda, Eric Cancès, Emmanuel Giner, and Laurent Vidal. "Self-consistent field algorithms in Restricted Open-Shell Hatree-Fock". *Submitted*.

Software

Our contributions are the following:

- For the works described in chapter 1 and chapter 2, we implemented some Riemannian direct minimization methods and standard self-consistent field algorithms for restricted open-shel Hartree-Fock in the `Julia` package `ROHFToolkit`[1];
- the modified operator approach of chapter 4 has been implemented in the `DFTK` [HLC21] software[2];
- as described in the first part of chapter 5, we developped a code for the simulation of twited-bilayer graphene as the `Julia` package `TwistedBilayerGraphene`;
- the work presented in the second part of chapter 5 resulted in the development of an interface for `DFTK` with the `Wannier90` [Piz+20] program[3].

[1] https://github.com/LaurentVidal95/ROHFToolkit
[2] https://docs.dftk.org/stable/examples/energy_cutoff_smearing/
[3] https://docs.dftk.org/stable/examples/wannier/

Context and motivations

The contributions made during this PhD thesis are related to the numerical analysis and implementation of *ab initio* electronic structure methods. Electronic structure theory aims at describing the behavior of electrons in materials, ranging from isolated atoms to macroscopic solids, in order to derive some of their qualitative and quantitative properties. The *ab initio* methods, in particular, are derived from the first principles of quantum mechanics and depend solely on fundamental physical constants or well-known parameters such as nuclear masses.

The advent of quantum mechanics in the early 20th century, notably effective in predicting the behavior of individual particles, offered a glimpse of the possibility to understand matter at all scales from first principles. Yet, while the task is already arduous for a single particle, it becomes virtually impossible as the number of particles increases. Each atomic nucleus or electron in matter interacts with all the other particles, yielding a partial differential equation whose dimension grows linearly with the size of the system. Producing precise approximations for this problem, that can be used in real-world applications, has been a continuous effort of theoretical chemists and solid state physicists since the 1920's.

The contribution of the mathematical community, in particular in the analysis of numerical methods, is more recent. Mathematicians have contributed significantly to the theoretical understanding of electronic structure models, and the practical implementation of numerical methods.

Electronic structure computations represent a substantial portion of the computation time worldwide. Still, most numerical methods require the user to manually set input parameters (solvers, convergence criteria, ...), which strongly influence the accuracy and computational efficiency of these methods. The design of fast, stable and black-box algorithms should be highly beneficial to theoretical chemists and physicists, but also to the industry, where users may lack the expertise or time to question the validity of the calculations. This PhD thesis aims at being a small contribution to this ongoing effort.

Contents

Introduction

This opening chapter introduces the mathematical formalism for electronic structure models and related problems studied throughout this manuscript and delineates the specific contributions of this PhD.

Contents

1 The quantum many-body electronic ground state problem

In all what follows, matter is described at the atomic level using non-relativistic quantum mechanics. In order to reduce the dimensionality of the problem, all models are introduced in the Born-Oppenheimer approximation [CLBM06, Appendix A], in which atomic nuclei are considered as point-like classical particles. In turn, only the electrons are described at the quantum level. We also restrict to electronic states of lowest energy, whose physical properties are independent of time. We adopt the system of atomic units (a.u.) in which

$$m_e = 1, \qquad e = 1, \qquad \hbar = 1, \qquad \frac{1}{4\pi\varepsilon_0} = 1, \tag{1.1}$$

where m_e is the mass of the electron, e the elementary charge, $\hbar$ the reduced Planck constant and ε_0 the dielectric permittivity of vacuum. In particular, the a.u. distance unit is the Bohr (a_0), equal to the average distance between the nucleus and the electron in the ground-state of the hydrogen atom, approximately 5.29×10^{-11} m. The atomic unit of energy is the Hartree (Ha), roughly equal to 4.36×10^{-18} J.

1.1 Physical states in quantum mechanics

In classical Hamiltonian mechanics, the physical states of a given system are described by a tuple (q, p) of coordinates and associated momenta. This information is sufficient in the sense that two states $\mathcal{X}_1$ and $\mathcal{X}_2$ can be distinguished by the only data of their associated tuples (q_1, p_1) and (q_2, p_2). The space of all such tuples is called the *phase space*, and it is of finite dimension. In addition, physical quantities are functions defined on the phase space.

A rather different formalism is used in quantum mechanics [BDJ02]. A quantum state is described by a wave-function ψ. It belongs to a complex Hilbert space $\mathcal{H}$, called the *state space*, which can be of infinite dimension. Physical quantities are measured by means of observables. An observable A, associated to the physical quantity a, is a self-adjoint operator on $\mathcal{H}$, often unbounded with domain $D(A)$ dense in $\mathcal{H}$, and whose spectrum is real and consists of the set of all admissible values for a. In contrast to classical mechanics, the same experiment can provide different measurements of a. The probability that a belongs to the open set $E \subset \mathbb{R}$ for the given state ψ is given by

$$\|P_A(E)(\psi)\|^2 \tag{1.2}$$

where P_A is the spectral projection-valued measure of A defined by the spectral theorem. Statistically, the mean value of the quantity a, measured on the specific wave-function $\psi \in \mathcal{H}$, is given by the Rayleigh quotient

$$q_A(\psi) := \frac{\langle \psi \,|\, A\psi \rangle_{\mathcal{H}}}{\|\psi\|^2}. \tag{1.3}$$

The self-adjointness of A implies in particular that for all $\psi \in Q(A)$, the form domain of A,

$$\langle \psi | A\psi \rangle_{\mathcal{H}} = \langle A\psi | \psi \rangle_{\mathcal{H}} = \langle \psi | A | \psi \rangle_{\mathcal{H}} \tag{1.4}$$

in the Dirac "bra-ket" notation. We will omit the index $\mathcal{H}$ when a single scalar product is considered. The main focus of this PhD will be the computations of the mean value q_A for the energy of a quantum system.

Let us mention two additional features of quantum mechanics that will be used in the sections below:

1. let $\psi \in \mathcal{H} \backslash \{0\}$. For all observable A and $\alpha \in \mathbb{R}$, the sesquilinearity of the scalar product of $\mathcal{H}$ implies $q_A(\alpha\psi) = q_A(\psi)$. Indeed, the whole line $\{\alpha\psi \mid \alpha \in \mathbb{C}^*\} \subset \mathcal{H}$ represents the same physical state. In other words, a quantum state is a point of the projective space $\mathbf{P}\mathcal{H}$. Still, it is common practice to represent a state $\mathcal{X}$ by a normalized wave-function ψ, i.e. such that $\|\psi\| = 1$, rather than an element of $\mathbf{P}\mathcal{H}$;

2. right after measurement, a given state ψ belongs to the eigenspace of A corresponding to the measured value of a. As a result, two observables that do not commute are incompatible, in the sense that they affect each-other's measurements. A *Complete Set of Commuting Observables* (CSCO) is a set of observable that commute and whose combined measurement allow to fully differentiate two states $\mathcal{X}_1$ and $\mathcal{X}_2$.

1.2 Quantum description of electrons

We now briefly state how this formalism applies to the description of electrons moving in $\mathbb{R}^3$. From a mathematical point of view, we introduce the N-electron state space and the observables related to energy and spin.

1.2.1 One-electron state space

In addition to its Cartesian coordinates $r = (x, y, z) \in \mathbb{R}^3$, an electron possesses an intrinsic spin $\sigma \in \mathbb{C}^2$. It is a purely quantum degree of freedom which is involved in a wide range of phenomena, from ferromagnetism and superconductivity to the arrangement of the periodic table. Actually, the spin parameter characterizes the irreducible projective representations of the rotation group $SO(3)$. In this manuscript, we will simply introduce spin as an ad-hoc degree of freedom, as it has been in the early days of quantum mechanics: let us denote by

$$\uparrow := \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \downarrow := \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

the canonical basis of $\mathbb{C}^2$. From the Stern-Gerlach experiment (e.g. [LL19, Chapter 1]), it is natural to postulate that one-electron states have a "spin-up" and a "spin-down" component. With this assumption, the configuration space of a single electron is the product $\mathbb{R}^3 \times \{\uparrow, \downarrow\}$ and the one-electron state space is the Hilbert space

$$\mathcal{H}_1 := L^2(\mathbb{R}^3 \times \{\uparrow, \downarrow\}; \mathbb{C}). \tag{1.5}$$

It is endowed with the scalar product

$$\langle \psi_1 | \psi_2 \rangle_{\mathcal{H}_1} = \int_{\mathbb{R}^3 \times \{\uparrow,\downarrow\}} \overline{\psi_1(r,\sigma)} \psi_2(r,\sigma) \mathrm{d}\mu(r,\sigma) \quad \text{with} \quad \mu = \lambda_{\mathbb{R}^3} \otimes (\delta_\uparrow + \delta_\downarrow), \tag{1.6}$$

where $\lambda_{\mathbb{R}^3}$ is the Lebesgue measure on $\mathbb{R}^3$. It is sometime preferable to work with the isomorphic expression

$$\boxed{\mathcal{H}_1 = L^2(\mathbb{R}^3 \times \{\uparrow, \downarrow\}; \mathbb{C}) \simeq L^2(\mathbb{R}^3; \mathbb{C}^2)} \tag{1.7}$$

where the spin degree of freedom appears in the co-domain of the wave-functions. In that case, wave-functions are vector-valued and are called *spinors*. We will use the two representations in the following exposition.

1.2.2 N-electron state space

Now consider a system X_N of N electrons. By generalizing the above description, the state space of X_N should simply read

$$L^2\left(\left(\mathbb{R}^3 \times \{\uparrow, \downarrow\} \right)^N ; \mathbb{C} \right). \tag{1.8}$$

However, in addition to the above-mentioned constraints, multi-electron states ψ have to verify the anti-symmetry property

$$\forall p \in \mathfrak{S}_N \quad \forall (\mathbf{x}_1, \dots, \mathbf{x}_N) \in \left(\mathbb{R}^3 \times \{\uparrow, \downarrow\} \right)^N, \qquad \psi(\mathbf{x}_{p(1)}, \dots, \mathbf{x}_{p(N)}) = \varepsilon(p) \psi(\mathbf{x}_1, \dots, \mathbf{x}_N) \tag{1.9}$$

with $\mathfrak{S}_N$ the group of all permutations of $\{1, \dots, N\}$ and $\varepsilon(p)$ the signature of the permutation p. This property, known as the *Pauli principle*, expresses the fact that electrons are fermions. In particular it implies that two electrons cannot be found in the same configuration $\mathbf{x} = (r, \sigma) \in \mathbb{R}^3 \times \{\uparrow, \downarrow\}$, since for all ψ verifying (1.9):

$$\exists\, i, j \in \{1, \dots, N\} \quad \text{s.t.} \quad \mathbf{x}_i = \mathbf{x}_j \quad \Longrightarrow \quad |\psi(\mathbf{x}_1, \dots, \mathbf{x}_N)|^2 = 0, \tag{1.10}$$

where we recall that $|\psi(\mathbf{x}_1, \dots, \mathbf{x}_N)|^2$ is interpreted as the probability to find the N electrons in respective configurations $\mathbf{x}_i$. The corresponding N-electron state space reads

$$\mathcal{H}_N := \left\{ L^2\left(\left(\mathbb{R}^3 \times \{\uparrow, \downarrow\} \right)^N ; \mathbb{C} \right) \,\middle|\, \psi \text{ verifies (1.9)} \right\}. \tag{1.11}$$

In practice, the properties of $\mathcal{H}_N$ are deduced from $\mathcal{H}_1$. For all $(\psi_1, \dots, \psi_N) \in \mathcal{H}_1^N$, and configuration $(\mathbf{x}_1, \dots, \mathbf{x}_N) \in (\mathbb{R}^3 \times \{\uparrow, \downarrow\})^N$, we denote

$$\psi_{i_1} \wedge \cdots \wedge \psi_{i_N}(\mathbf{x}_1, \dots, \mathbf{x}_N) = \frac{1}{\sqrt{N!}} \sum_{p \in \mathfrak{S}_N} \varepsilon(p) \psi_{p(i_1)}(\mathbf{x}_1) \times \cdots \times \psi_{p(i_N)}(\mathbf{x}_N) \tag{1.12}$$

the normalized antisymmetric product of $(\psi_1, \dots, \psi_N)$, also known as *Slater determinant*. Then $\mathcal{H}_N$ is isomorphic to the antisymmetric tensor product of N copies of $\mathcal{H}_1$:

$$\boxed{\mathcal{H}_N \simeq \bigwedge_1^N \mathcal{H}_1 = \mathrm{Span}\left\{\psi_1 \wedge \psi_2 \wedge \cdots \wedge \psi_N, \quad (\psi_1, \dots, \psi_N) \in \mathcal{H}_1^N\right\}.} \tag{1.13}$$

1.2.3 Energy and spin observables

In the scope of this manuscript, the only observables of interest are related to the energy and the spin of electrons.

The observable measuring the total energy E of X_N is the Hamiltonian. It is a sum of terms, each of which models a specific contribution to the total energy. In the absence of magnetic field, the non-relativistic Born-Oppenheimer N-electron Hamiltonian has the general form

$$\boxed{\hat{\mathrm{H}}_N = \hat{T} + \hat{V} + \hat{W}_{\mathrm{ee}} := -\frac{1}{2} \sum_{i=1}^N \Delta_{r_i} + \sum_{i=1}^N V(r_i) + \sum_{1 \leqslant i < j \leqslant N} \frac{1}{|r_i - r_j|}.} \tag{1.14}$$

It acts on $\mathcal{H}_N$ with domain $\mathcal{H}_N \bigcap H^2\left((\mathbb{R}^3 \times \{\uparrow, \downarrow\})^N; \mathbb{C}\right)$. The three terms are the kinetic energy of the electrons, the interaction of the electrons with an external potential V, and the electron-electron repulsion. The terms $\hat{V}$ and $\hat{W}_{\mathrm{ee}}$ are multiplication operators. The self-adjointness of $\hat{\mathrm{H}}_N$ and its spectrum, hence the accessible energies for X_N, depends on N and V, as discussed in the next section.

When it comes to spin, observables can be computed in various ways. Among the recent references, let us cite [LL19, Chapter 1] for a phenomenological introduction from the Stern-Gerlach experiment, and [Lew22, Section 4.1.7] for an algebraic viewpoint. In short, a one-electron spin state $\mathcal{X}$, represented by a normalized spinor $\psi \in L^2(\mathbb{R}^3; \mathbb{C}^2)$, can be mapped to a vector $\vec{\mathcal{X}}$ of $\mathbb{R}^3$ whose components are given by

$$[\vec{\mathcal{X}}]_\mu = 2\langle \psi \,|\, \hat{S}_\mu \psi \rangle_{L^2(\mathbb{R}^3; \mathbb{C}^2)}, \quad \mu \in \{x, y, z\}. \tag{1.15}$$

In the above expression, the x, y, and z-projected spin operators are defined by

$$\hat{S}_\mu = \frac{1}{2}\sigma_\mu \quad \text{with} \quad (\sigma_x, \sigma_y, \sigma_z) = \left(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\right) \tag{1.16}$$

acting on $L^2(\mathbb{R}^3; \mathbb{C}^2)$. The matrices σ_μ are called the Pauli matrices. The vector $\vec{\mathcal{X}}$ is unique for each spin state. Unfortunately, the operators $\hat{S}_x$, $\hat{S}_y$ and $\hat{S}_z$ do not commute. It is therefore impossible to have a precise measurement of the three components of $\vec{\mathcal{X}}$ at the same time. For that reason we introduce the spin and spin squared observables

$$\hat{S} = [\hat{S}_x, \hat{S}_y, \hat{S}_z], \qquad \hat{S}^2 = \hat{S}_x^2 + \hat{S}_y^2 + \hat{S}_z^2 = \frac{3}{4}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{1.17}$$

The operator S^2 commutes with all components of spin, and the data of a tuple $(s, m_z) \in \sigma\left(S^2\right) \times \sigma(\hat{S}_z)$ is, in practice, sufficient to discriminate spin states. From the $\hat{S}_\mu$ operators, we construct the j-th electron spin observables for a N-electron system as

$$\hat{S}_{\mu,j} = 1 \otimes \cdots \otimes \underbrace{\hat{S}_\mu}_{j-\mathrm{th}} \otimes \cdots \otimes 1, \qquad \hat{S}_j = [\hat{S}_{x,j}, \hat{S}_{y,j}, \hat{S}_{z,j}] \qquad \hat{S}_j^2 = \hat{S}_{x,j}^2 + \hat{S}_{y,j}^2 + \hat{S}_{z,j}^2 \tag{1.18}$$

which are bounded operators on $\mathcal{H}_N = \bigwedge_1^N L^2(\mathbb{R}^3; \mathbb{C}^2)$. We also introduce the total spin operators

$$\boxed{\hat{S}_{\mu,N} = \sum_{j=1}^{N} \hat{S}_{\mu,j}, \qquad \hat{S}_N = [\hat{S}_{x,N}, \hat{S}_{y,N}, \hat{S}_{z,N}] \quad \text{and} \quad \hat{S}_N^2 = \sum_{\mu\in\{x,y,z\}} \hat{S}_{\mu,N}^2.} \tag{1.19}$$

The spectra of the total spin operators can be computed from (1.16) and (1.19). They only depend on the number N of electrons

$$\sigma(\hat{S}_{\mu,N}) = \left\{-\frac{N}{2}, -\frac{N}{2}+1, \cdots, \frac{N}{2}-1, \frac{N}{2}\right\}, \quad \sigma\left(\hat{S}_N^2\right) = \left\{|m_z|(|m_z|+1) \;\middle|\; m_z \in \sigma(\hat{S}_{z,N})\right\}. \tag{1.20}$$

Looking back at the Hamiltonian, we see that $\hat{\mathrm{H}}_N$ does not explicitly involve the electronic spin components. For that reason, $\hat{\mathrm{H}}_N$ commutes with all spin observables, and it can be shown that $\hat{\mathrm{H}}_N$, $\hat{S}_N^2$ and $\hat{S}_{z,N}$ form a CSCO as described in Section 1.1.

1.3 The many-body electronic ground state problem

1.3.1 General formulation

The N-electron ground-state problem consists in solving the optimization problem

$$\mathcal{E}_* := \inf\left\{\mathcal{E}(\psi) = \langle\psi|\hat{\mathrm{H}}_N|\psi\rangle \;\middle|\; \psi \in \mathcal{H}_N, \quad \|\psi\| = 1\right\}. \tag{1.21}$$

If the spectrum of $\hat{\mathrm{H}}_N$ is bounded below, then (1.21) admits a solution $\mathcal{E}_* > -\infty$, called the ground state energy, which can be of two distinct kinds.

1. First, $\mathcal{E}_*$ can be an eigenvalue of $\hat{\mathrm{H}}_N$. Then by definition there exists a state $\psi_* \in \mathcal{H}_N$ solution of the eigenvalue problem
$$\hat{\mathrm{H}}_N\psi_* = \mathcal{E}_*\psi_*. \tag{1.22}$$
By the Rayleigh-Ritz formula, it is equivalent to the fact that the infimum in (1.21) is attained with $\mathcal{E}_* = \mathcal{E}(\psi_*)$. The wave-function ψ_* is called a ground state of the system X_N. Since ψ_* has a finite $\mathcal{H}_N$ norm, it is often referred to as a *bound state*: it physically describes states localized in space, i.e. electrons that are trapped by the electronic potential generated by the nuclei.

2. Otherwise $\mathcal{E}_*$ is not an eigenvalue of $\hat{\mathrm{H}}_N$. By the above point, the infimum in (1.21) is not attained and $\mathcal{E}_*$ is the bottom of the continuous spectrum of $\mathcal{H}_N$. In addition, there are no wave-functions in $\mathcal{H}_N$ describing a state of energy $\mathcal{E}_*$. In some instances (see e.g. [Sim81]), there exists generalized eigenvectors ψ_*, that belong to some space larger that $\mathcal{H}_N$, which are solutions to the eigenvalue equation (1.22). In particular, this kind of function has an infinite $\mathcal{H}_N$ norm. Consequently, it proves useful to describe delocalized physical states, such as free electrons, and is sometimes called a *scattering state*.

An essential distinction between the two scenarios lies in the fact that numerical methods used to compute the spectrum of a bounded-below self adjoint operators differ significantly when applied to point spectrum or continuous spectrum. The problems considered in this manuscript always boil down to the first case.

1.3.2 The ground state problem for molecular systems

Let us now make one step further toward chemistry. Consider a molecule in $\mathbb{R}^3$, containing N electrons and M atoms whose nuclei have respective positions $R_1, \cdots, R_M \in \mathbb{R}^3$ and charges $Z_1, \cdots, Z_M \in \mathbb{N}^*$. In the molecular case, the potential V in the Hamiltonian contains the electron-nuclei Coulomb attraction and nuclei-nuclei Coulomb repulsion

$$V(r) = -\sum_{\alpha=1}^{M} \frac{Z_\alpha}{|R_\alpha - r|} + \sum_{1\leqslant\alpha<\beta\leqslant M} \frac{Z_\alpha Z_\beta}{|R_\alpha - R_\beta|} \tag{1.23}$$

In that setting, the molecular Hamiltonian writes

$$\hat{\mathrm{H}}_N = -\frac{1}{2}\sum_{i=1}^{N}\Delta_{r_i} - \sum_{i=1}^{N}\sum_{\alpha=1}^{M}\frac{Z_\alpha}{|R_\alpha - r_i|} + \sum_{1\leqslant i<j\leqslant N}\frac{1}{|r_i - r_j|} + \sum_{1\leqslant \alpha<\beta\leqslant M}\frac{Z_\alpha Z_\beta}{|R_\alpha - R_\beta|}. \tag{1.24}$$

Remark that the last term is independent of the electronic configuration, and acts as a simple shift in energies. The characterization of the spectrum of the molecular Hamiltonian, ignoring the electronic spin, is a classical result that can be found e.g. in [RS78, Chapter XIII], [Zhi60]. Let us also refer to the more recent work [Lew22]. Since the presence of electronic spin does not affect the nature of the spectrum of $\hat{\mathrm{H}}_N$, we can write the following

Theorem 1.1 (**Spectrum of $\hat{\mathrm{H}}_N$ in the molecular case.**).
Let us denote by $Z = \sum_{\alpha=1}^{M} Z_\alpha$ the total charge of the system and suppose that $N \leqslant Z$. Then

- *the molecular Hamiltonian $\hat{\mathrm{H}}_N$ is self adjoint on $\mathcal{H}_N$ with domain $\mathcal{H}_N \bigcap H^2\left((\mathbb{R}^3 \times \{\uparrow,\downarrow\})^N;\mathbb{C}\right)$, and its spectrum is bounded from below (from [Lew22, Theorem 6.1]);*
- *there exists $\Sigma_N \in \mathbb{R}$ such that the essential spectrum of $\hat{\mathrm{H}}_N$ is $\sigma_{\mathrm{ess}}(\hat{\mathrm{H}}_N) = [\Sigma_N, +\infty[$ (from [Lew22, Theorem 6.5]);*
- *there exists an infinity of eigenvalues $(\varepsilon_{i,N})_{i\in\mathbb{N}}$ of finite multiplicity below the essential spectrum, accumulating at Σ_N. (from [Lew22, Theorem 6.14]).*

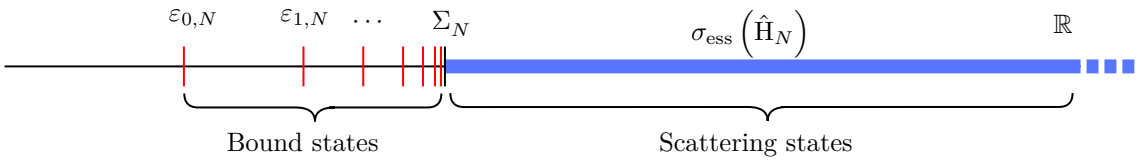


Figure 1 – Illustration of Theorem 1.1. If $N \leqslant Z$, the spectrum of $\hat{\mathrm{H}}_N$ is bounded below. The discrete spectrum consists of a sequence of eigenvalues $(\varepsilon_{i,N})$, which accumulate at the bottom of the essential spectrum Σ_N.

From Theorem 1.1, a molecular system possesses a ground state $\psi_* \in \mathcal{H}_N$ as long as the number of electrons N does not exceed the total charge Z. In its current form, however, the ground state problem has a complexity that grows exponentially with the number of electrons, making it unsolvable in practice. This natural barrier, known as the *curse of dimensionality*, motivates the introduction of approximation methods in the following Section 2.

Remark 1.1 (Excited states). *The molecular Hamiltonian possesses an infinite number of eigenvalues below its essential spectrum, hence an infinite number of* bound states. *Apart from the ground state, which corresponds to the smallest eigenvalue, these states are called excited states. They describe physical states of the system that can be reached after an excitation by a external source of energy (in the approximation when the atoms stay fixed).*

1.3.3 The case of the non-interacting-electron atom

Yet, there exists an important case in which eigenvalues and eigenvectors can be computed analytically. This is the set of N non-interacting electrons, subject to the attraction of a single nucleus of charge Z, i.e.:

$$V(r) = -\frac{Z}{|r|} \quad \text{and} \quad \hat{W}_{\mathrm{ee}} = 0. \tag{1.25}$$

The most simple example $N = Z = 1$, corresponds to the hydrogen atom. From (1.25), the N-electron Hamiltonian for the non-interacting-electron atom is the operator defined on $\mathcal{H}_N$ as

$$\hat{\mathrm{H}}_N = \sum_{i=1}^{N} \left(-\frac{1}{2}\Delta_{r_i} - \frac{Z}{|r_i|} \right) = \sum_{i=1}^{N} \mathbf{1} \otimes \cdots \otimes \underbrace{\hat{\mathrm{H}}_1(Z)}_{i-th} \otimes \cdots \otimes \mathbf{1}. \tag{1.26}$$

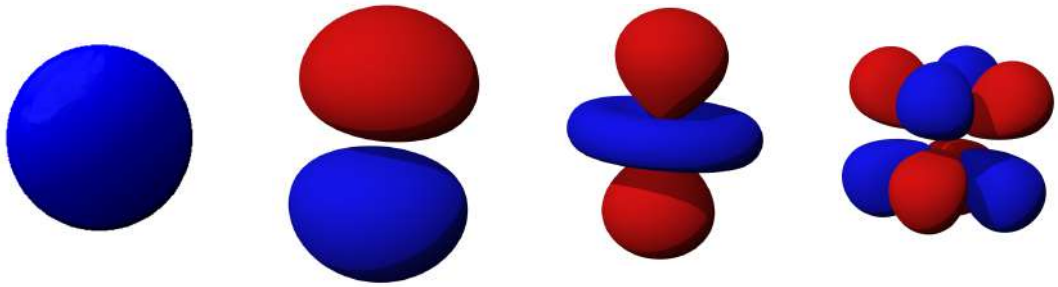
As seen in the above right-most expression, the electronic properties of the non-interacting-electron atom are encoded in the one-electron Hamiltonian $\hat{\mathrm{H}}_1(Z) = -\frac{1}{2}\Delta - \frac{Z}{|r|}$. From Theorem 1.1, the discrete spectrum of $\hat{\mathrm{H}}_1(Z)$ is an infinite sequence of eigenvalues. Since $\hat{\mathrm{H}}_1(Z)$ is radially symmetric, in the sense that it commutes with all the spatial rotations, the *bound states* of $\hat{\mathrm{H}}_1(Z)$ can be computed in spherical coordinates, as a product of a radial part and an angular part [CLBM06]

$$\psi_{nlm\sigma}(r, \theta, \phi, \sigma') = e^{-Zr} R_{nl}(Zr) Y_{lm}(\theta, \phi) \delta_{\sigma\sigma'}, \tag{1.27}$$

where $n \in \mathbb{N}^*$, $l \in \{0, 1, \ldots, n-1\}$, $m \in \{-l, -l+1, \ldots, l-1, l\}$ and $\sigma \in \{\uparrow, \downarrow\}$. The Y_{lm} are the spherical harmonics and each R_{nl} is a polynomial of degree $n - l - 1$. For a fixed $n \in \mathbb{N}^*$, all the states $\psi_{nlm\sigma}$ belong to the same eigenspace of energy

$$\varepsilon_n = -\frac{Z}{2n^2}. \tag{1.28}$$

In quantum chemistry, these eigenvectors are called *atomic orbitals* (AOs). They are the building blocks of atomic basis sets, presented in the Section 3.4 of this introduction, which are the most popular reduced bases for the resolution of molecular ground state problems. In chemistry, AOs are labeled using the first quantum number n, followed by a letter corresponding to the quantum number l with the rule: $l = 0 \longrightarrow$ s, $l = 1 \longrightarrow$ p, $l = 2 \longrightarrow$ d, $l = 3 \longrightarrow$ f, etc. Some s, p, d and f type of AOs are shown in Figure 2.



(a) 1s orbital (b) 2p orbital (c) 3d orbital (d) 4f orbital

Figure 2 – Illustration of the first four type of atomic orbitals, denoted with the standard (spdf…) convention. Each plot is a level set, where the colors represent the sign of the function. *Source: adapted from Wikipedia Commons.*

From (1.26) the ground-state of the N-electron Hamiltonian $\hat{\mathrm{H}}_N$ is obtained by summing the smallest N eigenvalues $\varepsilon_1 \leqslant \cdots \leqslant \varepsilon_N$ of $\hat{\mathrm{H}}_1(Z)$. Another way to write the ground-state energy is to introduce a chemical potential μ defined by

$$\mu = \frac{1}{2}(\varepsilon_{\mathrm{HO}} + \varepsilon_{\mathrm{LU}}) \tag{1.29}$$

where $\varepsilon_{\mathrm{HO}}$ and $\varepsilon_{\mathrm{LU}}$ are the respective energies of the highest occupied (HO) state and lowest unoccupied (LU) (or virtual) state. When the ground state is non-degenerate (i.e. $\varepsilon_{\mathrm{HO}} < \varepsilon_{\mathrm{LU}}$) the associated energy reads as the sum

$$\mathcal{E}_* = \sum_{n \in \mathbb{N}} \varepsilon_n \mathbb{1}(\varepsilon_n \leqslant \mu). \tag{1.30}$$

Otherwise, if N_μ denotes the number of electrons in the system with energy μ, then

$$\mathcal{E}_* = \sum_{n \in \mathbb{N}} \varepsilon_n \mathbb{1}(\varepsilon_n < \mu) + N_\mu \mu. \tag{1.31}$$

This is known as the *Aufbau* principle. It justifies the common representation that, in the ground state of a system X_N, the electrons "fill" the first energy levels, pictured as the rungs of a ladder.

1.3.4 The ground state problem for solid material

The ground state problem is way more challenging for solid materials, where the number of variable virtually tends to infinity. As an example, a one micrometer cube of silicon contains roughly 10^{10} atoms and of the order of 10^{11} electrons. This renders the formulation and analysis of any N-electron Hamiltonian or wave-function practically impossible. A way around the problem relies on the introduction of additional approximations, as well as the Bloch transform, a mathematical tool whose description is postponed to Section 4.

2 Variational approximations of ground states

In order to get around the *curse of dimensionality*, quantum chemists and physicists have produced numerous solvable approximations of the ground state problem (1.21). The most classical approximations fall into four categories: wave-function methods (WFM), methods based on popular density functional theory (DFT), quantum Monte-Carlo methods (QMC) and Green's function methods (GFM). In this manuscript, we have focused on variational WFM and DFT methods, of which we provide a brief introduction bellow. The central object of interest is the wave function of the system in the first category and the electronic density in the second one. For a broader view, we advise the reader to consult the literature, in particular the books [CLBM06; LL19; CF23] for a general mathematical introduction. We also refer to the comprehensive chemistry "pink bible" [HJO14] for wave-functions methods and to the review [Tou22] for density functional theory.

2.1 Variational Wave-function methods

2.1.1 Configuration Interaction and Multi-configuration problems

In variational wave-functions methods, the energy in (1.21) is minimized on a finite dimensional subspace of $\mathcal{H}_N$. Defining this variational space amounts to selecting a *discretization basis* (understand an orthonormal family of finite rank) of $\mathcal{H}_N$, which can be built systematically from discretization basis sets of $\mathcal{H}_1$. Indeed, let $\Phi := (\phi_i)_{i\in\mathbb{N}}$ be an orthonormal basis of $\mathcal{H}_1$ and

$$\mathcal{I}_N := \left\{(i_1, \cdots, i_N) \in \{1, \dots, \mathbb{N}\}^N, \quad i_1 < \cdots < i_N\right\}. \tag{2.1}$$

For a given $I \in \mathcal{I}_N$, let $\Phi_I := \phi_{i_1} \wedge \cdots \wedge \phi_{i_N}$ abbreviate the Slater determinant of $(\phi_{i_1}, \dots, \phi_{i_N})$, as defined in (1.12). From (1.13) (see e.g. [Lew04, Lemma 1]), all normalized wave-function $\psi \in \mathcal{H}_N$ can be expanded as an infinite sum of Slater determinants of N basis functions in Φ

$$\boxed{\psi = \sum_{I\in\mathcal{I}_N} \lambda_I \Phi_I, \qquad \text{with} \sum_{I\in\mathcal{I}_N} |\lambda_I|^2 = 1.} \tag{2.2}$$

As often in the chemistry literature, we will refer to the functions ϕ_i as *molecular orbitals*, since they serve as the analogue, for molecular systems, to the atomic orbitals introduced for the non-interacting-electron atom (Section 1.3.3).

From this point, approximations of ψ can be constructed either by taking a discretization basis set of Φ of rank $K \geqslant N$, or by truncating the sum, selecting a finite set of Slater determinants in $\mathcal{I}_N$. For all $K \geqslant N$, let

$$\begin{aligned} \mathcal{B}_K &:= \left\{\Phi = (\phi_1, \cdots, \phi_K) \in \mathcal{H}_1^K, \quad \langle\phi_i|\phi_j\rangle_{\mathcal{H}_1} = \delta_{ij} \quad \forall 1 \leqslant i, j \leqslant K\right\}, \\ \mathcal{I}_N^K &:= \left\{(i_1, \cdots, i_N) \in \{1, \cdots, K\}^N, \quad i_1 < \cdots < i_N\right\}. \end{aligned} \tag{2.3}$$

Given a basis $\Phi \in \mathcal{B}_K$ and a set of determinants $\mathcal{J} \subset \mathcal{I}_N^K$, the variational space we obtain writes

$$\mathcal{W}_N^K(\Phi, \mathcal{J}) := \left\{\psi = \sum_{I\in\mathcal{J}} \lambda_I \Phi_I, \quad \lambda_I \in \mathbb{C}, \quad \sum_{I\in\mathcal{J}} |\lambda_I|^2 = 1\right\}, \tag{2.4}$$

with dimension $\dim(\mathcal{W}_N^K) = |\mathcal{J}| \leqslant \binom{K}{N}$. The corresponding ground state energy is called the *Configuration Interaction* (CI) energy

$$\boxed{\mathcal{E}_{\mathrm{CI}}(\Phi, \mathcal{J}) := \min_{\psi \in \mathcal{W}_N^K(\Phi,\mathcal{J})} \langle \psi | \, \hat{\mathrm{H}}_N \, | \psi \rangle.} \tag{2.5}$$

By construction, it holds

$$\mathcal{E}_{\mathrm{CI}}(\Phi, \mathcal{J}) \geqslant (\mathcal{E}_{\mathrm{exact}})_* \tag{2.6}$$

and the approximation is exact in the limit $\mathcal{J} = \mathcal{I}_N^K$, $K \to +\infty$, which is standard with Galerkin methods.

For a fixed basis $\Phi = (\phi_i)_{1 \leqslant i \leqslant K}$, the list of determinants $\mathcal{J}$ is either set by a systematic procedure, as in the CISD (*CI Single Double*) approach [HJO14, Section 5.6], or more general selected-CI (see [Eva14] and references therein). Sometimes, it is chosen by hand, following chemical intuition and experiments. The Full-CI (FCI) method uses the complete set of determinant $\mathcal{J} = \mathcal{I}_N^K$. It is the best possible approximation for a given basis Φ, though the factorial growth of the number of determinants with the number K makes it challenging to use in practice. For these reasons, selecting an expansion that is both accurate and computationally efficient is an ongoing area of research in quantum chemistry.

Of course, the quality of the CI method depends on the chosen basis set $\Phi = (\phi_i)_{1 \leqslant i \leqslant K}$. For a fixed expansion $\mathcal{J}$, the best approximation is obtained by optimizing over all possible basis sets $\Phi \in \mathcal{B}_K$, which reads as a nested minimization problem called the *Multi-Configuration Self-Consistent Field* problem

$$\boxed{\min_{\Phi \in \mathcal{B}_K} \min_{\Lambda \in \mathbb{C}^{|\mathcal{J}|}} \mathcal{E}_{\mathrm{MCSCF}}(\Phi, \Lambda) := \left\{ \langle \psi \, | \, \mathrm{H}_N \, | \, \psi \rangle \;\middle|\; \Lambda = [\lambda_I]_{I \in \mathcal{J}}, \quad \psi = \sum_{I \in \mathcal{J}} \lambda_I \Phi_I, \quad \|\Lambda\|^2_{\mathbb{C}^{|\mathcal{J}|}} = 1 \right\}.} \tag{2.7}$$

The above presentation is inspired by [Lew04], which provides a mathematical proof of the existence of solutions to the multi-configuration problem.

Remark 2.1. *Beware that, in quantum chemistry, the term* self-consistent field (SCF) *refers both to the general problem of optimizing the molecular orbitals, i.e. the outer minimization appearing in* (2.7)*, and to a class of numerical methods to solve this problem. The SCF class of numerical methods is the subject of Section 3.3*

2.1.2 The Hartree-Fock approximation

The Hartree-Fock (HF) method can be seen as the simplest truncation of the MCSCF problem. It consists in solving (2.7) for single determinant wave-functions (i.e. setting $K = N$). Although this might seem a crude approximation, the Hartree-Fock ground state is exact in the case $\hat{W}_{\mathrm{ee}} = 0$, where the electrons are non-interacting. For interacting electrons, the accuracy of the HF approximation is measured by the correlation energy

$$\mathcal{E}_{\mathrm{c}} = (\mathcal{E}_{\mathrm{exact}})_* - (\mathcal{E}_{\mathrm{HF}})_* < 0. \tag{2.8}$$

If $\mathcal{E}_{\mathrm{c}}$ is very small, the system is *weakly correlated*, and the Hartree-Fock determinant is already a good approximation of the ground state. Otherwise, the system is said to be *strongly correlated*, and one needs to resort to other methods to capture the missing correlation.

Let $\Phi = (\phi_1, \cdots, \phi_N) \in \mathcal{B}_N$ be a given discretization basis. Setting $\psi = \phi_1 \wedge \cdots \wedge \phi_N$ in (2.7), a standard computation provides the Hartree-Fock energy

$$\boxed{\mathcal{E}_{\mathrm{HF}}(\Phi) = \sum_{i=1}^{N} \int_{\mathbb{R}^3 \times \{\uparrow,\downarrow\}} |\nabla \phi_i|^2 + \int_{\mathbb{R}^3} \rho_\psi V + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho_\psi(r) \rho_\psi(r')}{|r - r'|} \mathrm{d}r \mathrm{d}r' - \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{|\gamma_\psi(r, r')|^2}{|r - r'|} \mathrm{d}r \mathrm{d}r'} \tag{2.9}$$

with the density matrix of order 1 and the electronic density

$$\gamma_\psi(r, r') = \sum_{\sigma \in \{\uparrow,\downarrow\}} \sum_{i=1}^{N} \overline{\phi_i(r', \sigma)} \phi_i(r, \sigma), \quad \rho_\psi(r) = \gamma_\psi(r, r). \tag{2.10}$$

2.1.3 Separate molecular orbitals

The formulation we have given for the MCSCF (2.7) and HF (2.9) problems are sometimes called General MCSCF (GMCSCF) and General HF (GHF), since the molecular orbitals ϕ_i can be any functions of $\mathcal{H}_1$. In our specific case, where the Hamiltonian $\hat{\mathrm{H}}_N$ is time-reversal symmetric and commutes with the spin operators, we can assume without any loss of generality that the functions ϕ_i are real-valued, and separate the spatial and spin coordinates

$$\forall (r,\sigma) \in \mathbb{R}^3 \times \{\uparrow,\downarrow\}, \quad \forall i \in \{1,\cdots,N\}, \qquad \phi_i(r,\sigma) = \varphi_i(r)\tau_i(\sigma), \tag{2.11}$$

which we can also write $\phi_i = \varphi_i \otimes \tau_i$. The spatial part is a real square integrable function $\varphi_i \in L^2(\mathbb{R}^3;\mathbb{R})$ and the spin part τ_i is equal to either $\alpha := \sigma \mapsto \delta_{\uparrow\sigma}$ or $\beta := \sigma \mapsto \delta_{\downarrow\sigma}$ (this case where τ_i takes its values in $\{0,1\}$ is also known as *collinear spins*).

2.1.4 Imposing the right symmetry with Configuration State Functions.

We have seen in the Section 2 that the exact N-electron ground state is an eigenfunction of the spin observables $\hat{S}_N^2$ and $\hat{S}_{z,N}$. While Slater determinants are eigenfunctions of $\hat{S}_{z,N}$, they are not in general eigenfunctions of the total spin operator $\hat{S}_N^2$ [HJO14, Chapter 2]. For that reason, approximate CI wavefunctions, built as a sum of Slater determinants, do not necessarily satisfy the proper spin symmetry that is expected for the true ground state.

To look for a solution in a targeted spin configuration $(s,m_z) \in \sigma(\hat{S}_N^2) \times \sigma(\hat{S}_{z,N})$, a natural strategy is to expand the N-electron wave function ψ on a basis $(\psi_k^{(s,m_z)})_{1\leqslant k\leqslant L}$ of L spin-eigenfunctions

$$\psi = \sum_{k=1}^{L} c_k \psi_k^{(s,m_z)} \tag{2.12}$$

where for all $1 \leqslant k \leqslant L$

$$\hat{S}_N^2 \psi_k^{(s,m_z)} = s(s+1)\psi_k^{(s,m_z)} \quad \text{and} \quad \hat{S}_{z,N}\psi_k^{(s,m_z)} = m_z \psi_k^{(s,m_z)}. \tag{2.13}$$

The functions $\psi_k^{(s,m_z)}$ are called *Configuration State Functions* (CSFs) for the spin configuration (s,m_z). The CI states obtained as a sum of CSFs are called *spin-restricted*, while they are labeled *spin-unrestricted* for Slater determinants. Configuration state functions provide an alternative basis for the CI approach, that produces states with the proper symmetry, at the cost of being more complicated to handle.

In practice the two approaches, using respectively Slater determinants or CSFs, are linked. Configuration state functions can be systematically built from a basis $\Phi \in \mathcal{B}_K$ of molecular spin-orbitals, as the sum of several determinants

$$\psi_k^{(s,m_z)} = \sum_{I\in\mathcal{J}_k} \lambda_I^k \Phi_I \tag{2.14}$$

where the coefficients $(\lambda_I^k)_{I\in\mathcal{J}_k}$ and the expansion $\mathcal{J}_k$ are imposed by the spin symmetry requirement (see [HJO14, Section 2.6] on the systematic construction of CSFs with the genealogical coupling scheme). For separate spin orbitals (2.11), the spin restrictions (2.13) directly translate as restrictions on the spatial and spin parts of the basis Φ. Indeed let $\Phi = (\varphi_1 \otimes \tau_1, \cdots, \varphi_K \otimes \tau_K)$ and consider a CSF $\psi^{(s,m_z)}$ as in (2.14). We introduce for all determinants $\Phi_I = \varphi_{i_1}\otimes\tau_{i_1} \wedge \ldots \wedge \varphi_{i_N}\otimes\tau_{i_N}$ the occupation number operators

$$\forall\, i \in \{1,\cdots,K\}, \quad \forall\, \tau \in \{\alpha,\beta\}, \quad \hat{N}_{i\tau}\Phi_I = n_{i\tau}\Phi_I, \quad n_{i\tau} = \left| \begin{array}{ll} 1 & \text{if } \exists n \in I \text{ s.t. } \varphi_n = \varphi_i \text{ and } \tau_n = \tau \\ 0 & \text{otherwise} \end{array} \right. \tag{2.15}$$

and $\hat{N}_i = \hat{N}_{i\alpha} + \hat{N}_{i\beta}$. The operators $\hat{N}_i$ commute with $\hat{S}^2$ and S_z [HJO14], so that one can impose that $\psi^{(s,m_z)}$ is a common eigenfunction of all $\hat{N}_i$. Therefore $\psi^{(s,m_z)}$ can be associated to an occupation vector $n(\psi^{(s,m_z)}) = (n_1,\ldots,n_K) \in \{0,1,2\}^K$ such that

$$\sum_{i=1}^{K} n_i = N \quad \text{and} \quad \hat{N}_i\psi^{(s,m_z)} = n_i\psi^{(s,m_z)}. \tag{2.16}$$

Going one step further, consider an approximate CI state in the spin-restricted setting

$$\psi = \sum_{k=1}^{L} c_k \psi_k^{(s,m_z)} \tag{2.17}$$

for a given basis $(\psi_1^{(s,m_z)}, \dots, \psi_L^{(s,m_z)})$ of CSFs. Up to reordering the molecular orbitals, one can assume that:

$$\boxed{\begin{aligned} &\forall\, 1 \leqslant i \leqslant N_i, && \forall k \in \{1,\dots,L\}, && [n(\psi_k^{(s,m_z)})]_i = 2, \\ &\forall\, N_i + 1 \leqslant i \leqslant N_i + N_a, && \exists k \in \{1,\dots,L\}, && [n(\psi_k^{(s,m_z)})]_i = 1, \\ &\forall\, N_i + N_a + 1 \leqslant i \leqslant K, && \forall k \in \{1,\dots,L\}, && [n(\psi_k^{(s,m_z)})]_i = 0. \end{aligned}} \tag{2.18}$$

Then the first $2N_i$ spin-orbitals of Φ are called *doubly-occupied* or *internal orbitals*. They verify

$$\forall\, 1 \leqslant i \leqslant N_i, \quad \varphi_{2i-1} = \varphi_i, \quad \tau_{2i-1} = \alpha \quad \text{and} \quad \tau_{2i} = \beta. \tag{2.19}$$

The next N_a orbitals are referred to as *active* orbitals, without specific constraints. They can be separated into N_α spin-up and N_β spin-down orbitals

$$\begin{aligned} &\phi_i = \varphi_i \otimes \alpha \quad \forall\, 2N_i + 1 \leqslant i \leqslant N_i + N_\alpha, \\ &\phi_i = \varphi_i \otimes \beta \quad \forall\, 2N_i + N_\alpha + 1 \leqslant i \leqslant N_i + \underbrace{N_\alpha + N_\beta}_{N_a}. \end{aligned} \tag{2.20}$$

Finally, the last orbitals are known as *external* or *virtual* orbitals, which are unoccupied for every configuration. The numbers N_i, N_a, s and m_z verify

$$2N_i + N_a = N \quad \text{and} \quad |m_z| \leqslant s \leqslant \frac{1}{2} N_a. \tag{2.21}$$

Configuration state functions are generally the sum of several determinants. In this manuscript, we will restrict to the particular *high-spin* and *closed-shell* cases, respectively defined by $s = m_z = \frac{1}{2}N_a$ and $s = 0$, for which single Slater determinants are CSFs. The generalization to arbitrary spin-states is a straightforward yet tedious exercise (see again [HJO14, Section 2.6]).

2.1.5 Unrestricted and spin-restricted Hartree-Fock

Additional information can be given when it comes to the Hartree-Fock method in spin-restricted setting. Since the HF ground state is a single Slater determinant, all the $N_a = N_\alpha + N_\beta$ active orbitals are necessarily singly-occupied. The high-spin *spin-restricted* Hartree-Fock corresponds to the choice $N_\beta = 0$. It is called Restricted Hartree-Fock (RHF) for $N_\alpha = 0$ and Restricted Open-shell Hartree-Fock (ROHF) for $N_\alpha > 0$. The Unrestricted Hartree-Fock (UHF) corresponds to $N_i = 0$. The spin constraints on molecular orbitals in the generalized, spin-restricted and spin-unrestricted Hartree-Fock models are pictured in the Figure 3.

Inserting (2.19) and (2.20) into (2.9), we see that the Hartree-Fock energy in the spin-restricted setting only depends on the $N_i + N_a$ first spatial parts of the molecular orbitals, through the the *internal* and *active* density matrices:

$$\gamma_i(r,r') = 2\sum_{i=1}^{N_i} \varphi_i(r')\varphi_i(r), \quad \gamma_a(r,r') = \sum_{i=N_i+1}^{N_a} \varphi_i(r')\varphi_i(r). \tag{2.22}$$

2.1.6 Post Hartree-Fock methods

Even in the case of *strongly* correlated system, it is common practice to use a Hartree-Fock ground state Φ_{HF} as a starting guess for other, more elaborate methods, which are then commonly referred to as *Post Hartree-Fock* methods (PHF).

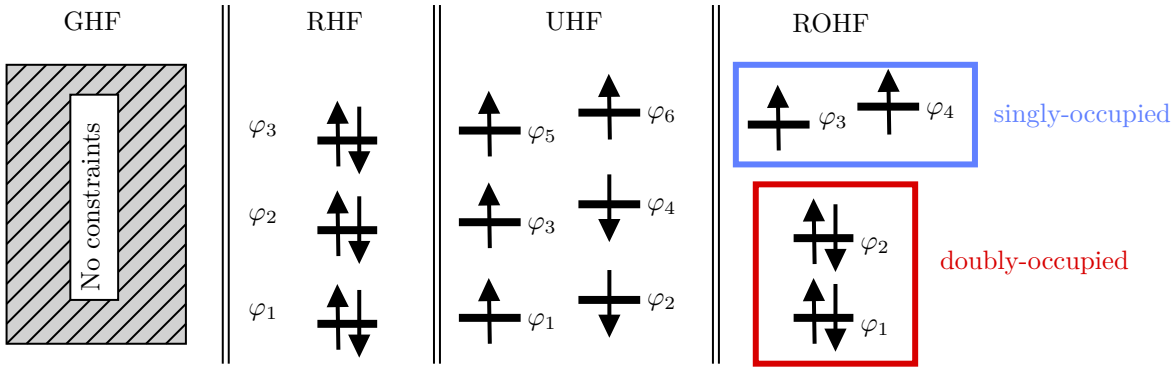


Figure 3 – Pictorial representation of the constraints on molecular orbitals in the generalized, spin-restricted and spin-unrestricted Hartree-Fock models. The RHF and ROHF models correspond to the spin-restricted Hartree-Fock in the closed-shell ($s = 0$) and open-shell ($s = m_z > 0$) settings. The α and β spin-parts are respectively pictured as up and down arrows.

Variational PHF methods simply consist in solving a MCSCF problem starting from a basis optimal at the Hartree-Fock level. For example, the popular *Complete Active Space* SCF (CASSCF) divides some HF orbitals into internal, active and virtual orbitals. The wave-functions is then expanded on the set of CSFs constructed by distributing a number N_a^e of valence (or active) electrons among the active orbitals in all possible ways, which is usually abbreviated to CAS(N_v^e,N_a).

Another important family of PHF methods are non variational approaches. Some methods, as Møller-Plesset theory (MP) [HJO14], rely on perturbation theory to compute correction terms to the HF energy. The most successful non-variational PHF method for the calculation of energies is the *single reference coupled cluster* (CC) [BM07; HMW23], that provides a non linear approximation of the true ground state, as the result of the action of a parameter-dependent operator on Φ_{HF} (note however that other reference wave-functions can be used). The coupled cluster operator is built as the exponential of a linear combination of so-called excitation operators, whose truncation gives rise to various flavors of the CC methods.

2.2 Methods based on Density Functional Theory

Exact Density Functional Theory (DFT) offers a theoretical tool for computing the ground state energy (1.21) that relies solely on one-body electronic densities, leading to a dramatic reduction in the dimensionality of the problem. Applicable to a broad range of systems, this method relies on the introduction of a universal functional of which no explicit expression is known. Various approximations of this exact functional have been suggested, with the Kohn-Sham DFT being the most popular, giving rise to a wide variety of approximate DFT methods.

2.2.1 Exact DFT

The following exposition is based on [Tou22]. The main results are the work of Hohenberg and Kohn [HK64], Levy [Lev79] and Lieb [Lie02]. We define the set of N-representable densities as

$$\mathscr{I}_N = \left\{ \rho \in L^1(\mathbb{R}^3;\mathbb{C}), \quad \rho \geqslant 0, \quad \sqrt{\rho} \in H^1(\mathbb{R}^3;\mathbb{C}), \quad \int_{\mathbb{R}^3} \rho = N \right\} \tag{2.23}$$

For a given $\rho \in \mathscr{I}_N$, let

$$\begin{aligned} \mathcal{H}_N^\rho = \Big\{ & \psi \in \mathcal{H}_N \cap H^1\left((\mathbb{R}^3 \times \{\uparrow,\downarrow\})^N;\mathbb{C}\right), \\ & \rho_\psi(r) := N \int_{\{\uparrow,\downarrow\}\times(\mathbb{R}^3\times\{\uparrow,\downarrow\})^{N-1}} |\psi\left((r,\sigma),\mathbf{x}_2,\cdots,\mathbf{x}_N\right)|^2 \mathrm{d}\sigma\mathrm{d}\mathbf{x}_1\cdots\mathrm{d}\mathbf{x}_N = \rho \Big\}. \end{aligned} \tag{2.24}$$

be the set of wave-functions ψ with density ρ. For all $\rho \in \mathscr{I}_N$, the set $\mathcal{H}_N^\rho$ is not empty, and we can define the Levy-Lieb density functional as the mapping $F : \mathscr{I}_N \to \mathbb{R}$ defined by

$$F(\rho) = \min_{\psi \in \mathcal{H}_N^\rho} \langle \psi | \hat{T} + \hat{W}_{\text{ee}} | \psi \rangle = \langle \psi[\rho] | \hat{T} + \hat{W}_{\text{ee}} | \psi[\rho] \rangle \tag{2.25}$$

where $\hat{T}$ and $\hat{W}_{\text{ee}}$ are the N-electron Hamiltonian terms defined (1.14). The minimizer $\psi[\rho] \in \mathcal{H}_N^\rho$ exists but is not necessary unique [Lie02; LLS22]. The universal functional allows to re-write the ground state problem (1.21) as a minimization problem on $\mathscr{I}_N$

$$\mathcal{E}_* = \min_{\rho \in \mathscr{I}_N} \left\{ F(\rho) + \int_{\mathbb{R}^3} \rho V \right\}. \tag{2.26}$$

This formulation proves highly advantageous. In (2.26), the ground state energy is found by minimization over densities, which are functions of 3 variables, instead of many-body wave-functions. In particular, the dimensionality of the DFT ground state problem is independent of the number of electrons. However, there are no explicit formula for the universal functional F. The most popular scheme to circumvent the problem has been proposed by Kohn and Sham.

2.2.2 Kohn-Sham DFT

The idea of Kohn-Sham density functional theory (KS-DFT) is to decompose the universal functional F as the sum of an explicit term, computed in a single determinant approximation, plus a correction energy which can be approximated.

Let us introduce the set of Slater determinants $\mathcal{S}_N^\rho \subset \mathcal{H}_N^\rho$ with density ρ. Again, from [Lie02], this set is not empty and the minimization problem in the definition (2.25) of the universal functional F restricted to $\mathcal{S}_N^\rho$, has a solution $\Phi[\rho]$

$$F_{\mathcal{S}}(\rho) = \min_{\Phi \in \mathcal{S}_N^\rho} \langle \Phi | \hat{T} + \hat{W}_{\text{ee}} | \Phi \rangle = \langle \Phi[\rho] | \hat{T} + \hat{W}_{\text{ee}} | \Phi[\rho] \rangle. \tag{2.27}$$

The energy (2.26) is explicit by replacing F with $F_{\mathcal{S}}$ and is simply given by the Hartree-Fock energy (2.9) evaluated at $\Phi[\rho]$. As for Hartree-Fock, this approximation is exact for non-interacting electrons. In the general case, the inclusion $\mathcal{S}_N^\rho \subset \mathcal{H}_N^\rho$ implies

$$F(\rho) \leqslant F_{\mathcal{S}}(\rho) \tag{2.28}$$

and the difference between the non-interacting electronic system and the real one is measured by the correlation functional

$$\mathcal{E}_{\text{c}}(\rho) = \langle \psi[\rho] | \hat{T} + \hat{W}_{\text{ee}} | \psi[\rho] \rangle - \langle \Phi[\rho] | \hat{T} + \hat{W}_{\text{ee}} | \Phi[\rho] \rangle \leqslant 0. \tag{2.29}$$

In KS-DFT, the universal functional is written in terms of the correlation energy as

$$F(\rho) = \sum_{i=1}^N \int_{\mathbb{R}^3 \times \{\uparrow,\downarrow\}} |\nabla \phi_i(\mathbf{x})|^2 \mathrm{d}\mathbf{x} + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(r)\rho(r')}{|r-r'|} \mathrm{d}r\mathrm{d}r' \underbrace{- \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{|\gamma_{\Phi[\rho]}(r,r')|^2}{|r-r'|} \mathrm{d}r\mathrm{d}r' + \mathcal{E}_{\text{c}}(\rho)}_{\mathcal{E}_{\text{xc}}(\rho)} \tag{2.30}$$

where the last two terms have been gathered in the *exchange-correlation* (XC) functional $\mathcal{E}_{\text{xc}}(\rho)$. With that expression, the KS-DFT minimization problem reads as a minimization on orbitals

$$\boxed{(\mathcal{E}_{\text{KS}})_* = \min \left\{ \mathcal{E}_{\text{KS}}(\Phi),\ \Phi \in \mathcal{B}_N \right\},} \tag{2.31}$$

with $\mathcal{B}_N$ as in (2.3) and where

$$\mathcal{E}_{\text{KS}}(\Phi) = \sum_{i=1}^N \int_{\mathbb{R}^3 \times \{\uparrow,\downarrow\}} |\nabla \phi_i(\mathbf{x})|^2 \mathrm{d}\mathbf{x} + \int_{\mathbb{R}^3} \rho_\Phi V + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho_\Phi(r)\rho_\Phi(r')}{|r-r'|} \mathrm{d}r\mathrm{d}r' + \mathcal{E}_{\text{xc}}(\rho_\Phi). \tag{2.32}$$

with ρ_Φ the density of the Slater determinant built from Φ. Minimizers $\Phi_* = (\phi_{1*}, \dots, \phi_{N*})$ for (2.32) are usually called *Kohn-Sham orbitals*, and the associated Slater determinants $\phi_{1*} \wedge \cdots \wedge \phi_{N*}$ are called *Kohn-Sham (KS) determinants*. The KS-DFT model for zero exchange-correlation $\mathcal{E}_{\mathrm{xc}} = 0$ is called *reduced Hartree-Fock* (rHF).

In KS-DFT, the difficulty of evaluating the functional F has been transferred on the XC functional, and the accuracy of the model depends on the approximation of $\mathcal{E}_{\mathrm{xc}}$. As for wave-function methods, one often imposes that the molecular spin-orbitals read as the product of a spatial part and a spin part, yielding to *spin-restricted* or *spin-unrestricted* KS-DFT.

2.2.3 Approximations of the KS exchange-correlation functional

There exists a wide-variety of approximate exchange-correlation functionals. They are usually classified according to the rungs of the so-called "Jacob's ladder" of exchange-correlation functionals (Figure 4), ranging in complexity from no exchange-correlation ($\mathcal{E}_{\mathrm{xc}} = 0$) at ground level, to exact exchange correlation at the top. The first rung of the ladder is the LDA functional [KS65]. Introduced by Kohn and Sham in 1965, it is defined as

$$\mathcal{E}_{\mathrm{xc}}^{\mathrm{LDA}}(\rho) = \int_{\mathbb{R}^3} e_{\mathrm{xc}}^{\mathrm{UEG}}(\rho(r))\mathrm{d}r, \tag{2.33}$$

where $e_{\mathrm{xc}}^{\mathrm{UEG}} : \mathbb{R}^+ \to \mathbb{R}$ is the exchange-correlation energy density of the infinite uniform electron gas (UEG). It is a local approximation in the sense that it only uses local values of the density ρ.

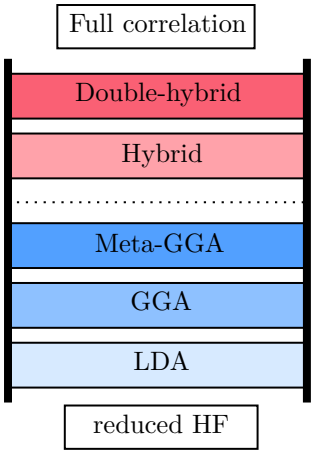


Figure 4 – Jacob's ladder of exchange-correlation functionals. It ranges from the reduced Hartree-Fock at ground level, to exact exchange-correlation at the heaven level. The first two rungs only depend on the density ρ and its derivatives. The next rungs also incorporate fractions of the exact exchange for the KS determinant $\Phi[\rho]$.

On the second and third rungs, the *Generalized Gradient Approximation* (GGA) and Meta-GGA approximations have energy densities that depend on the density ρ, but also on the first derivative of ρ for GGA, and first and second derivatives for Meta-GGA. For that reason they are often called semi-local approximations. Among the GGA approximation, let us cite the PBE functional [PBE96], widely used in solid state physics, and a central tool for the computations of the Part II of this manuscript.

Finally, the fourth and fifth rungs target the two main known deficiencies of local and semi-local approximations, which are the self-interaction error and the absence of long-range van der Waals interactions. Since they incorporate a fraction of the exact exchange energy of the HF determinant

$$\mathcal{E}_{\mathrm{x}}(\Phi_{\mathrm{HF}}) = -\frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{|\gamma_{\Phi_{\mathrm{HF}}}(r,r')|^2}{|r-r'|}\mathrm{d}r\mathrm{d}r', \tag{2.34}$$

they are referred to as hybrid-functionals. Again, we refer to [Tou22] for a complete review of approximate XC functionals.

On the ladder, the functional are sorted by increasing complexity, which should not be confused with increasing accuracy. There is no direct link between rung level and the overall accuracy of the approximations. Functionals are often optimized to target specific quantities or specific systems. For practical applications, it is necessary to refer to XC-functional benchmarks in order to choose functional(s) best suited for the task.

3 Solving the approximate ground state problems

In the preceding section, we presented a series of "solvable" problems, in the sense that their reduced dimensionality makes their numerical resolution feasible. In that regard, the next natural step is to translate these approximate models in a discrete framework, allowing for the use of general minimization algorithms.

We have seen that the approximate models consist in two minimization problems.

1. Given a reduced basis of rank $K \geqslant N$ and an expansion $\mathcal{J} \in \mathcal{I}_N^K$, the CI problem finds the coefficients $\Lambda_* \in \mathbb{C}^{|\mathcal{J}|}$ that minimize the CI energy.

2. For a given Λ, the self-consistent field problem finds the best basis $\Phi_* \in \mathcal{B}_N^K$ in terms of energy.

It is clear that only the second problem has to be discretized. In quantum chemistry, the elements of a discretization basis $\mathcal{X}$ for the SCF problem are usually *atomic orbitals* (AO), as introduced for the non-interacting-electron atom in Section 1.3.3. Further details on the construction of AO basis sets are provided in Section 3.4.

3.1 Discretization of the self-consistent field problem

3.1.1 Parametrizations of discrete states in molecular orbital and density matrices formalisms

In this manuscript, where the spatial and spin coordinates can be separated, only $N_i + N_a$ occupied spatial parts $\varphi_1, \ldots, \varphi_{N_i+N_a}$ of $L^2(\mathbb{R}^3; \mathbb{R})$ have to be optimized. Let $N_o = N_i + N_a$ and $\mathcal{X} = (\chi_1, \cdots, \chi_{N_b})$ be a discretization basis of $L^2(\mathbb{R}^d; \mathbb{R})$ of size $N_b \geqslant N_o$. After discretization, a trial wave-function ψ is represented by a matrix $C = (C_i \,|\, C_a) \in \mathbb{R}^{N_b \times N_o}$ containing the coefficients of the spatial parts φ_i in the basis $\mathcal{X}$:

$$\begin{aligned} \varphi_i(\mathbf{r}) &= \sum_{\mu=1}^{N_b} [C_i]_{\mu i} \chi_\mu(\mathbf{r}), \qquad && 1 \leq i \leq N_i, \\ \varphi_{N_i+i}(\mathbf{r}) &= \sum_{\mu=1}^{N_b} [C_a]_{\mu i} \chi_\mu(\mathbf{r}), \qquad && 1 \leq i \leq N_a. \end{aligned} \tag{3.1}$$

In practice, the χ_μ's are non-orthogonal atomic orbitals. In order to simplify the presentation, we will however assume here that the basis $\mathcal{X}$ is orthonormal, or equivalently that the overlap matrix is the identity matrix:

$$S_{\mu\nu} := \int_{\mathbb{R}^3} \chi_\mu(\mathbf{r}) \chi_\nu(\mathbf{r}) \, d\mathbf{r} = \delta_{\mu\nu}. \tag{3.2}$$

Let us emphasize that we make this simplification for pedagogical purposes only; extending our arguments to non-orthogonal basis sets is a simple exercise. In that setting, the orthonormality constraints on the orbitals imply that C is a rectangular orthogonal matrix; in other words, a point of the Stiefel manifold

$$\boxed{C \in \mathrm{St}(N_o; N_b) := \{C \in \mathbb{R}^{N_b \times N_o} \text{ s.t. } C^T C = I_{N_o}\}} \tag{3.3}$$

where I_{N_o} denotes the identity matrix in $\mathbb{R}^{N_o \times N_o}$. We call this formalism occupied MOs (OMO).

When the number N_b of basis functions is relatively small, it proves advantageous to keep track of additional $N_e = N_b - N_o$ virtual orbitals, although they do not contribute to the energy. The matrix

$C = (C_i|C_a|C_e) \in \mathbb{R}^{N_b \times N_b}$ becomes a point of the orthogonal group $\mathcal{O}_{N_b}$. We obtain the all MOs (AMO) formalism:

$$\boxed{C \;=\; (C_i|C_a|C_e) \;\in\; \mathcal{O}_{N_b}.} \tag{3.4}$$

The structure of $\mathcal{O}_{N_b}$ is simpler that the Stiefel manifold $\mathrm{St}(N_o; N_b)$. It is a smooth Lie group with Lie algebra $\mathbb{R}^{N_b \times N_b}_{\mathrm{skew}}$, the set of $N_b \times N_b$ real skew-symmetric matrices, so that there exists $\kappa \in \mathbb{R}^{N_b \times N_b}_{\mathrm{skew}}$ such that $C = \exp(\kappa)$. In quantum chemistry, the matrix κ is usually referred to as a *rotation operator*, and provides another parametrization for discrete states, equivalent to OMO

$$\boxed{C \;\in\; \left\{ e^{\kappa}, \quad \kappa \in \mathbb{R}^{N_b \times N_b}_{\mathrm{skew}} \right\}.} \tag{3.5}$$

Alternatively, a state can be represented by a pair of matrices $(P_i, P_a) \in \mathbb{R}^{N_b \times N_b} \times \mathbb{R}^{N_b \times N_b}$ collecting the coefficients of the one-particle density matrices (DM) on spatial parts,

$$\gamma_i = \sum_{i=1}^{N_i} |\varphi_i\rangle\langle\varphi_i| \quad \text{and} \quad \gamma_a = \sum_{i=N_i+1}^{N_o} |\varphi_i\rangle\langle\varphi_i| \tag{3.6}$$

in the basis set $\mathcal{X}$:

$$\gamma_i = \sum_{\mu,\nu=1}^{N_b} [P_i]_{\mu\nu} |\chi_\mu\rangle\langle\chi_\nu| \quad \text{and} \quad \gamma_a = \sum_{\mu,\nu=1}^{N_b} [P_a]_{\mu\nu} |\chi_\mu\rangle\langle\chi_\nu|.$$

A point $C \in \mathrm{St}(N_o; \mathbb{R}^{N_b})$ and a couple (P_i, P_a) are related by

$$P_i = C_i C_i^T, \quad P_a = C_a C_a^T, \tag{3.7}$$

from which it follows that

$$\begin{cases} P_i^2 = P_i = P_i^T, \quad \mathrm{Tr}(P_i) = N_i, \\ P_a^2 = P_a = P_a^T, \quad \mathrm{Tr}(P_a) = N_a, \\ P_i^T P_a = 0. \end{cases} \tag{3.8}$$

We deduce the form of the set of discrete states in DM formalism

$$\boxed{\begin{aligned} \mathcal{M}_{\mathrm{DM}}(N_i, N_a; \mathbb{R}^{N_b}) := \Big\{ & (P_i, P_a) \in \mathbb{R}^{N_b \times N_b}_{\mathrm{sym}} \times \mathbb{R}^{N_b \times N_b}_{\mathrm{sym}} \text{ s.t. } P_i^2 = P_i, \\ & \mathrm{Tr}(P_i) = N_i \text{ for } i \in \{1,2\} \text{ and } P_i^T P_a = 0 \Big\}. \end{aligned}} \tag{3.9}$$

3.1.2 Adding gauge invariance

An important difference between the DM and MO formalisms (OMO and AMO) is that a trial state is represented by one and only one point of $\mathcal{M}_{\mathrm{DM}}(N_i, N_a; \mathbb{R}^{N_b})$, which is not the case in MO formalisms. Indeed one notices that for all $C = (C_i|C_a) \in \mathrm{St}(N_o; \mathbb{R}^{N_b})$, the MCSCF of KS-DFT approximate energies are invariant under any unitary transformations on orbitals of the same type i or a. For that reason, a trial state, discretized as $C \in \mathrm{St}(N_o; \mathbb{R}^{N_b})$, is represented by an infinity of points in $\mathrm{St}(N_o; \mathbb{R}^{N_b})$, namely the set

$$[C] = \left\{ CU := C \begin{pmatrix} U_i & 0 \\ 0 & U_a \end{pmatrix} = (C_i U_i | C_a U_a), \text{ where } U = (U_a|U_i) \in \mathcal{O}_{N_i} \times \mathcal{O}_{N_a} \right\} \subset \mathrm{St}(N_o; \mathbb{R}^N_b). \tag{3.10}$$

One way to recover the unicity of representation of trial states in MO formalism relies on the notion of quotient manifold. We introduce the equivalence relation on $\mathrm{St}(N_o; \mathbb{R}^{N_b})$ defined by

$$C \sim C' \Leftrightarrow \exists U \in \mathcal{O}_{N_i} \times \mathcal{O}_{N_a} \text{ such that } C = C'U \tag{3.11}$$

and we define the MO discrete state manifold as the quotient

$$\boxed{\mathcal{M}_{\mathrm{OMO}}(N_i, N_a; \mathbb{R}^{N_b}) := \mathrm{St}(N_o; \mathbb{R}^{N_b})/\sim = \mathrm{St}(N_o; \mathbb{R}^{N_b})/(\mathcal{O}_{N_i} \times \mathcal{O}_{N_a}).} \tag{3.12}$$

Then $\mathcal{M}_{\mathrm{OMO}}(N_i, N_a; \mathbb{R}^{N_b})$ is diffeomorphic to both $\mathcal{M}_{\mathrm{DM}}(N_i, N_a; \mathbb{R}^{N_b})$ and the approximate N-electron state space. Similarly, we can define the AMO manifold as the quotient

$$\boxed{\mathcal{M}_{\mathrm{AMO}}(N_i, N_a; N_b) \;:=\; \mathcal{O}_{N_b}/(\mathcal{O}_{N_i} \times \mathcal{O}_{N_a} \times \mathcal{O}_{N_e}).} \tag{3.13}$$

A summary of these three standard parametrizations is given in Table 1.

	Occupied MOs $C = (C_i \vert C_a) \in \mathbb{R}^{N_b \times N_o}$	Density Matrices $(P_i, P_a) \in \mathbb{R}^{N_b \times N_b}_{\mathrm{sym}} \times \mathbb{R}^{N_b \times N_b}_{\mathrm{sym}}$	All MOs $C = (C_i \vert C_a \vert C_e) \in \mathbb{R}^{N_b \times N_b}$
Constraints	$C^{\mathrm{T}} C = \mathrm{I}_{N_o}$	$P_j^2 = P_j$, $\mathrm{Tr}(P_j) = N_j$, $P_i P_a = 0$	$C^T C = I_{N_b}$
Manifold	$\mathrm{St}(N_o, \mathbb{R}^{N_b})$	$\mathcal{M}_{\mathrm{DM}}(N_i, N_a; \mathbb{R}^{N_b})$	$\mathcal{O}_{N_b}$
Extra variables	None	None	virtual MOs
Gauge invariance group	$\mathcal{O}(N_i) \times \mathcal{O}(N_a)$	$\{e\}$	$\mathcal{O}_{N_i} \times \mathcal{O}_{N_a} \times \mathcal{O}_{N_e}$

Table 1 – Three equivalent parametrizations of trial states for variational WFM and DFT ground state problems. All states are parametrized as a point of a smooth matrix manifold, defined by a set of differentiable constraints. The corresponding energy is invariant by the action of a gauge invariance group. The extra variables in AMO parametrization do not contribute to the energy, but impacts the geometry of the AMO manifold.

3.2 Optimization algorithms: direct minimization

Let us now discuss the two main classes of algorithms used to solve discrete GS problems starting with direct minimization. On the one hand, direct minimization algorithms are iterative methods that consist in following a series of steps in the search space, starting from an initial guess, so that the value of the energy functional decreases at each step. Let us recall the formulation of some famous direct minimization algorithms in the case of constrained minimization. They are usually formulated in cases where the search space is the linear space $\mathbb{R}^d$, endowed with the canonical scalar product $\langle \cdot | \cdot \rangle$. In that case, each iteration decomposes as a three-step process:

1. from an initial point $x_n \in \mathbb{R}^d$, choose a descent direction d_n along which the energy is locally decreasing. Of the numerous algorithms available, *gradient methods* are the most standard. For a given symmetric positive definite preconditioner P_n, they consist in following the direction d_n opposite to the gradient of the energy for the scalar product $\langle \cdot | P_n^{-1} \cdot \rangle$. This amounts to solve the quasi-newton equation
$$P_n d_n = -\nabla_{x_n} E \tag{3.14}$$
at each iteration, where the gradient of the energy on the right-hand side is computed for the canonical scalar product. In general, P_n is a symmetric positive definite approximation of the Hessian of the energy at current point $\mathrm{Hess}_{x_n} E$. The two extreme examples are the *steepest descent* (SD) algorithm ($P_n = \mathrm{Id}$) and the *Newton method* ($P_n = \mathrm{Hess}_{x_n} E$), which are respectively first and second order methods. While quadratically convergent when close to a minimum, the Newton method needs more computational resources and is unstable when starting far from a minimum. In that case, $\mathrm{Hess}_{x_n} E$ is not positive definite, and d_n might not be a descent direction. On the other hand, the steepest descent always produces descent directions, but it is usually very slow to converge.

Among intermediate approximations, let us mention the *non-linear conjugate gradient* (CG) and *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) schemes, which construct an approximation of the inverse Hessian of E by combining previous directions $\{d_{n-1}, d_{n-2}, \dots\}$ to the current direction computed with (3.14). These methods usually converge much faster than the steepest descent algorithm, while being much cheaper to implement and more stable than the Newton method (in practice, the low-memory version of the BFGS algorithm, or L-BFGS, is preferred). More details can be found in [NW99].

2. choose a step-size α_n and make a step in the direction d_n by setting $x_{n+1} = x_n + \alpha_n\, d_n$. This second phase of the direction minimization process, called *line-search*, amounts to solve the one-dimensional minimization problem
$$\alpha_n = \min_{\alpha \in \mathbb{R}^+} f_n(\alpha) := E(x_n + \alpha d_n). \tag{3.15}$$
Between the *fixed step* line-search, where the step size is the same for all iterations, and the *optimal step* method, which solves (3.15) exactly, most methods simply ask for a sufficient decrease of the function f_n. This requirement takes the form of a relaxed optimality condition, which sometimes involve the derivatives of f_n. Among the most commonly used methods, let us cite the Backtracking [NW99], the More-Thuente [MT94] or standard Hager-Zhang method [HZ05], originally designed for conjugate gradient.

3. given a tolerance $\varepsilon > 0$, test convergence by checking that the gradient of the energy is smaller in norm than ε (first order optimality condition).

Another class of direct optimization methods is the family of Riemannian optimization algorithms, that can be applied to minimization problems posed on a smooth manifold $\mathcal{M}$. These methods can be constructed from standard direct minimization algorithms in $\mathbb{R}^d$, by slightly modifying the above three steps as described in [**absil2009optimization**; EAS98]. These methods are the main subject of chapter 1, where they are discussed with more details.

3.3 Optimization algorithms: self-consistent field

In some cases, the discrete GS problem is equivalent to a generalized eigenvalue problem, wherein the matrix to be diagonalized depends on the ground state density matrix. It that case, the problem is typically solved using a fixed-point algorithm known in chemistry as a self-consistent field (SCF) method.

3.3.1 Standard SCF method

In that section, let us only consider a single group of orbital by assuming that $N_i = 0$ or $N_a = 0$. Let again N_o be the number of occupied orbitals. In that case, a trial state is parametrized in DM formalism by a single density matrix in the Grassmann manifold
$$\mathrm{Grass}(N_o; \mathbb{R}^{N_b}) := \underbrace{\left\{P \in \mathbb{R}^{N_b \times N_b}_{\mathrm{sym}}, \quad P^2 = P^T = P, \quad \mathrm{Tr}(P) = N_o\right\}}_{\text{DM}} = \underbrace{\mathcal{O}_{N_b}/(\mathcal{O}_{N_o} \times \mathcal{O}_{N_e})}_{\text{AMO}}. \tag{3.16}$$
On the Grassmann, it can be shown [CLBM06] that the first order optimality conditions read as the generalized eigenvalue problem (we still assume orthonormality of the basis set for simplicity)
$$\boxed{F(P_*)C_* \;=\; \left(C_*^T F(P_*) C_*\right) C_*.} \tag{3.17}$$
The *Fock matrix* F depends on the chosen approximation and C_* is the rectangular matrix in $\mathbb{R}^{N_b \times N_o}$ such that $P_* = C_* C_*^T$. This alternative form of the approximate ground state problem allows to introduce a new class of solvers, that can be formulated in terms of density matrices only.

First introduced by Roothaan [Roo60] in 1960 for the RHF model, it consists in assembling the Fock matrix for the current iterate $P_n \in \mathrm{Grass}(N_o; \mathbb{R}^{N_b})$, diagonalize it and select the N_o lowest energy eigenvectors C_{n+1} to form the next iterate $P_{n+1} = C_{n+1} C_{n+1}^T$. Note that this last step assumes that the *Aufbau* principle is verified. This procedure can be interpreted as a fix point method on the function
$$\boxed{g(P_n) \;:=\; P_{n+1}.} \tag{3.18}$$

At convergence, the mean-field potentials produced by the density matrices P_n and P_{n+1} become identical, which is why the method is commonly referred to as a *self-consistent field* procedure. While easy to implement (since it requires no knowledge about the geometry of $\mathcal{M}_{\mathrm{DM}}$) the simple SCF procedure (3.18) suffers from a lack of stability. The simple SCF map g has been shown to display chaotic behaviors [CKL21], and generally, the convergence of the SCF algorithms is highly dependent on the gap between occupied and virtual states, through the Jacobian of g. As a result the initial method have been refined to showcase faster and more stable convergence, yielding a variety of SCF algorithms, which we briefly present below. Mathematical studies of the convergence of standard Roothaan SCF can be found in [CLB00b; Liu+14; CKL21].

3.3.2 Stabilized / accelerated SCF

Among many corrections, it is common to add the following features to the standard SCF iteration $P_{n+1} = g(P_n)$:

1. a damping parameter $\alpha \in [0, 1]$:

$$P_{n+1} = P_n + \alpha(g(P_n) - P_n). \tag{3.19}$$

   In that case the density matrix produced by the diagonalization of the Fock matrix is linearly mixed with the previous iterate. We emphasize that the linear combination P_{n+1} has no reason to be on the discrete state manifold $\mathrm{Grass}(N_o; \mathbb{R}^{N_b})$. In order to retrieve an admissible density, it is customary to add a last simple SCF iteration ($\alpha = 1$) at the end of the procedure, since g takes its values in the manifold $\mathrm{Grass}(N_o; \mathbb{R}^{N_b})$. Note however that $g(P)$ can be very far from P in the cases where $\mathrm{Grass}(N_o; \mathbb{R}^{N_b})$ is highly curved. The convergence of the damped-SCF has been studied in [CKL21].

2. an acceleration method $\mathcal{A}$ which uses the previous iterates up to a certain *depth*

$$P_{n+1} = g(\mathcal{A}(P_n, \cdots, P_{n-\mathrm{depth}})). \tag{3.20}$$

   The most common methods are the Anderson-Pulay Acceleration (APA) techniques. This terminology, recently coined in [Chu+21], regroups various acceleration schemes into a general framework, including the *Anderson acceleration* [And65], introduced in the general context of integral equations, and the *Direct Inversion of the Iterative Subspace* (DIIS), introduced independently by Pulay [Pul80] in quantum chemistry. The APA schemes are based on linear combinations of the current iterate with the previous ones. The coefficients for the linear combination are solution of a low dimensional least square problem, which is solved at each iteration. Mathematical studies on the convergence of DIIS algorithms can be found in [RS11; CKL21; Chu+21].

3. a preconditioner (sometimes referred to as *mixing*)

$$P_{n+1} = M^{-1} g(P_n), \tag{3.21}$$

   where M approximates the Jacobian of the SCF function g. Some examples of preconditioning are given in [Ker81; HL20].

All three add-ons can be combined in the compact formulation

$$\boxed{P_{n+1} = P_n + \alpha M^{-1} \left(g(\mathcal{A}(P_n, \cdots, P_{n-\mathrm{depth}})) - P_n\right).} \tag{3.22}$$

3.4 Choosing a discretization basis

Let us now briefly discuss the choice of discretization basis $\mathcal{X}$ in quantum chemistry. We have seen that the standard WFM and DFT models construct an approximation of true ground-state wave-function or density built from a basis of molecular spin-orbitals. To produce a computationally tractable model, each spatial parts of the MOs is then expanded on a discretization basis $\mathcal{X} = (\chi_\mu)_{1 \leqslant \mu \leqslant N_b}$, which ultimately impacts the quality of the approximate ground-state energy and wave-function or density.

After discretization, the computation of approximate energies notably requires to evaluate matrix elements that depend on the so-called *kinetic* and *electron-repulsion* integrals

$$\int_{\mathbb{R}^3} \nabla\chi_\mu(r)\nabla\chi_\nu(r)\mathrm{d}r \quad \text{and} \quad \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\chi_\mu(r')\chi_\nu(r')\chi_\kappa(r)\chi_\lambda(r)}{|r-r'|}\mathrm{d}r\mathrm{d}r' \tag{3.23}$$

for all $(\mu,\nu,\kappa,\lambda) \in \{1,\dots,N_b\}^4$. This imposes two constraints on $\mathcal{X}$ (we refer to [HJO14] for a detailed discussion on the subject):

1. the integrals (3.23) should to be easy to evaluate;
2. the basis $\mathcal{X}$ should provide a relatively accurate approximation with a small number of basis functions, as the number of integrals (3.23) grows as N_b^4 with the number of basis function N_b (in practice refined strategies allow to reduce this complexity).

While finite elements or plane-wave methods (e.g. [Leh19c; Das+22]) have been applied to electronic structure problems for molecules, they are typically slow to converge with respect to the basis size N_b. For that reason they are seldom used for molecular systems, in comparison to the following approach.

3.4.1 Linear combination of atomic orbitals

The most common method, that offers a good compromise regarding the two constraints 1) and 2), is the *linear combination of atomic orbitals* (LCAO) method. Consider a molecule with M nuclei, with respective positions $R_1,\dots,R_M$ and charges $Z_1,\dots,Z_M$. In LCAO, one starts by constructing for each individual atom a basis of nucleus-centered fast-decaying functions

$$\left\{\chi_\mu^{Z_i}(\cdot - R_i) \in H^1(\mathbb{R}^3;\mathbb{C}),\ \ 1 \leqslant \mu \leqslant n^{Z_i}\right\}. \tag{3.24}$$

The $\chi_\mu^{Z_i}$ are usually chosen to mimic the atomic orbitals of the non-interacting-electron atom, as introduced in Section 1.3.3, which read in spherical coordinates

$$\chi_\mu^{Z_i}(r,\theta,\phi) = f_{nl\zeta}^{\mu}(r)Y_{lm}^{\mu}(\theta,\phi). \tag{3.25}$$

The radial part $f_{nl\zeta}^{\mu} : \mathbb{R}^3 \to \mathbb{R}$ depends on quantum parameters n and l, as well as a spread parameter $\zeta > 0$, and Y_{lm}^{μ} is a standard spherical harmonic. The parameters ζ, n, l and m usually depend on μ, but we have omitted that dependence for readability. The LCAO basis for a molecule is then obtained as the union of all bases for the individual atoms

$$\mathcal{X} = \left\{\chi_{\mu_n}^{Z_n}(\cdot - R_n),\ 1 \leqslant \mu_n \leqslant n_{Z_n},\ 1 \leqslant n \leqslant M\right\}. \tag{3.26}$$

For each atom, the number of atomic orbitals and the values of the parameters ζ, n, l and m are set to accurately describe the core electronic structure of the atom, as well as covalent bonds and electronic polarization in the molecule. As a result, only a small number of AOs per atom (typically a dozen) are necessary to obtain a relatively accurate approximation of most quantities of interest.

There exist essentially three flavors of LCAO methods that use AOs of the form (3.25), depending on the choice of the radial function $f_{nl\zeta}^{\mu}$. The first historically introduced variant of LCAO uses *Slater type orbitals* (STO), which are defined (up to a normalization constant) by

$$f_{nl\zeta}^{\mu,\mathrm{STO}}(r) = r^{n-1}e^{-\zeta r}. \tag{3.27}$$

The STOs conserve the exponential decay of the original AOs (1.27) as well as their characteristic cusp at nuclei positions. However, the integrals in (3.23) are difficult to compute for such functions, which motivated the introduction of *Gaussian-type orbitals* (GTOs), that use a Gaussian exponential part

$$f_{nl\zeta}^{\mu,\mathrm{GTO}}(r) = r^l e^{-\zeta r^2} \tag{3.28}$$

(where again we omitted the normalization constant). The primary advantage of GTOs lies in the fact that the integrals in (3.23) can be computed analytically. Nevertheless, unlike STOs, these functions provide a poor approximation of the real atomic orbitals, thereby limiting their practical utility. This limitation can be mitigated by employing *contracted* GTOs, as we discuss in the following section.

Last, *numerical atomic orbitals* (NAO) use numerical radial functions, typically tabulated on a fine grid and obtained as the minimizer (or approximate minimizer) of some optimality criterion for a given data set of atomic and molecular configurations. While STOs and GTOs have been historically preferred, recent developments in the generation of fully numerical AOs made NAOs an increasingly popular scheme in modern electronic structure codes [Lin+24; Leh24; Leh19a].

3.4.2 Contracted Gaussian-type bases

Contracted GTOs (CGTOs) are currently the most popular atomic basis sets for the LCAO method. Contracted GTOs radial functions are linear combinations of N_{ctr} individual GTOs

$$f_{nl}^{\mu,\text{CGTO}}(r) = r^l \left(\sum_{i=1}^{N_{\text{ctr}}} \alpha_i^\mu e^{-\zeta_i r^2} \right) \tag{3.29}$$

with contracting coefficients $\alpha_i^\mu \in \mathbb{R}$ and exponents $\zeta_i > 0$. In most cases, these functions address the deficiencies of GTOs while retaining their computational efficiency. As a result, they are commonly employed in molecular calculations, to the extent that the term GTOs typically denotes *contracted* GTOs in the chemistry literature. In turn, the gaussians used in the linear combination (3.29) are called *primitive* gaussians. Contracted GTOs fall into two categories based on the structure of the contraction coefficients matrix $M_{\text{ctr}} = [\alpha_i^\mu] \in \mathbb{R}^{N_{\text{ctr}} \times N_b}$: *segmented* contracted bases feature a block-diagonal matrix M_{ctr}, whereas *general* contracted bases use a full matrix M_{ctr}.

The variety of molecular configurations and properties that need to be approximated, along with the necessity of maintaining small basis sets, complicates the systematic development of *good* basis sets. A single AO basis set often lacks global accuracy, resulting in numerous basis sets tailored to specific calculations or systems. This can be seen for instance in the *Basis Set Exchange* (BSE) database, which catalogs 689 distinct types of atomic basis sets (each applicable to several atoms) at the time of writing of this manuscript. Recent reviews discussing the optimization of AO basis sets include the following references [Jen13; Leh19a; Per21].

4 Electronic structure of crystalline materials

Let us now return to the topic of crystalline materials. Given their virtually infinite number of electrons, modeling the electronic structure of solids requires introducing a specific mathematical framework.

4.1 The ground state problem for crystals

4.1.1 Mathematical description of a perfect infinite crystal

The mathematical description of crystalline materials is made easier by two physical hypotheses: first, the crystalline material is modeled as a perfect infinite crystal in dimension $d = 1, 2, 3$. This physically amounts to neglect the boundary effects and the presence of eventual defects in the solid. Conveniently, this structure is invariant by translation of the periodic lattice $\mathcal{R} = \sum_{i=1}^d \mathbf{a}_i \mathbb{Z}$, where $(\mathbf{a}_i)_{1 \leqslant i \leqslant d}$ are vectors of $\mathbb{R}^d$. Given a set of M atomic positions $(\mathbf{R}_1, \cdots, \mathbf{R}_M)$ in the unit cell $\Omega = \sum_{i=1}^d \mathbf{a}_i [0, 1)$, the full atomic configuration of the crystal is the set

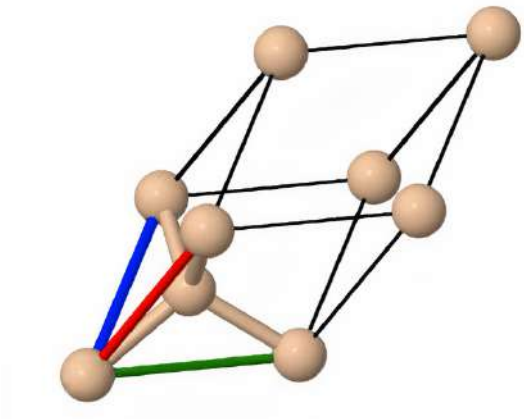
$$\big\{ \mathbf{R}_i + \mathbf{R}, \;\; 1 \leqslant i \leqslant M, \;\; \mathbf{R} \in \mathcal{R} \big\}. \tag{4.1}$$

Second, the electron cloud can be described by a set of non-interacting quasi-particles, evolving under the influence of an effective (mean-field) potential V_{per}, which is $\mathcal{R}$-periodic by the above approximation. Since the electrons do not interact, all the electronic properties of the crystal can be recovered by studying the one-electron Hamiltonian

$$\boxed{\hat{\mathrm{H}} \;=\; -\frac{1}{2}\Delta + V_{\text{per}} \quad \text{on} \quad L^2(\mathbb{R}^d \times \{\uparrow, \downarrow\}; \mathbb{C}).} \tag{4.2}$$

Under suitable regularity assumptions on V_{per}, depending on the dimension d, the Hamiltonian $\hat{\mathrm{H}}$ is proven to be bounded-bellow and self-adjoint [Lew22, Theorem 7.1]. However, it can be shown that

(a) A piece of silicon, a widely used semi-conductor, studied in Part II. *Source: Wikipedia Commons.*



(b) Unit cell Ω of face-centered cubic (FCC) silicon. The vectors $\mathbf{a}_i$ are pictured in blue, red and green. The two atoms in the cell are reported on the edge of the cell by periodicity. *Produced with the SeeK-path online tool [Hin+17].*

Figure 5 – Representation of silicon as a three dimensional perfect infinite crystal in face-centered cubic (FCC) configuration.

$\hat{\mathrm{H}}$ has no eigenvalues, which makes the direct computation of its spectrum difficult. This problem is usually handled with Bloch theory, that allows to compute the spectral properties of periodic operators. We outline bellow the essential aspects of the Bloch transform needed in this manuscript. A thorough introduction to Bloch theory and the proofs of these results can be found in [RS78, Chapter XIII.16].

4.1.2 Bloch theory

For the sake of simplicity, we will neglect here the electronic spin. From a mathematical point of view, Bloch theory consists in the block-decomposition of $\mathcal{R}$-periodic operators, which are operators $\hat{\mathrm{A}} \in \mathcal{L}(L^2(\mathbb{R}^d;\mathbb{C}))$ verifying

$$[\hat{\mathrm{A}}, \tau_{\mathbf{R}}] = 0 \tag{4.3}$$

for all translation operators of the lattice $\mathcal{R}$

$$\tau_{\mathbf{R}} : f \in L^2(\mathbb{R}^d;\mathbb{C}) \mapsto f(\cdot - \mathbf{R}), \quad \text{with} \quad \mathbf{R} \in \mathcal{R}. \tag{4.4}$$

Informally, the commuting relations (4.3) imply that $\hat{\mathrm{A}}$ can be block-diagonalized in a basis of pseudo-eigenvectors of the translation operators $(\tau_{\mathbf{R}})_{\mathbf{R}\in\mathcal{R}}$, as it would happen for a family of commuting matrices in finite dimensions. Let

$$\mathcal{R}^* = \sum_{i=1}^{d} \mathbf{a}_i^* \mathbb{Z}, \qquad \Omega^* = \sum_{i=1}^{d} \mathbf{a}_i^* [0,1) \tag{4.5}$$

be the reciprocal lattice of $\mathcal{R}$ and the corresponding reciprocal unit cell, where the reciprocal basis vectors are defined by the relation $\langle \mathbf{a}_i^* | \mathbf{a}_j^* \rangle = 2\pi\delta_{ij}$, $1 \leqslant i,j \leqslant d$. Let us also denote the space of $\mathcal{R}$-periodic functions

$$L^2_{\mathrm{per}}(\Omega;\mathbb{C}) = \left\{ u \in L^2_{\mathrm{loc}}(\mathbb{R}^d;\mathbb{C}),\ \text{such that } \tau_{\mathbf{R}} u = u,\ \forall \mathbf{R} \in \mathcal{R} \right\}. \tag{4.6}$$

The pseudo-eigenvectors of the translations of the lattice $\mathcal{R}$ are given by the functions

$$\psi_{\mathbf{k}} : r \in \mathbb{R}^d \mapsto u_{\mathbf{k}}(r) e^{i\mathbf{k}\cdot r}, \quad u_{\mathbf{k}} \in L^2_{\mathrm{per}}(\Omega;\mathbb{C}), \quad \mathbf{k} \in \Omega^* \tag{4.7}$$

usually called *Bloch waves* in solid state physics. Starting from the standard inverse Fourier transform $\mathcal{F}^{-1}$, we find that all functions of $u \in L^2(\mathbb{R}^d;\mathbb{C})$ can be decomposed as an integral of Bloch waves

$$u(x) = \int_{\mathbb{R}^d} \mathcal{F}(u)(q) e^{iq\cdot x} \mathrm{d}q = \int_{\Omega^*} \left(\sum_{\mathbf{G}\in\mathcal{R}^*} \mathcal{F}(u)(\mathbf{k}+\mathbf{G}) e^{i\mathbf{G}\cdot r} \right) e^{i\mathbf{k}\cdot r} \mathrm{d}\mathbf{k} := \int_{\Omega^*} u_{\mathbf{k}}(x) e^{\mathbf{k}\cdot x} \mathrm{d}\mathbf{k} \tag{4.8}$$

where

$$u_{\mathbf{k}}(x) = \sum_{\mathbf{G}\in\mathcal{R}^*} \mathcal{F}(u)(\mathbf{k}+\mathbf{G}) e^{i\mathbf{G}\cdot x}. \tag{4.9}$$

From the above informal reasoning, the action of $\hat{\mathrm{A}}$ on a function $u \in L^2(\mathbb{R}^d;\mathbb{C})$ should be recovered by its action on each individual Bloch wave $u_{\mathbf{k}}$ defined by (4.9), which is the case in practice. We refer to [Lew22, Chapter 7] or to [Lev20] for broader intuitive, pedagogical introductions to Bloch theory.

Formally, we introduce the Bloch transform as the isometry $\mathcal{B} : L^2(\mathbb{R}^d;\mathbb{C}) \longrightarrow L^2(\Omega_*, L^2_{\mathrm{per}}(\Omega))$ defined for all $u \in L^2(\mathbb{R}^d;\mathbb{C})$, $\mathbf{k} \in \Omega^*$ and $x \in \mathbb{R}^d$ by

$$\boxed{\begin{aligned} \mathcal{B}(u)(\mathbf{k},x) &:= u_{\mathbf{k}}(x) = \sum_{\mathbf{G}\in\mathcal{R}^*} \mathcal{F}(u)(\mathbf{k}+\mathbf{G}) e^{i\mathbf{G}\cdot x} \\ \left(\mathcal{B}^{-1} u_\bullet\right)(x) &= \fint_{\Omega^*} u_{\mathbf{k}}(x) e^{i\mathbf{k}\cdot x} \mathrm{d}\mathbf{k}. \end{aligned}} \tag{4.10}$$

Then a $\mathcal{R}$-periodic operator $\hat{\mathrm{A}}$ is decomposed by the Bloch transform [RS78], which means that there exists an operator-valued function $\hat{\mathrm{A}}_\bullet \in L^\infty(\Omega^*, \mathcal{L}(L^2_{\mathrm{per}}(\Omega)))$ such that

$$\mathcal{B}(\hat{\mathrm{A}}u)(\mathbf{k}) = \hat{\mathrm{A}}_{\mathbf{k}} u_{\mathbf{k}}. \tag{4.11}$$

The operators $(\hat{\mathrm{A}}_{\mathbf{k}})$ are called the Bloch fibers of $\hat{\mathrm{A}}$. Additionally, the spectrum of $\hat{\mathrm{A}}$ is recovered with $\sigma(\hat{\mathrm{A}}) = \overline{\bigcup_{\mathbf{k}\in\Omega^*} \sigma(\hat{\mathrm{A}}_{\mathbf{k}})}$. Using the direct integral notations (see [RS78, Section XIII.16]), the decomposition (4.11) reads

$$\boxed{L^2(\mathbb{R}^3;\mathbb{C}) \simeq \fint_{\Omega^*}^{\oplus} L^2_{\mathrm{per}}(\Omega)\mathrm{d}\mathbf{k} \quad \text{and} \quad \mathcal{B}\hat{\mathrm{A}}\mathcal{B}^{-1} = \fint_{\Omega^*}^{\oplus} \hat{\mathrm{A}}_{\mathbf{k}}\mathrm{d}\mathbf{k}.} \tag{4.12}$$

4.1.3 Band diagrams

The decomposition (4.12) can be applied to the one-body electronic Hamiltonian $\hat{\mathrm{H}}$ as defined in (4.2). In that case, the fibers $\hat{\mathrm{H}}_{\mathbf{k}} = (-i\nabla + \mathbf{k})^2 + V_{\mathrm{per}}$ are *unbounded* self-adjoint operators acting on

$$L^2_{\mathrm{per}}(\Omega \times \{\uparrow,\downarrow\};\mathbb{C}) = \left\{ \psi \in L^2_{\mathrm{loc}}(\mathbb{R}^d \times \{\uparrow,\downarrow\};\mathbb{C}), \quad \text{s.t. } \tau_{\mathbf{R}}(\psi) = \psi, \quad \forall \mathbf{R} \in \mathcal{R} \right\} \tag{4.13}$$

with domain $H^2_{\mathrm{per}}(\Omega \times \{\uparrow,\downarrow\};\mathbb{C})$. We can write the following

Theorem 4.1 (Spectrum of $\hat{\mathrm{H}}$)**.** [Lew22, Chapter 7], [RS78, Chapter XIII.16] .

Consider the one-electron Hamiltonian $\hat{\mathrm{H}}$ with Bloch fibers $\hat{\mathrm{H}}_{\mathbf{k}}$, for all $\mathbf{k} \in \Omega^*$. Suppose that the potential V_{per} is a periodic potential for which $\hat{\mathrm{H}}$ is self-adjoint with domain $H^2(\mathbb{R}^d \times \{\uparrow,\downarrow\};\mathbb{C})$. Then:

1. each $\hat{\mathrm{H}}_{\mathbf{k}}$ is self-ajoint with domain $H^2_{\mathrm{per}}(\Omega \times \{\uparrow,\downarrow\};\mathbb{C})$, bounded-bellow and has compact resolvent. Therefore each $\hat{\mathrm{H}}_{\mathbf{k}}$ has a purely discrete spectrum with eigenvalues accumulating at $+\infty$ and eigenfunctions that form an orthonormal basis of $L^2_{\mathrm{per}}(\Omega \times \{\uparrow,\downarrow\})$;

2. for each $\mathbf{k}$, let $\varepsilon_{1,\mathbf{k}} \leqslant \varepsilon_{2,\mathbf{k}} \leqslant \cdots$ be the eigenvalues of $\hat{\mathrm{H}}_{\mathbf{k}}$ in increasing order. Then for all $n \in \mathbb{N}^*$, the mapping $\mathbf{k} \mapsto \varepsilon_{n,\mathbf{k}}$ is Lipschitz continuous and $\mathcal{R}^*$-periodic. It is analytic away from band crossings, which are wave-vector $\mathbf{k}_0$ such that

$$\exists\, n, m \in \mathbb{N}^* \quad \text{s.t} \quad \varepsilon_{n,\mathbf{k}_0} = \varepsilon_{m,\mathbf{k}_0}.$$

3. the operator $\hat{\mathrm{H}}$ has purely continuous spectrum and

$$\sigma(\hat{\mathrm{H}}) = \bigcup_{n\in\mathbb{N}} \left[\min_{\mathbf{k}\in\overline{\Omega^*}} \varepsilon_{n,\mathbf{k}},\ \max_{\mathbf{k}\in\overline{\Omega^*}} \varepsilon_{n,\mathbf{k}} \right].$$

From the above theorem, the spectrum of the one-electron Hamiltonian $\hat{\mathrm{H}}$ can be computed by solving the family of eigenvalue problems for each $\mathbf{k}$-fiber

$$\hat{\mathrm{H}}_{\mathbf{k}} u_{n,\mathbf{k}} = \varepsilon_{n,\mathbf{k}} u_{n,\mathbf{k}}, \qquad \langle u_{n\mathbf{k}} | u_{m\mathbf{k}} \rangle_{L^2_{\mathrm{per}}(\Omega;\mathbb{C})} = \delta_{nm}, \quad \forall n, m \in \mathbb{N}. \tag{4.14}$$

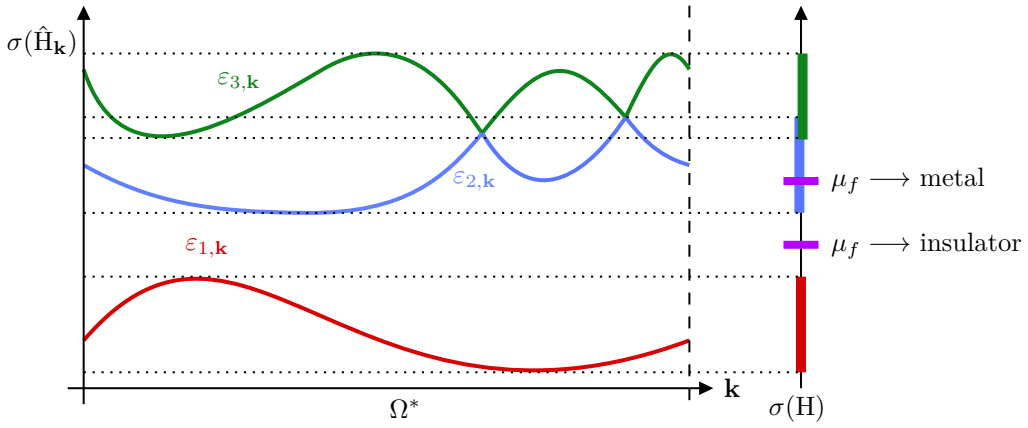


Figure 6 – Illustration of Theorem 4.1. The spectrum of $\hat{\mathrm{H}}$ is composed of a set of bands, that are the union of the individual eigenvalues of the Bloch fibers $\hat{\mathrm{H}}_{\mathbf{k}}$. The Fermi level μ_f, which depends on the number of electrons per unit cell, allows to differentiate materials. If μ_f lies in a gap between bands of $\sigma(\hat{\mathrm{H}})$, the material is called an insulator or semiconductor. If μ_f belongs to a band, it is called a metal.

A representation of $\sigma(\hat{\mathrm{H}})$ via the map $\mathbf{k} \mapsto (\varepsilon_{n,\mathbf{k}})_{n\in\mathbb{N}}$, called a *band diagram*, is pictured in Figure 6. From $\sigma(\hat{\mathrm{H}})$, we can retrieve information on the full crystal. As for the non-interacting-electron atom, the independent electrons within the crystal occupy the lowest energy levels of $\hat{\mathrm{H}}$, "filling" them until the number of electrons per unit cell reaches the wanted value. In the crystalline case with an infinite number of electrons, the sum over the energies translates, through a thermodynamic limit [LBL05], as the integral over the reciprocal unit cell Ω^*. We introduce the integrated density of states per unit cell:

$$\forall \mu \in \mathbb{R}, \quad \mathcal{N}(\mu) = \sum_{n\in\mathbb{N}} \fint_{\Omega^*} \mathbb{1}(\varepsilon_{n,\mathbf{k}} \leqslant \mu) \mathrm{d}\mathbf{k}. \tag{4.15}$$

and the integrated density of energy per unit cell

$$\forall \mu \in \mathbb{R}, \quad \mathcal{E}(\mu) = \sum_{n\in\mathbb{N}} \fint_{\Omega^*} \varepsilon_{n,\mathbf{k}} \mathbb{1}(\varepsilon_{n,\mathbf{k}} \leqslant \mu) \mathrm{d}\mathbf{k}. \tag{4.16}$$

Let N_Ω be the number of electrons per unit cell. The crystal Fermi level μ_f is given by the equation

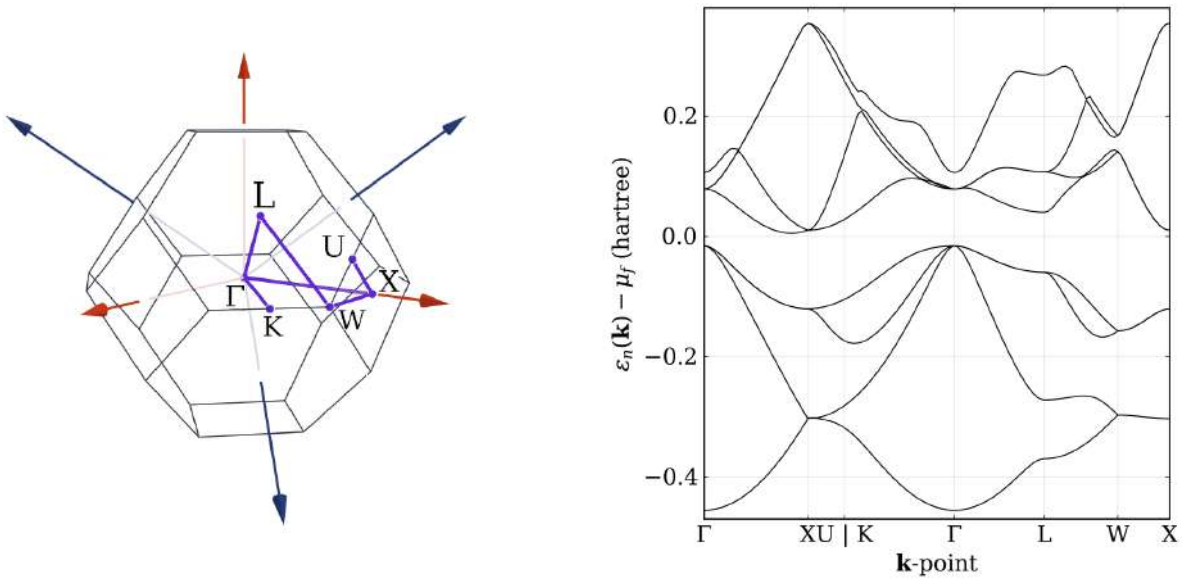
$$\mathcal{N}(\mu_f) = \frac{N_\Omega}{2} \tag{4.17}$$

where the factor 2 comes from the degeneracy of energies related to spin. It is the equivalent for solids of the chemical potential introduced for molecular systems. From μ_f, the ground state energy per unit cell reads

$$\mathcal{E}_* = \mathcal{E}(\mu_f). \tag{4.18}$$

The Fermi level of crystalline solids does not necessarily fall between occupied and virtual energies, which provides a qualitative mean to distinguish materials. As illustrated in Figure 6, if the Fermi level falls within a gap between two bands in the spectrum of $\hat{\mathrm{H}}$, the material is categorized as an insulator (or a semiconductor if the gap width is small). As in the atomic case, it is often placed in the middle of the gap by convention. Conversely, if it lies within one of the bands in $\sigma(\hat{\mathrm{H}})$, the material is classified as a metal.

In most cases, where $d = 3$, the full band structure of the system, which is a four-dimensional surface, cannot be shown. A representation such as that shown in Figure 6 is obtained by following a particular path $t \in [0, 1) \mapsto \mathbf{k}(t)$ in the reciprocal unit cell Ω^*. Since the Bloch decomposition is independent of the reciprocal unit cell, the standard approach is to take $\Omega^* = B$, the first Wigner-Seitz cell of the reciprocal lattice, known in physics as the (first) *Brillouin* zone. This cell inherits many symmetry properties from the crystal lattice. In the Brillouin zone, the paths $t \in [0, 1] \mapsto \mathbf{k}(t) \in \mathcal{B}$ are usually built as piece-wise linear paths joining *high-symmetry* $\mathbf{k}$-points, that are typically the corners, edges, and faces of the Brillouin zone. As an example, a given high-symmetry $\mathbf{k}$-path and associated band diagram of face-centered cubic silicon are pictured in Figure 7.



(a) A given high-symmetry $\mathbf{k}$-path in the Brillouin zone of face-centered cubic silicon.

(b) The band structure along this path.

Figure 7 – One high-symmetry $\mathbf{k}$-path of face-centered cubic crystalline silicon (left) and the corresponding band structure (right). Red and blue arrows display the Cartesian coordinate axes and the reciprocal basis vectors respectively. All bands are shifted so that the Fermi level appears at zero Hartree on the graph. *Source: produced with* `DFTK`

4.2 Solving the ground state problem for crystalline materials

From the above theorem, the electronic quantities of interest for crystalline materials can be expressed in terms of integrals (over the reciprocal unit cell) or derivatives involving a set of energy bands $(n, \mathbf{k}) \mapsto \varepsilon_{n,\mathbf{k}}$. The numerical estimation of important quantities, such as the energy per unit-cell (4.16) is therefore a two-step discretization process:

1. first select a finite sampling of Ω^* and a suitable numerical quadrature method to approximate the integral over Ω^*;

2. then discretized and solve approximately the $\mathbf{k}$-fiber eigenvalue problem, for all $\mathbf{k}$ in the sampling.

Concerning the first step, the famous Monkhorst-Pack numerical scheme [MP76] is widely used to

select the specific $\mathbf{k}$-points at which the eigenvalue problem is to be solved. It consists in a regular grid

$$\Omega^*_{\mathrm{MP}}(n_1, \cdots, n_d) = \left\{ \mathbf{k} = \sum_{i=1}^{d} \frac{c_i}{n_i} \mathbf{a}_i^*, \quad c_i \in \{0, \cdots, n_i - 1\} \right\}, \tag{4.19}$$

which benefits from exponential convergence rates in the number of $\mathbf{k}$-points, at least in the case of insulators [GL16]. Additionally, a number of numerical quadrature methods for integration have been proposed including the well-known linear tetrahedron method (see, e.g., [LT72]) and the improvement due to Blöchl et al. [BJA94], and smearing methods (see, e.g., [Mor+18; PP99; Hen01; MP89]). We refer to [Can+20] for a mathematical study of Brillouin-zone integration methods.

When it comes to the second step, two classes of methods prevail, which we briefly describe bellow.

4.2.1 Plane-wave discretization methods

Since the eigenvalue problems (4.14) are posed on a periodic domain, Fourier discretization method are a first natural choice. The $\mathbf{k}$-fiber states $u_{n,\mathbf{k}}$ are indeed $\mathcal{R}$-periodic, and can be expanded in the plane-wave (PW) basis

$$\mathcal{X}^\tau = \left\{ e^\tau_{\mathbf{G}}(r, \sigma) = \frac{1}{\sqrt{|\Omega|}} e^{i\mathbf{G}\cdot r} \tau(\sigma), \quad \mathbf{G} \in \mathcal{R}^* \right\}, \quad \tau \in \{\alpha, \beta\}. \tag{4.20}$$

Note that for the cases studied in this manuscript, we can neglect the electronic spin and consider the plane-waves $e_{\mathbf{G}} := e^\alpha_{\mathbf{G}} + e^\beta_{\mathbf{G}}$, $\forall\, \mathbf{G} \in \mathcal{R}^*$. The Fourier coefficients of $u_{n,\mathbf{k}}[\mathbf{G}]$ go to zero with $|\mathbf{G}| \to \infty$ so that the basis (4.20) is usually truncated using a scalar cut-off $\mathrm{E_c} > 0$

$$\mathcal{X}_0^{\mathrm{E_c}} = \left\{ e_{\mathbf{G}} \;\middle|\; \mathbf{G} \in \mathcal{R}^*, \quad \frac{1}{2}|\mathbf{G}|^2 < \mathrm{E_c} \right\}. \tag{4.21}$$

Two discretization strategies now arise:

- **Uniform Galerkin discretization:** the first strategy uses the same discretization basis $\mathcal{X}_0^{\mathrm{Ec}}$ for all eigenvalue problems. Let $\mathrm{X}_0^{\mathrm{E_c}} = \operatorname{Span} \mathcal{X}_0^{\mathrm{E_c}}$ and $\Pi_0^{\mathrm{E_c}}$ be the orthogonal projector on $\mathrm{X}_0^{\mathrm{E_c}}$ for the L^2_{per} canonical scalar product. Given a fiber $\hat{\mathrm{H}}_{\mathbf{k}}$, one solves the eigenvalue problem

$$\Pi_0^{\mathrm{E_c}} \hat{\mathrm{H}}_{\mathbf{k}} \Pi_0^{\mathrm{E_c}} u^{\mathrm{E_c}}_{n,\mathbf{k}} = \varepsilon^{\mathrm{E_c}}_{n,\mathbf{k}} u^{\mathrm{E_c}}_{n,\mathbf{k}} \tag{4.22}$$

  where $\langle u_{n,\mathbf{k}} | u_{m,\mathbf{k}} \rangle_{L^2_{\mathrm{per}}(\Omega;\mathbb{C})} = \delta_{nm}$ for all $n, m \in \{1, \cdots, \operatorname{rank} \Pi_0^{\mathrm{E_c}}\}$.

- **k-dependent Galerkin discretization:** in the second strategy, the discretization basis depends on the wave-vector $\mathbf{k}$

$$\mathcal{X}_{\mathbf{k}}^{\mathrm{E_c}} = \left\{ e_{\mathbf{G}} \;\middle|\; \mathbf{G} \in \mathcal{R}^*, \quad \frac{1}{2}|\mathbf{G} + \mathbf{k}|^2 < \mathrm{E_c} \right\}. \tag{4.23}$$

  The discrete problem is essentially the same as above, by replacing $\Pi^{\mathrm{E_c}}$ by $\Pi_{\mathbf{k}}^{\mathrm{E_c}}$, the orthogonal projector on $\mathrm{X}_{\mathbf{k}}^{\mathrm{E_c}} = \operatorname{Span} \mathcal{X}_{\mathbf{k}}^{\mathrm{E_c}}$.

These two strategies are pictured in Figure 8. While plane-waves are a natural choice of discretization basis for periodic systems, they typically require a very large number of basis functions to accurately describe the sharp variations of the wave-functions or density, such as cusp observed at the nuclear positions [Kat57]. In addition, the orthogonality constraints imposed on the orbitals $u_{n,\mathbf{k}}$ imply that the orbitals related the valence electrons have to quickly oscillate near atomic sites.

These two well-known caveats of plane-wave methods are commonly dealt with by using pseudo-potentials, that approximate the Coulomb interaction generated by the core electrons. At present day, plane-wave methods using norm-conserving pseudo-potentials [HSC79] or the projector augmented wave method [Blö94] are a very convenient and popular choice for the electronic structure of crystalline materials. We will not give further information on pseudo-potential, that are used as a mere tool in this manuscript, and refer to [Dup18] for a clear introduction to the subject.

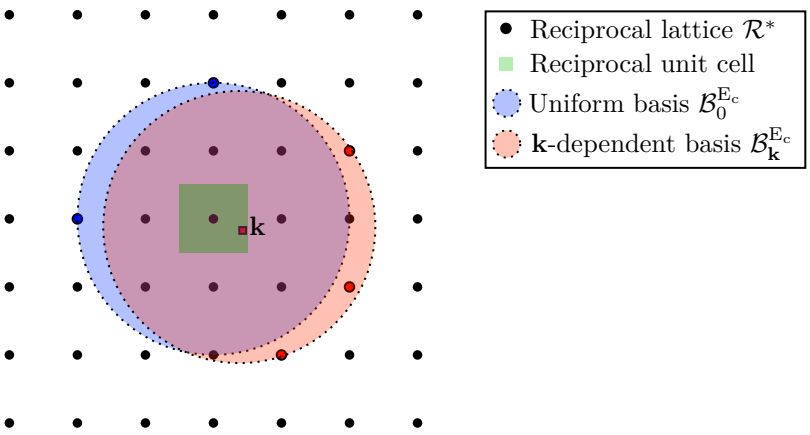


Figure 8 – An example of the uniform and $\mathbf{k}$-dependent basis sets. The reciprocal lattice $\mathcal{R}^*$ is a square lattice indicated with black dots, with a reciprocal unit-cell shaded in light green. The blue disk contains all $\mathbf{G} \in \mathcal{R}^*$ which belong to the uniform basis set, while the red disk contains all vectors in the $\mathbf{k}$-dependent basis. Notice that both bases contain additional $\mathbf{G}$ vectors that are missing from the other basis. These points are respectively pictured in blue and red.

4.2.2 Tight-Binding approximation

An alternative approach to solving the ground state problem for crystalline materials consists in first introducing a discretization basis $\mathcal{X}$ in real space before applying the Bloch decomposition, or its equivalent in a discrete setting. Such methods are referred to as *tight-binding* (TB) approximations. They are the analogue for periodic systems to the LCAO method, introduced for isolated molecular systems in the previous Section 3.4. In TB, one introduces a set of N_b fast-decaying nuclei-centered functions $\chi_1, \dots, \chi_{N_b} \in H^1(\mathbb{R}^d; \mathbb{C})$ in the unit cell of the crystal. The full discretization basis is then obtained by translation

$$\mathcal{X}^{\mathrm{TB}} = \{\chi_i(\cdot - \mathbf{R}), \quad \mathbf{R} \in \mathcal{R}, \quad 1 \leqslant i \leqslant N_b\} \tag{4.24}$$

and a tight-binding state read as an infinite sum

$$\psi^{\mathrm{TB}}(x) = \sum_{i=1}^{N_b} \sum_{\mathbf{R}\in\mathcal{R}} [C(\mathbf{R})]_i \chi_i(x - \mathbf{R}), \quad C(\mathbf{R}) \in \mathbb{C}^{N_b}. \tag{4.25}$$

Solving the TB approximate problem require to compute the matrix elements

$$[\mathbf{H}_{\mathbf{R}}]_{jj'} = \langle \chi_i | \hat{\mathrm{H}} | \tau_{\mathbf{R}} \chi_j \rangle, \quad \text{and} \quad [\mathbf{S}_{\mathbf{R}}]_{jj'} = \langle \chi_i | \tau_{\mathbf{R}} \chi_j \rangle \tag{4.26}$$

for all $1 \leqslant j, j' \leqslant N_b$ and $\mathbf{R} \in \mathcal{R}$, which is one of the subject of chapter 5.

4.2.3 Wannier functions in crystalline materials

While the Bloch decomposition allows to compute the accessible energies of the electrons of a crystal, the wave-functions associated with the Bloch fibers, the Bloch waves, are periodic of the lattice and delocalized over the whole crystal. Hence they do not provide an intuitive representation of the electronic localization or bonds typically observed in the unit cell. Wannier functions, on the other hand, are localized wave-functions that offer an alternative representation of the electronic structure of crystalline materials, derived by a unitary transform of a set of Bloch waves. They are the analogue for crystals of the atomic orbitals basis sets, discussed in Section 3.4, and are therefore closely related to tight-binding methods.

Let again $(\hat{\mathrm{H}}_{\mathbf{k}})_{\mathbf{k}\in\Omega^*}$ be the Bloch fibers of the one-electron crystalline Hamiltonian with associated energy bands $(\varepsilon_{n,\mathbf{k}})_{n\in\mathbb{N}^*,\mathbf{k}\in\Omega^*}$. To construct Wannier functions we start by defining an energy window of interest $\mathscr{E} = (\varepsilon^-, \varepsilon^+) \subset \mathbb{R}$. For the sake of simplicity, we suppose that $\mathscr{E}$ contains a set $\mathcal{I} = (i_1, \dots, i_J) \subset \mathbb{N}^J$ of J isolated bands in the sense that

1. $\forall n \in \mathcal{I} \quad \mathrm{Ran}(\varepsilon_{n,\bullet}) \subset \mathscr{E}$ (the n-th band is entirely contained in the energy window);

2. $\inf_{\mathbf{k}\in\Omega^*, n\in J, m\notin\mathcal{J}} |\varepsilon_{n,\mathbf{k}} - \varepsilon_{m,\mathbf{k}}| > 0$ (bands in the energy window are isolated from other bands).

An example of such a window is given by $\mathscr{E} = (-\infty, \mu_{\mathrm{f}})$, where μ_f is the Fermi level of an insulator. The construction of Wannier functions generalizes to the case of non-isolated bands (or entangled bands) as described in [Mar+12; DLL19]. For all $k \in \Omega^*$, let $P_{\mathscr{E}}(\mathbf{k})$ be the spectral projector of $\hat{\mathrm{H}}_{\mathbf{k}}$ associated to $\sigma(\hat{\mathrm{H}}_{\mathbf{k}}) \cap \mathscr{E}$. A family of Wannier functions are obtained from a basis $\{u_{n,\mathbf{k}}\}_{n\in\mathcal{I}}$ of $\mathrm{Ran}(P_{\mathscr{E}}(\mathbf{k}))$ with

$$\boxed{\forall \mathbf{R} \in \mathcal{R} \quad w_{n,\mathbf{R}}(x) = \tau_{\mathbf{R}} \left(\mathcal{B}^{-1}(u_{n,\bullet})\right)(x) = \fint_{\Omega^*} u_{n,\mathbf{k}}(x) e^{i\mathbf{k}\cdot(x-\mathbf{R})} \mathrm{d}\mathbf{k}} \tag{4.27}$$

with $\mathcal{B}^{-1}$ the inverse Bloch transform (4.10). By orthonormality of the Bloch waves, and since $\mathcal{B}$ is an isometry, one obtains that the Wannier functions $\{w_{n,\mathbf{R}},\, n \in \mathcal{I},\, \mathbf{R} \in \mathcal{R}\}$ are orthonormal and span the same Hilbert space as $\{u_{n,\mathbf{k}},\, n \in \mathcal{I},\, \mathbf{k} \in \Omega^*\}$. It can be seen from (4.27) that Wannier functions are not uniquely determined with respect to the energy window $\mathscr{E}$ and depend on the choice of basis of $P_{\mathscr{E}}(\mathbf{k})$. We can make that dependence appear by considering for all family of unitary matrices $\{U(\mathbf{k}),\, \mathbf{k} \in \Omega^*\}$, usually called a *gauge*, the action

$$\forall \mathbf{k} \in \Omega^* \quad u_{n,\mathbf{k}} \cdot U(\mathbf{k}) = \sum_{m\in\mathcal{I}} u_{n,\mathbf{k}} U_{m,n}(\mathbf{k}) \tag{4.28}$$

and denote $w_{n,R}(U(\mathbf{k}))$ the Wannier functions constructed via (4.27) with the basis $\{u_{n,\mathbf{k}} \cdot U(\mathbf{k})\}$.

By standard Fourier duality, the localization in space of a Wannier functions $w_{n,\mathbf{R}}$ is related to the regularity of $\mathbf{k} \mapsto u_{n,\mathbf{k}}$. For isolated bands, the map $\mathbf{k} \mapsto P_{\mathscr{E}}(\mathbf{k})$ is analytic [PP13] and one can build a map $(u_{n,\bullet})_*$ (e.g. the projection $P_{\mathscr{E}}(\mathbf{k})u$ of any guess function u) that is analytic. The corresponding Wannier functions decrease exponentially. In the general case however, the construction of exponentially localized Wannier functions might be hindered by topological obstructions [Pan07].

The spectral and localization properties of exponentially localized Wannier functions makes them a central tool of condensed matter physics, used in various applications such as energy bands interpolation, tight-binding parametrizations, and low-scaling methods. This makes the numerical computation of exponentially decreasing Wannier function an important topic.

The most standard algorithm to compute localized Wannier functions has been introduced by Marzari and Vanderbilt in 1997 [MV97] and consists in minimizing the functional

$$\Omega^{\mathrm{MV}}(\{U(\mathbf{k})\}) = \sum_{n\in\mathcal{I}} \left(\int_{\mathbb{R}^3} |x|^2 |w_{n,\mathbf{0}}(U(\mathbf{k}))|^2(x) \mathrm{d}x - \left| \int_{\mathbb{R}^3} x |w_{n,\mathbf{0}}(U(\mathbf{k}))|^2 \mathrm{d}x \right|^2 \right) \tag{4.29}$$

measuring the spread of Wannier functions in the unit cell for the gauge $\{U(\mathbf{k})\}$. Wannier functions obtained by the Marzari-Vanderbilt (MV) procedure are usually called *maximally localized Wannier functions* (MLWFs). The MV algorithm is notably implemented in the software `Wannier90` [Piz+20], which is currently one of the most commonly used program for wannierization. For isolated bands, MLWFs are known to be exponentially decreasing [Bro+07]. For entangled bands however, the MV algorithm expresses as a two step minimization problem (with a pre-computation step called *disentanglement*) which can fail to converge and showcases several local minima, making it tied to the choice of a good initial gauge. More recent approaches, using the variational formulation for Wannier functions of [DLL19] with *selected columns of the density matrix* (SCDM) initial guesses [DLY15; DL18], or approaches using another spread functional as in [Li+23], seem to provide robust black-box methods for the computation of MLWFs.

4.2.4 The case of moiré materials

Moiré materials, which are the subject of the last chapter of this PhD thesis, have gathered the attention of condensed matter physicist in recent years due to their unique electronic properties offering rich opportunities for both fundamental research and technological innovation. They are created by stacking *two-dimensional materials*, atomically thin crystals such as graphene or transition metal dichalcogenides

(TMDs), on top of each other with a slight rotation of the individual layers. The moiré pattern formed by the superposition of two or more periodic lattices, as appearing in Figure 9, can lead to the emergence of new electronic properties that are not present in the individual constituent layers.

Small changes in the twist angle between layers can lead to significant modifications in the moiré pattern and hence in electronic properties. In addition, the interlayer interactions of moiré materials are typically mediated by van der Waals forces, much weaker that covalent bonds, allowing for precise control over the stacking configuration of the constituent layers. These two factors make the moiré electronic properties highly tunable, promising for a wide range of applications.

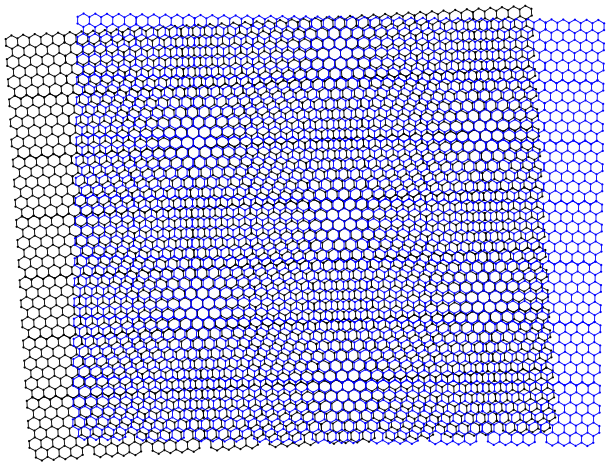


Figure 9 – A sample of twisted-bilayer graphene (TBG). The twist angle between the two graphene layers (respectively black and blue on the picture) creates a characteristic moiré pattern. *Source: adapted from Wikipedia Commons.*

While moiré materials are usually not periodic (except for a countable set of twist angles) they can be approximated as a crystal at mesoscopic scale, with the associated lattice known as the *moiré lattice*. This allows to describe moiré materials in the framework of crystalline electronic structure theory, introduced in the above section. However, the unit cell of the moiré lattice typically contains of the order of thousands of atoms, which makes the direct application of atomic-scale numerical methods such as DFT or tight-binding approximations difficult.

Another class of approach are continuous models, that treat the moiré pattern as a smooth, periodic modulation of the electronic potential arising from the interference between the constituent layers. They are computationally efficient, and capture the essentials physics of moiré systems. Yet, they are typically based on heuristics, and can fail to predict the atomic-scale details and short-range interactions of the material, such as the interlayer interaction.

4.2.5 DFTK: a Julia-based PW-DFT package

This PhD thesis resulted in several numerical contributions, almost all of them in `Julia` language [Bez+17]. In particular the package *Density-functional ToolKit* (`DFTK`) [HLC21] has been a central component in the work presented in Part II of this manuscript, devoted to crystalline materials.

Actively developed since 2019, mainly by Michael Herbst and Antoine Levitt, `DFTK` is designed as a prototyping platform for plane-wave DFT simulations. Remarkably the core features of the package, allowing for the routine treatment of small to medium size systems (up to 1.000 electrons), consists in less than 7.000 lines of codes, entirely in `Julia` language, and publicly available on github.

The current version of `DFTK` implements many features such as an interface with the exchange-correlation functionals library `libxc`, norm-conserving pseudo-potentials (in Kleinman-Bylander form), wannierization, phonon diagrams, among others. A comprehensive description of the code's capabilities can be found in the documentation at `https://docs.dftk.org/stable/features`. Most features are either directly implemented within `DFTK` or are managed by `Julia` packages, rendering the code's inner workings accessible and facilitating the incorporation of novel methods.

DFTK also offers unique functionalities like built-in support for forward automatic differentiation and arbitrary precision. Additionally, the code enables the construction of arbitrary reduced models, such as the Cohen-Bergstresser model or Gross-Pitaevskii-type problems. This flexibility makes DFTK compatible with rigorous numerical analysis and uncertainty quantification, making it a powerful tool for interdisciplinary research, as appearing on the list of related publications https://docs.dftk.org/stable/publications/.

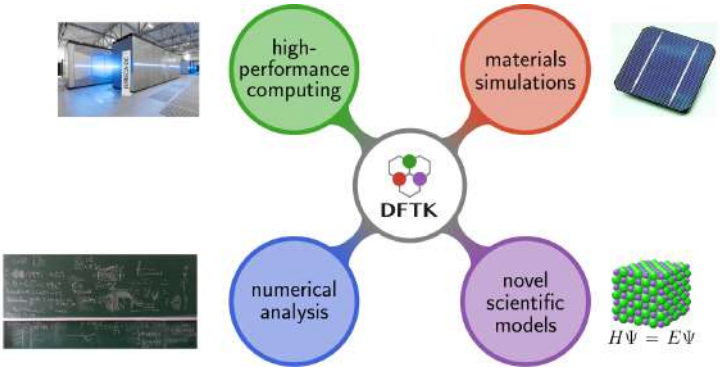


Figure 10 – Four key components of the DFTK software. *Source: Michael F. Herbst* https://dftk.org

5 Contributions of the thesis

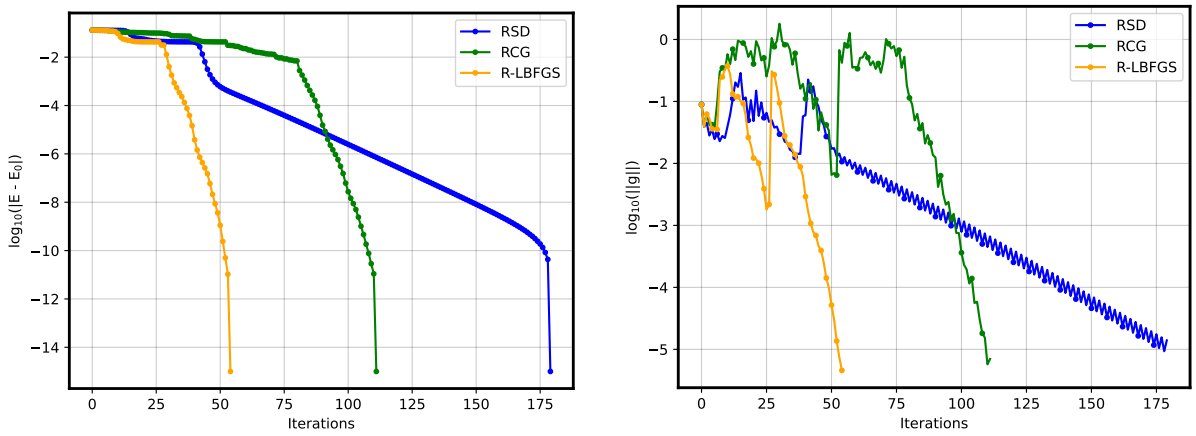
5.1 Results of chapter 1 - Direct minimization in quantum chemistry

Our first contribution is related to the study of direct minimization algorithms applied to the ROHF and CASSCF models, following on from Sections 2.1, 3.1 and 3.2.

When a single group of orbitals is considered (*i.e.* when the internal or the active space is empty), it is well known that the Hartree-Fock and Kohn-Sham DFT models can be formulated as optimization problems on Stiefel (molecular orbital formalism) or Grassmann (density matrix formalism) matrix manifolds, after discretization in a finite basis set. The differential geometry quantities, appearing in the formulation of Riemannian optimization algorithms on these manifolds, have analytic closed form expressions [EAS98; AMS08]. Notably, they can be computed with simple algebraic operations (QR or singular value decomposition) on the MO or DM matrices. As a result, these formulations led to enlightening geometric interpretations of the HF and KS-DFT equations and to the design and implementation of robust and efficient direct minimization algorithms. Nonetheless, due to the low computational cost, satisfactory performance, and simplicity of SCF algorithms, the use of direct minimization methods for HF and KS-DFT is rather marginal in quantum chemistry.

The situation contrasts significantly for ROHF and CASSCF, parametrized by two orthogonal subspaces of orbitals. For these models, SCF algorithms often exhibit a lack of robustness with respect to convergence parameters or initial starting points, and in some cases, they may fail to converge altogether (the stability of SCF algorithms for ROHF is the subject of chapter 2). Direct minimization algorithms are commonplace for CASSCF. They are derived in MO formalism by parametrizing the MO matrix as the exponential of a rotation operator κ, with zero diagonal blocks. By expanding the energy in terms of small variations of κ, one identifies the Riemannian gradient and Hessian on the MO manifold, in terms of one and two-body reduced density matrices. Still, this formulation of Riemannian algorithms hides their geometrical aspects. In ROHF, where the energy is cheaper to evaluate, SCF algorithms still prevail in most quantum chemistry codes. Some direct minimization approaches for ROHF include geometric direct minimization (GDM) methods [DVHG02; VHG02] and the QC-SCF implementation of [NGL21b]. Let us also mention direct minimization applied to the Constrained Unrestricted Hartree-Fock (CUHF) method of [TS10].

From a mathematical point of view, the MO and DM manifolds for ROHF and CASSCF are flag manifolds, which has been recently studied in the context of optimization [YWL22], allowing to adopt a geometric perspective, as in the case of HF and KS-DFT.



Algorithm	$\langle \text{It.} \rangle^{\text{RHF}}$	σ^{RHF}	$\langle \text{It.} \rangle^{\text{UNO}}$	σ^{UNO}
RSD	115.9	112.9	32.3	12.5
RCG	31.9	11.6	15.4	3.0
R-LBFGS(dyn)	37.9	11.3	19.2	3.5
R-LBFGS(fix)	35.6	15.2	19.1	3.7
SCI(DIIS)	38.4	26.0	14.9	6.2
SCI(no DIIS)	61.5	25.1	21.6	9.9
NEO	12.2	1.8	5.1	0.5

Figure 11 – (Up) ROHF calculations on Ti_2O_4 in cc-pVTZ basis from a guess 1 Ha from the expected energy for Riemannian optimization routines. (Up-left) Energy difference with respect to the converged energy along the iterations. (Up-right) Frobenius norm of the Riemannian gradient along the iterations. (Down) Average number of iterations ($\langle \text{It.} \rangle$) and standard deviation (σ) for CASSCF calculation with each tested algorithm, starting with two different guess orbitals corresponding to a *bad* and a *good* initial guess. Details for all methods and initial guesses are given in chapter 1.

In chapter 1, our contribution consists in investigating the ROHF and CASSCF models as optimization problems on a flag manifold. Through this approach, we establish the formulation for Riemannian optimization algorithms tailored to these specific models, providing a unified framework for the direct minimization methods found in the chemistry literature and introducing new ones.

In particular, we discuss the role of parallel transport for these algorithms. This differential geometry toolbox offers a proper way to compare two directions (such as gradients of the energy) that belong to different tangent spaces. In HF and KS-DFT, parallel transport on the Grassmann manifold is trivial. On the other hand, this is not the case on the flag manifold, a fact that that has apparently not been identified in the theoretical chemistry literature and may contribute to some observed instabilities.

The rest of our contribution is devoted to numerical experiments. We test the viability of first-order optimization methods for ROHF and CASSCF. We have implemented these methods in a `Julia` code, which is interfaced with `PySCF` and `CFOUR` to evaluate the respective ROHF and CASSCF energies and gradients. For ROHF we tested our methods on Ti_2O_4 in its D_{2h} geometry, which is employed as a template for addressing SCF convergence issues[4] in the Amsterdam Density Functional (ADF) quantum-chemistry package [TV+01]. For CASSCF we used a subgroup of the benchmark set used in [MKW16] and [NGL21a]. In all cases, Riemannian optimization methods show robust convergence properties, and do so without requiring the user to finely tune the parameters that control the optimization. Even in our naive implementation, they demonstrate that they can be competitive with other traditional implementations in terms of number of iterations, and thus overall computational cost (Figure 11).

It would be a natural progression of our study to explore alternative methods of retraction and transport, distinct from the exponential retraction and parallel transport used in chapter 1. These alternative approaches have the potential to significantly enhance the performance of Riemannian optimization algo-

[4] See https://www.scm.com/doc/ADF/Examples/SCF_Ti2O4.html

rithms. Some examples of other retractions and transports on flag manifolds are given for Riemannian conjugate gradient in the recent paper [ZS24].

5.2 Results of chapter 2 - SCF algorithms for ROHF

Our second contribution is related to the study of SCF algorithms in Restricted Open-shell Hartree-Fock (ROHF), following on from Sections 2.1.2, 2.1.5 and 3.3.

Self-consistent field algorithms for the Hartree-Fock problem in the restricted closed-shell and unrestricted open-shell settings are well understood. Several flavors of SCF algorithms have been proposed for RHF and UHF in the past 70 years, the most common being the Roothaan's algorithm [Roo51] endowed with level shifting [SH73] or DIIS acceleration methods [Pul80; Pul82; HP86; RS11]. In [KSC02], the authors propose a robust and efficient method to solve the RHF and UHF problems using their EDIIS algorithm for the first iterations, and switching to DIIS to accelerate convergence when the iterates are close enough to the solution. This method always work for UHF and most of the time for RHF. The SCF algorithms for RHF and UHF have also been studied from a mathematical viewpoint [Can+03; CKL21; Chu+21].

As mentioned above, the situation is radically different for the ROHF model, where existing SCF algorithms fail to converge in many cases, notably for radicals and molecular systems containing transition metals.

In chapter 2, we investigate SCF algorithms for ROHF. We start by writing the ROHF problem in DM and OMO formalisms, as the minimization of an energy functional on the flag manifolds $\mathcal{M}_{\mathrm{OMO}}(N_d, N_s; \mathbb{R}^{N_b})$ (3.12) and $\mathcal{M}_{\mathrm{DM}}(N_d, N_s; \mathbb{R}^{N_b})$ (1.2.10). Here N_d and N_s denote the respective numbers of *doubly-occupied* and *singly-occupied* orbitals. We then discuss the fact that, in contrast to RHF and UHF, the first-order optimality conditions for ROHF (the ROHF equations) cannot be *naturally* formulated as a nonlinear eigenvalue problem.

Standard SCF (parameter dependent). As a result, standard SCF algorithms for ROHF are based on the construction and diagonalization of a non-physical effective Hamiltonian $H_{A,B}$, depending on six user-defined parameters $A = (A_{dd}, A_{ss}, A_{vv}) \in \mathbb{R}^3$ and $B = (B_{dd}, B_{ss}, B_{vv}) \in \mathbb{R}^3$ characterizing the SCF algorithm. Notably, the effective Hamiltonian $H_{A,B}$ does not always satisfy the *Aufbau* principle, which is a key ansatz for the stability and convergence of SCF algorithms.

New SCF (parameter free). Following this observation, we propose a new SCF scheme that better respects the structure of the ROHF equations, and that does not rely on the *Aufbau* principle. We show that optimal points for this new SCF scheme are optimal for the ROHF problem. From an initial point $(P_d^{(k)}, P_s^{(k)})$, the next iterate is chosen as a minimizer of a linear functional on $\mathcal{M}_{\mathrm{DM}}(N_d, N_s; \mathbb{R}^{N_b})$

$$\boxed{(P_d^{(k+1)}, P_s^{(k+1)}) \in \operatorname{argmin} \left\{ \operatorname{Tr}(F_d^{(k)} P_d + F_s^{(k)} P_s), \quad (P_d, P_s) \in \mathcal{M}_{\mathrm{DM}}(N_d, N_s; \mathbb{R}^{N_b}) \right\}} \tag{5.1}$$

which can be evaluated by a few iterations of a direct minimization procedure on $\mathcal{M}_{\mathrm{DM}}(N_d, N_s; \mathbb{R}^{N_b})$, as described in chapter 1. The formula (5.1) is obtained by analogy to the formulation of the standard SCF iteration for RHF as a minimization problem on the Grassmann manifold $\mathrm{Grass}(N_o; \mathbb{R}^{N_b})$ (3.16). Using this new SCF iteration, free of the *Aufbau* ansatz, we also extend the relaxed-constrained ODA algorithm [CB00] to the ROHF setting.

To assess their performance, we tested our new algorithms against state-of-the-art SCF algorithms for some challenging chemical systems, such as organic ligands chelating - or simply interacting with - transition metals. A sample of our results is given in Table 2.

Performances of standard DIIS. Numerical experiments show that standard SCF methods, even endowed with DIIS acceleration, fail to converge for most choices of coefficients A and B in the literature. In general, the standard methods are very sensitive to the choice of initial guess and DIIS parameters. Starting from a good initial guess (very close to a minimum) does not always provide better convergence results.

We observe however that the standard SCF method with *Guest and Saunders* coefficients converges for almost all our test cases, when applying DIIS acceleration from the first iteration. This is at odds with the default implementation of most quantum chemistry codes, in which DIIS only activates when the iterations reach the attraction basin of a local minimum. To our knowledge, the fact that DIIS can stabilize SCF iterations far from a local minimum remains unexplained.

Performance of our new methods. As for the EDIIS + DIIS method in the closed shell case, numerical experiments show that applying a few iterations of ODA, followed by a new SCF endowed with DIIS acceleration provides a robust black-box method that converges in all our test cases.

Standard SCF-DIIS	Pyridine–Fe^{2+}	Pyridine–Fe^{3+}	Porphyrin model–Fe^{2+}
Guest and Saunders	✓(100)	✓(187)	
Roothaan	✓(212)	✓(139)	✓(52)
Euler	✓(68)		✓(72)
McWeeny	✓(271)		✓(187)
Other coefficients			
ODA + new SCF-DIIS	✓(60)	✓(144)	✓(28)

Table 2 – Convergence results of standard SCF methods (designated by their names in the literature) and of our new black-box method for some of our test cases, starting from an extended Hückel guess in 6-31G basis set. The results for standard SCF are from our own implementation. The details of implementation are given in chapter 2. The cell contains the number of iterations to reach microHartree convergence. The cell is barred when the method do not converge.

In their current state, our new methods are computationally more demanding than traditional SCF algorithms (when the latter converge), as they require to approximately solve the minimization problem (5.1) at each iteration. Yet they are a promising step toward a black-box algorithm for open-shell systems. A natural follow up to our study is the realization of a high-throughput calculations to assess the validity of our methods on a large set of test cases. One should also do a complete analysis of the local minima respectively found with standard SCF and our methods, as we only compared their respective energies in our analysis. Lastly, as discussed in chapter 2, the fixed depth of the DIIS history sometimes hinder convergence by keeping information that should be discarded. Hence, it would be interesting to test on challenging cases the use of the adaptative depth DIIS procedure, as introduced in [Chu+21], together with our black-box routine.

5.3 Results of chapter 3 - General criteria for the optimization of LCAO bases

Our third contribution is related to the mathematical formalization and to the choice of optimality criteria for the optimization of atomic basis sets in quantum chemistry, following on from Section 3.4.

The LCAO approach is a standard discretization strategy in electronic structure calculations: for a large number of cases, it offers a computationally efficient discretization method, while requiring a small number of basis functions. To produce *good* approximations, AO basis sets are typically the result of an optimization process. However, in view of the vast number of molecular configurations and quantities to approximate, choosing an appropriate reference data-set for the optimization is a difficult task.

This yields many different approaches. Classical methods are restricted to gaussian-type orbitals basis sets, thus reducing the dimensionality of the problem by optimizing a reduced set of contracting coefficients and/or exponents. Among them, the "energy-based" basis sets are the most common. These include the standard Pople [BPH80] and Dunning [Dun89], as well as the more recent Karlsruhe [WA05] basis sets, which are optimized with respect to Hartree-Fock and/or post-Hartree Fock energies, mostly for single atoms. Let us also cite Jensen [Jen01] polarization consistent GTO basis sets which involve reference energies for polyatomic systems.

Another class of methods which we might call "state-based", more scarcely represented in the literature, directly involve reference wave-functions or density matrices. An example is given by [SPAS95; SPAS96] where GTOs are fitted to reference states obtained with plane-wave calculations. The recent development

of fully numerical atomic orbitals, which are free of many constraints imposed on GTOs, produced yet another class of methods.

The diverse range of methods employed for basis set optimization raises methodological questions. While "energy-based" approaches are typically simpler to implement, certain physical quantities are not directly correlated with energy. This raises the question of whether "state-based" methods should be developed, or if these approaches are somehow equivalent. Moreover, it remains unclear whether it is sufficient to optimize basis sets using atomic or diatomic configurations, as commonly practiced, or if one needs to include general polyatomic systems in the optimization data-set. Furthermore, LCAO basis sets inherently produce significant linear dependencies that must be addressed during optimization. Recent efforts aimed at curing the poor conditioning of AO basis sets include [Leh19b; Leh24].

These considerations point toward the need to investigate AO optimization from a general mathematical perspective. Mathematical studies proving convergence rates or proposing systematic enrichment of AO basis sets are so far quite limited. The approximability of solutions to electronic structure problems by Gaussian functions was studied in [Kut94], and later on in [SY17; Sha20]. An *a priori* error estimate on the approximation of Slater-type functions by Hermite and even-tempered Gaussian functions was derived in [BCS14]. A construction of Gaussian bases combined with wavelets was proposed on a one-dimensional toy model in [Pha17].

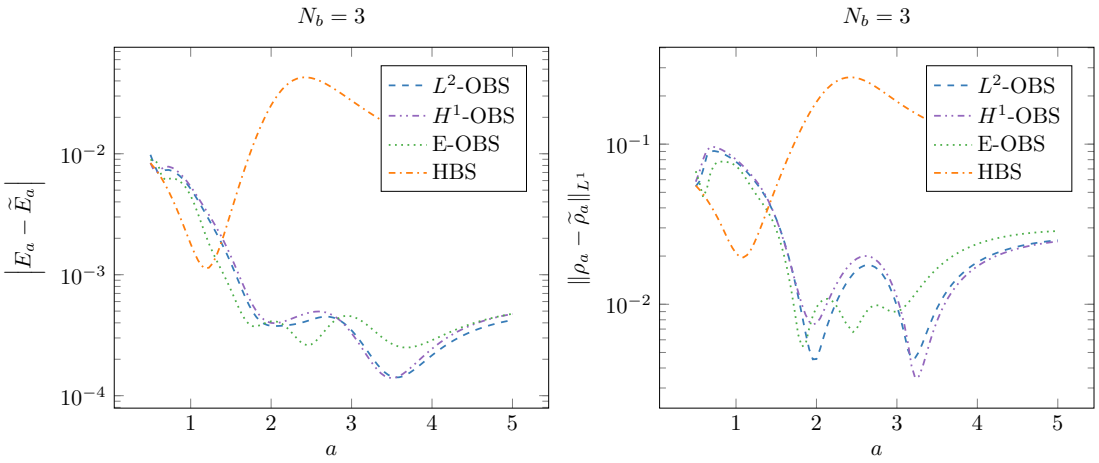


Figure 12 – Energy and densities error with basis functions optimized on $\mathcal{I} = [1.5, 5]$. The conventions are the following: the L^2-OBS and H^1-OBS basis are obtained for the J_A criterion, for A corresponding to the L^2 and H^1 norm. The E-OBS basis is obtained for J_E. HBS is the non-optimized GTO basis. The parameter a models the inter-atomic distance of our toy diatomic molecule.

In chapter 3, we introduce an abstract mathematical framework for the optimization of AO basis sets based on the choices of

1. a set of admissible atomic configurations Ω;
2. a probability measure $\mathbb{P}$ on Ω;
3. a set of admissible AO basis sets $\mathcal{B}$;
4. a criterion $j(\chi, \omega)$ quantifying the error between the exact values of the quantities of interest when the system has atomic configuration $\omega \in \Omega$ – for the continuous model under consideration – and the ones obtained after discretization in the basis set $\chi \in \mathcal{B}$.

In this formalism, we investigate possible choices of Ω, $\mathbb{P}$, $\mathcal{B}$ and criterion j. In particular, we define an "energy-based" and a "state-based" criterion, denoted J_E and J_A (with the operator A characterizing a choice of norm) and experiment with these criteria on a one-dimensional toy problem mimicking the optimization of GTO basis sets for diatomic configurations. The exact formulations for J_E and J_A are given in chapter 3.

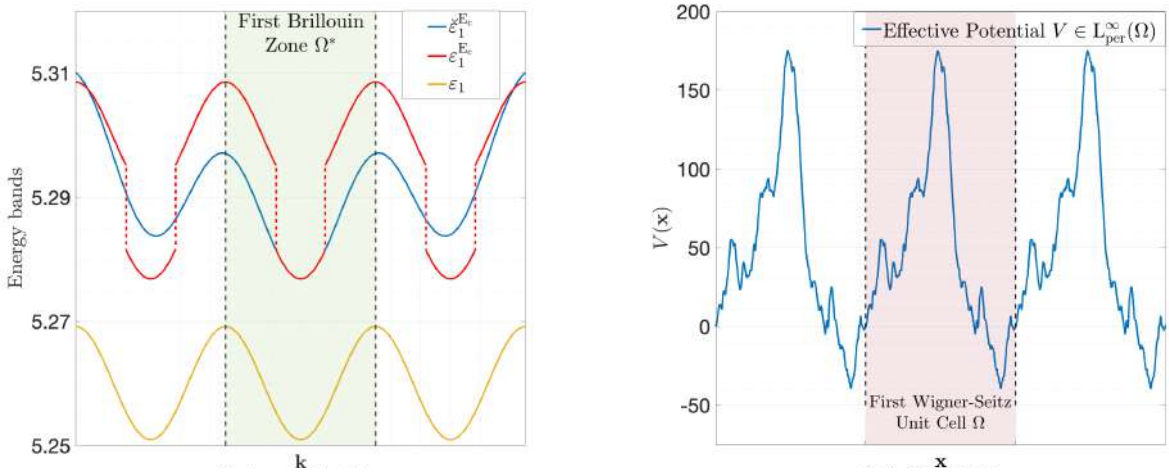
As seen in Figure 12, the optimized basis sets for the respective criteria provide approximations that are one to two orders of magnitude closer to the reference solutions than standard basis sets, for the same number of basis functions. Remarkably on this 1D toy model, the energy and state based optimization procedures seem to provide similar results, in terms of approximation of the ground-state energy and ground-state 1-RDM. We also observed that this improvement of the approximation holds for a small number of reference points close to the minimum of the dissociation curve.

In future work, we plan to generalize our study to small real systems. Reference solutions can be obtained for example using finite element approximations, as available in the `Helfem` software [Leh19c] for diatomic molecules, or in `DFT-FE` [Das+22] for larger systems. Some preliminary tests using `Helfem` yield encouraging results. Another study should be devoted to the bad conditioning of AO basis sets.

5.4 Results of chapter 4 - Modified operator for the computation of band diagrams

Our fourth contribution is related to the correction of discretization errors that arise when using truncated Fourier basis sets in the computation of band diagrams. It follows on from Section 4.

From the second point of Theorem 4.1, the exact energy bands $\mathbf{k} \mapsto \varepsilon_{n,\mathbf{k}}$ of a crystalline materials are $\mathcal{R}^*$-periodic and analytic away from band crossings. Those are two properties that are crucial for the convergence rate of quadrature methods used to compute integrals on Ω^* [Can+20]. However, as observed in Figure 4.2, classical uniform (4.21) and $\mathbf{k}$-dependent (4.23) plane-wave discretization strategies sometimes produce approximate energy bands that are discontinuous or aperiodic. Some corrective approaches have been proposed in the literature, for example the ones respectively implemented in `Abinit` [Abi] or `Qbox` [Qbo] software, all based on a specific modification of the kinetic operator of the Bloch fibers $\hat{\mathrm{H}}_{\mathbf{k}}$.



(a) Approximate and exact energy bands

(b) Corresponding $\mathcal{R}$-periodic potential $V \in L^\infty_{\mathrm{per}}(\Omega)$

Figure 13 – Lowest energy bands for a simple 1-D example with effective potential $V \in L^\infty_{\mathrm{per}}(\Omega)$ as shown on the left. The effective potential satisfies the regularity property $V \in H^{1-\epsilon}_{\mathrm{per}}(\Omega)$ for every $\epsilon > 0$.

In chapter 4, we describe a systematic modified operator approach that encompasses existing methods and which allows to produce periodic bands with arbitrary targeted regularity. Given an initial $\mathbf{k}$-dependent Fourier basis $\mathcal{X}^{\mathrm{E_c}}_{\mathbf{k}}$ with cut-off energy $\mathrm{E_c} > 0$, our approach introduces the modified Bloch fiber

$$\hat{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}} := \mathrm{E_c}\, \mathscr{G}\left(\frac{|-i\nabla + \mathbf{k}|}{\sqrt{2\mathrm{E_c}}}\right) + V_{\mathrm{per}} \tag{5.2}$$

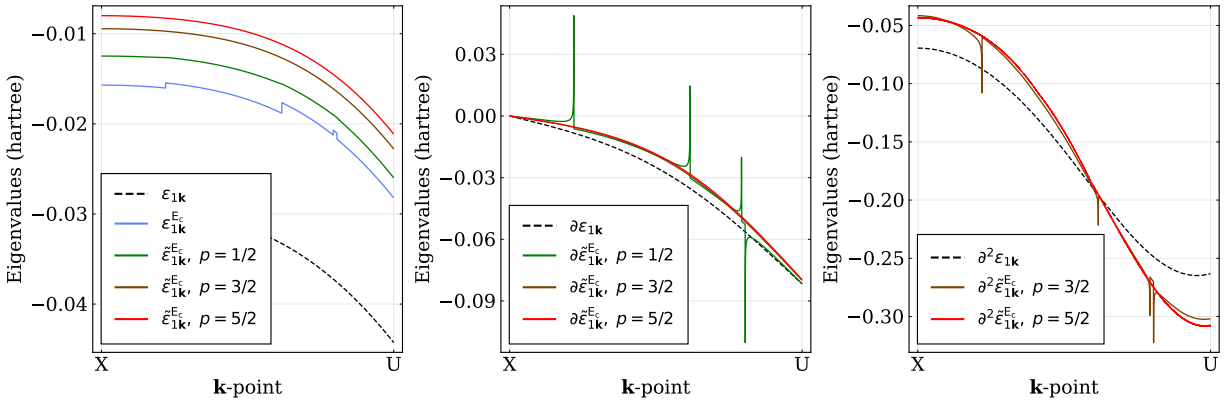
where $\mathscr{G}\colon \mathbb{R} \to \mathbb{R}$ is a one-dimensional "blow-up" function defined as

$$\mathscr{G}(x) = \begin{cases} f_2(x) & \text{for } |x| \in [0, \tfrac{1}{2}] \ \cup \ [1, \infty), \\ h_m(|x|) & \text{for } |x| \in \left(\tfrac{1}{2}, 1\right) \end{cases} \tag{5.3}$$

and where $h_m : [\frac{1}{2}, 1) \longrightarrow \mathbb{R}$, $h_m(|x|) \underset{|x|\to 1}{\to} +\infty$ is chosen to obtain a targeted regularity of the energy bands depending on $m \in \mathbb{N}$. Possible expressions for h_m are given in chapter 4. Under suitable regularity assumptions on V_{per}, we derive an error estimate (Theorem 4.5.1) for the approximate, modified energy bands $\mathbf{k} \mapsto \varepsilon_{n,\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}$. We then prove (Theorem 4.5.2) that the approximate bands $\mathbf{k} \mapsto \varepsilon_{n,\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}$ are of class C^m away from band crossings, and that they are Lipschitz continuous at crossings if $m \geqslant 1$ and only continuous otherwise.

The proof for the first theorem is an application of the Courant-Fischer min-max principle, which allows to obtain a first upper bound for the error introduced by the modified operator approach. The theorem is then proved by writing this upper bound as the sum of terms that either cancel by properties of $\mathscr{G}$, or that can be bounded in a controlled manner.

The second proof starts by showing the continuity of the modified approximate bands with respect to $\mathbf{k} \in \Omega^*$. The difficulty of the proof comes from the fact that the size of the discretization basis $\mathcal{X}_{\mathbf{k}}^{\mathrm{E_c}}$, hence the size of the matrix $\Pi_{\mathbf{k}}^{\mathrm{E_c}} \hat{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}} \Pi_{\mathbf{k}}^{\mathrm{E_c}}$ representing $\hat{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}$ in the truncated basis $\mathcal{X}_{\mathbf{k}}^{\mathrm{E_c}}$, varies with $\mathbf{k}$. For that reason we could not use the standard result from which a continuously parametrized family of fixed-size matrices have continuously parametrized eigenvalues. As detailed in chapter 4, we get around the problem by using a Schur complement to isolate the dimensions added to $\Pi_{\mathbf{k}}^{\mathrm{E_c}} \hat{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}} \Pi_{\mathbf{k}}^{\mathrm{E_c}}$ by an infinitesimal variation of $\mathbf{k}$, and by using the Hurwitz theorem from complex analysis. We prove higher regularity of the bands using a finite difference approximation.



(a) Energy band (b) First derivative (c) Second derivative

Figure 14 – Comparison of the first and second derivatives of the first band of face-centered cubic silicon on a small portion of the band-structure (detailed in chapter 4) for the **k**-dependent and modified discretization schemes.

Numerical experiments involving a toy model in 1D, graphene in 2D, and face-centered cubic silicon in 3D (shown in Figure 14) then validate our theoretical results and showcase the efficiency of the operator modification approach. The PW-DFT computations have been performed with `DFTK` (see Section 4.2.5), where we implemented the modified operator approach. The details of implementation can be found in chapter 4. Our modified operator approach also proves effective in regularizing the energy with geometry optimization.

Band diagrams are usually the first step in the evaluation of electronic properties and quantities of interest for materials, such as elastic constants, Wannier functions, Berry curvature, etc. A study of the impact of our modified operator approach on the convergence of these quantities with respect to real-space or Brillouin zone samplings would be an interesting topic of investigation. Preliminary results involving finite difference approximations of the energy with respect to the lattice parameter point toward practical applications.

5.5 Results of chapter 5 - Contributions to the Julia electronic structure ecosystem

Lastly, we describe two numerical contributions related to the simulation of twisted-bilayer graphene (TBG), following on from Section 4.

5.5.1 Developpment of a Julia package for the simulation of 2D materials

Our first contribution consisted in establishing the groundwork for a `Julia` code, `TwistedBilayerGraphene`, designed as a user-friendly playground for the simulation of twisted-bilayer graphene (TBG). As a starting point, we focused on two continuous models for the electronic structure of TBG. First, the Bistritzer-MacDonald (BM) model [BM11], a standard effective model introduced in 2011. Second, the more recent model introduced in [CGG23], that we call "CGG", derived from an approximate Kohn-Sham Hamiltonian for TBG.

In the first part of chapter 5, we compute the expressions for the respective BM and CGG models in a plane-wave discretization basis, and provide implementation details, to be shared as a public documentation. Our package, built as an overlay to the `DFTK` package, allows in its current state to compute the BM and CGG band diagrams in a few simple calls, with tunable geometry and convergence parameters (see Figure 15). It also includes automated tests to ensure the code's resilience over time, enabling adaptation to future updates of `Julia` and `DFTK`.

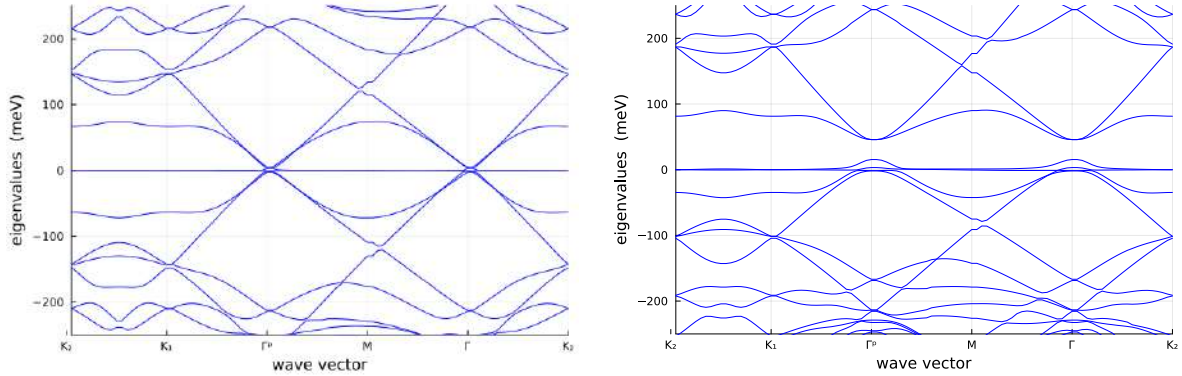


Figure 15 – (Left) BM and (right) CGG band diagrams of TBG, as introduced in [CGG23]. The precise definition of the **k**-path in horizontal axis and of the geometry used for these band diagrams are given in chapter 5.

Future work will be devoted to the implementation of additional features in our package. For example, the recent introduction of phonon mode calculations in `DFTK` will allow the integration of phonon or electron-phonon models for TBG in our code.

5.5.2 First steps toward large tight-binding simulation of multilayer graphene with compressed Wannier functions

Our last contribution focuses on the evaluation of matrix elements for large-scale tight-binding calculations on multilayer graphene. Applying a general compression procedure introduced in [Bak+18], we manage to expand the Wannier functions corresponding to the two lowest valence bands of graphene on a basis of symmetry-adapted gaussian-type orbitals (SAGTOs), for which tight-binding elements can be computed analytically.

The obtained compressed Wannier functions depend on 155 parameters only and approximate the reference Wannier functions with an error of 12% in H^1 norm and 5% in L^2 norm. Notably, our set of GTOs is larger than the one used in the original paper [Bak+18], the latter requiring more than twice as many parameters to achieve the same precision.

Unfortunately, we lacked the time to complete this study. In particular, the SAGTO basis produced by our naive implementation is ill-conditioned, preventing further investigation into the use of larger and

potentially more accurate basis sets. In forthcoming research, we should work on correcting the bad conditioning inherent to our implementation. We also should assess the time saved through the use of analytical integrals as opposed to standard quadrature methods and measure the effect of compression on the accuracy of tight-binding matrix element calculations.

Part I

Quantum Chemistry

Chapter 1

Geometric Optimization of Restricted-Open and Complete Active Space Self-Consistent Field Wavefunction

This chapter resulted in the preprint [LVp1]:

Laurent Vidal, Tommaso Nottoli, Filippo Lipparini, and Eric Cancès. "Geometric optimization of Restricted-Open and Complete Active Space Self-Consistent Field wavefunctions". Submitted

Abstract We explore Riemannian optimization methods for Restricted-Open-shell Hartree-Fock (ROHF) and Complete Active Space Self-Consistent Field (CASSCF) methods. After showing that ROHF and CASSCF can be reformulated as optimization problems on so-called flag manifolds, we review Riemannian optimization basics and their application to these specific problems. We compare these methods to traditional ones and find robust convergence properties without fine-tuning of numerical parameters. Our study suggests Riemannian optimization as a valuable addition to orbital optimization for ROHF and CASSCF, warranting further investigation.

Contents

1.1 Introduction

Orbital optimization is one of the most common task performed in quantum chemistry calculations. It is the numerical problem associated with Hartree-Fock (HF) [Har57] and Kohn-Sham Density Functional Theory (KS DFT) [KS65] as well as a component of Complete Active Space Self-Consistent Field (CASSCF) calculations [Wer87; She87; Roo87] and is further encountered in orbital optimized post-Hartree Fock methods [SSI87; She+98]. The various algorithms that have been propoposed to tackle this problem can be grouped into two families: fixed point methods, such as Roothaan's SCF algorithm [Roo51; Roo60], and direct optimization methods, such as quadratically convergent optimization strategies [Bac81; Bac82]. For HF and DFT, the former family is the most commonly employed, due to the existence of very robust implementation that exploit convergence acceleration techniques such as Pulay's Direct Inversion in the Iterative Subspace [Pul80; Pul82; HP85] (DIIS), constraint relaxation methods such as the Optimal Damping Algorithm [CLB00a; CB00; Can01] (ODA), or more sophisticated related techniques such as E-DIIS [KSC02] or A-DIIS [HY10]. Nevertheless, direct optimization techniques have received quite some attention due to their robustness and due to the possibility of implementing them avoiding dense linear algebra operations (e.g., diagonalization of the Fock matrix).

Direct minimization techniques for Restricted Open-Shell Hartree-Fock (ROHF) calculations are relatively scarce compared to Self-Consistent Field (SCF) methods, which predominantly feature in quantum chemistry software. Among direct minimization approaches, noteworthy methods include Geometric Direct Minimization (GDM) techniques [DVHG02] and the QC-SCF algorithm [NGL21b]. Additionally, the Second-Order SCF (SOSCF) algorithm [CSG97; Nee00] and the DIIS-GDM method [DVHG02; VHG02], which amalgamate aspects of both SCF and direct minimization strategies, merit mention. Moreover, the CUHF method, as introduced by Tsuchimochi and Scuseria [TS10], can be adapted for ROHF computations through the utilization of a direct minimization procedure designed for Unrestricted Hartree-Fock (UHF) calculations.

A difficulty associated with the formulation and implementation of direct minimization techniques is due to the fact that the quantity that needs to be optimized, such as the molecular orbitals (MO) coefficients, or the density matrix, needs to satisfy nonlinear constraints. In other words, the minimization set is not a vector space, but rather a differentiable manifold.

It is well-known that after discretization in a finite basis set, HF and KS models can be formulated as optimization problems on Stiefel (molecular orbital formalism) or Grassmann (density matrix formalism) manifolds [EAS98]. These formulations lead to enlightening geometric interpretations of the Hartree-Fock and Kohn-Sham equations, and to the design and convergence analysis [CLB00b; Lev12; CKL21] of robust and efficient direct minimization algorithms. The purpose of this article is to show that Restricted Open-Shell Hartree-Fock (ROHF) and Complete Active Space Self-Consistent Field (CASSCF) methods can be reformulated as optimization problems on so-called flag manifolds [YWL22]. This allows one to shed new light on the ROHF and CASSCF equations, and the direct minimization algorithms used to solve these problems. While the work presented in this article does not lead to game-changing improvements in orbital optimization, we hope that it will provide the community with a set of rigorous tools that can be used for further developments, as the ones recently proposed by some of us for the extrapolation of the SCF density matrix in the context of ab-initio molecular dynamics simulations [Pes+23].

This article is organized as follows. In Section 1.2, we briefly recall the high-spin ROHF and CASSCF orbital optimization problem in terms of both density-matrix (DM) and molecular orbitals (MO) formulations and provide a simple geometric interpretation of the ROHF and CASSCF equations. In Section 1.3, we review the basic concepts of geometric optimization (Riemannian gradient and Hessian, vector transport, affine connection, geodesic, retraction). In Section 1.4, we discuss more specifically geometric optimization for ROHF and CASSCF, providing also the tools to translate any algorithm formulated in the MO formalism into the DM formalism and viceversa. We then provide geometric interpretations of existing direct minimization algorithms for ROHF and CASSCF, and propose new ones. We also introduce a direct optimization method circumventing the use of virtual orbitals, which is useful for ROHF in large basis sets (i.e. planewaves, finite elements, or wavelets). Numerical results are reported in Section 2.4.

1.2 ROHF and CASSCF

In this section, we briefly recapitulate the orbital optimization problem for ROHF and CASSCF and introduce the manifolds associated with the MO and DM formalisms. ROHF and CASSCF methods indeed share common features. They both involve

- a set of N_I doubly-occupied molecular orbitals, often called internal orbitals;
- a set of N_A partially-occupied molecular orbitals, often called active orbitals,

the latter being orthogonal to the former. Consider a molecular system with N electrons discretized in a basis set of size $\mathcal{N}_\mathrm{b}$. We denote the electronic Hamiltonian by

$$\widehat{H}_N = -\frac{1}{2}\sum_{i=1}^{N} \nabla^2_{\mathbf{r}_i} + \sum_{i=1}^{N} V_\mathrm{nuc}(\mathbf{r}_i) + \sum_{1\leq i<j\leq N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}. \tag{1.2.1}$$

Notation To avoid possible misunderstandings, we clarify here the notation that will be adopted throughout the paper. Let us assume that we have discretized the problem using an orthonormal set of atomic orbitals (AO) $\{\chi_\mu\}_{\mu=1}^{\mathcal{N}_\mathrm{b}}$, obtained, for instance, by Löwdin orthogonalization of a usual Gaussian-type or Slater-type basis. This implies, for the overlap matrix:

$$S_{\mu\nu} = \langle \chi_\mu | \chi_\nu \rangle = \delta_{\mu\nu}.$$

We use greek letters $\mu, \nu, \ldots$ to label AOs. Molecular orbitals $\{\phi_p\}_{p=1}^{\mathcal{N}_\mathrm{b}}$ are written as linear combinations of atomic orbitals

$$\phi_p = \sum_{\mu=1}^{\mathcal{N}_\mathrm{b}} C_{\mu p} \chi_\mu$$

where the coefficient matrix $C \in \mathbb{R}^{\mathcal{N}_\mathrm{b}\times\mathcal{N}_\mathrm{b}}$ is, due to our choice of orthogonal AOs, an orthogonal matrix, i.e., $CC^T = C^TC = I_{\mathcal{N}_\mathrm{b}}$. We divide the molecular orbitals into three sets: N_I *internal* orbitals, that are always doubly occupied, N_A *active* orbitals, that are singly occupied in ROHF and have varying occupation for CASSCF, and N_E *external* orbitals, that are always empty. We use $i, j, \ldots$ to label internal orbitals, $u, v, \ldots$ to label active orbitals, $a, b, \ldots$ to label external orbitals, and $p, q, \ldots$ for generic ones. We call Π^I and Π^A the orthogonal projectors on the space spanned by internal and active orbitals, respectively, in the AO basis, i.e.,

$$\Pi^I_{\mu\nu} = \sum_{i=1}^{N_I} C_{\mu i} C_{\nu i}, \quad \Pi^A_{\mu\nu} = \sum_{u=N_I+1}^{N_I+N_A} C_{\mu u} C_{\nu u} \tag{1.2.2}$$

$P \in \mathbb{R}^{\mathcal{N}_\mathrm{b}\times\mathcal{N}_\mathrm{b}}$ denotes the one-body reduced density matrix (1-RDM) in the AO basis, while $\gamma \in \mathbb{R}^{\mathcal{N}_\mathrm{b}\times\mathcal{N}_\mathrm{b}}$ denotes the one-body reduced density matrix in the MO basis. We call m_α and m_β the number of α and β *active* electrons, respectively, so that the total number of electrons is given by $N = 2N_I + m_\alpha + m_\beta$. In high-spin ROHF, $m_\alpha = N_A$ and $m_\beta = 0$. Using these conventions, the high-spin ROHF density matrix is given by

$$P^\mathrm{ROHF} = 2\Pi^I + \Pi^A \tag{1.2.3}$$

where we note that the α and β one-body spin density matrices (1-SDM) are given by

$$P^{\mathrm{ROHF},\alpha} = \Pi^I + \Pi^A, \quad P^{\mathrm{ROHF},\beta} = \Pi^I \tag{1.2.4}$$

For CASSCF, the MO 1-RDM $\gamma^{\Psi^\mathrm{CAS}}$ has a block structure, with

$$\gamma^{\Psi^\mathrm{CAS}}_{ij} = 2\delta_{ij}, \quad \gamma^{\Psi^\mathrm{CAS}}_{uv} = \langle \Psi^\mathrm{CAS} | \hat{E}_{uv} | \Psi^\mathrm{CAS} \rangle, \quad \gamma^{\Psi^\mathrm{CAS}}_{ab} = 0, \tag{1.2.5}$$

where

$$\hat{E}_{pq} = \hat{a}^\dagger_{p\alpha}\hat{a}_{q\alpha} + \hat{a}^\dagger_{p\beta}\hat{a}_{q\beta}$$

is the spin-traced singlet excitation operator, and all other blocks vanishing. In the AO basis, the matrix $P^{\Psi^\mathrm{CAS}} = C\gamma^{\Psi^\mathrm{CAS}}C^T$ thus satisfies the following relation:

$$2\Pi^I \leqslant P^{\Psi^\mathrm{CAS}} \leqslant 2\Pi^I + \Pi^A. \tag{1.2.6}$$

To define 1-SDM for CASSCF, we need to introduce the active spin densities $\gamma_{uv}^{\Psi^{\mathrm{CAS}},\sigma}$, for $\sigma = \alpha, \beta$ which are defined as follows:

$$\gamma_{ij}^{\Psi^{\mathrm{CAS}},\sigma} = \delta_{ij}, \quad \gamma_{uv}^{\Psi^{\mathrm{CAS}},\sigma} = \langle \Psi^{\mathrm{CAS}} | \hat{a}_{u\sigma}^{\dagger} \hat{a}_{v\sigma} | \Psi^{\mathrm{CAS}} \rangle, \tag{1.2.7}$$

with all the other blocks (i.e., internal-active and all blocks with at least one external index) vanishing. The AO spin densities are then given by

$$P^{\Psi^{\mathrm{CAS}},\sigma} = C \gamma^{\mathrm{CAS},\sigma} C^T \tag{1.2.8}$$

and it holds that

$$\Pi^I \leqslant P^{\Psi^{\mathrm{CAS}},\sigma} \leqslant \Pi^I + \Pi^A. \tag{1.2.9}$$

Direct optimization methods for ROHF and CASSCF can be divided into two groups, depending on the degrees of freedom used for performing the optimization. In the MO formalism, the main variable is the coefficient matrix C. As mentioned previously, it is an orthonormal matrix, which can be seen as a point of the orthogonal group $O(\mathcal{N}_b)$. In the DM formulation, the main variable is the pair of orthogonal projectors $(\Pi^{\mathrm{I}}, \Pi^{\mathrm{A}})$, which can be identified with a point of the set

$$\mathcal{M}_{\mathrm{DM}} := \Big\{ (\Pi^{\mathrm{I}}, \Pi^{\mathrm{A}}) \in \mathbb{R}^{\mathcal{N}_{\mathrm{b}} \times \mathcal{N}_{\mathrm{b}}}_{\mathrm{sym}} \times \mathbb{R}^{\mathcal{N}_{\mathrm{b}} \times \mathcal{N}_{\mathrm{b}}}_{\mathrm{sym}} \text{ s.t. } (\Pi^{\mathrm{I}})^2 = \Pi^{\mathrm{I}},\ (\Pi^{\mathrm{A}})^2 = \Pi^{\mathrm{A}} \\ \mathrm{Tr}(\Pi^{\mathrm{I}}) = N_I,\ \mathrm{Tr}(\Pi^{\mathrm{A}}) = N_A,\ \text{and } \Pi^{\mathrm{I}}\Pi^{\mathrm{A}} = 0 \Big\}. \tag{1.2.10}$$

We will see later that the above set has a nice geometrical structure: it can be canonically identified with the flag manifold $\mathcal{M}_{\mathrm{Flag}} := \mathrm{Flag}(N_{\mathrm{I}}, N_{\mathrm{I}} + N_{\mathrm{A}}; \mathbb{R}^{\mathcal{N}_{\mathrm{b}}})$. The passage between MO and DM parameterization is done by the map

$$\zeta : O(\mathcal{N}_b) \ni C \mapsto \left(\Pi^{\mathrm{I}}, \Pi^{\mathrm{A}}\right) \in \mathcal{M}_{\mathrm{MO}} \qquad \text{with } \Pi^{\mathrm{I}}, \Pi^{\mathrm{A}} \text{ given by Eq. 1.2.2.} \tag{1.2.11}$$

The dimension of the MO manifold $O(\mathcal{N}_b)$ is $\frac{\mathcal{N}_b(\mathcal{N}_b-1)}{2}$ (i.e. the number of degrees of freedom in an orthogonal matrix), while the dimension of the DM manifold $\mathcal{M}_{\mathrm{DM}}$ can be shown to be $N_I N_A + N_I N_E + N_A N_E$. The discrepancy comes from the fact that rotations that mix orbitals of the same class do not affect the energy. In mathematical terms, this can be formulated as follows: the DM manifold $\mathcal{M}_{\mathrm{DM}}$ can be identified with the quotient of the MO manifold $O(\mathcal{N}_b)$ by the group

$$O(N_I) \times O(N_A) \times O(N_E).$$

This identification has very practical consequences on the design of direct optimization algorithms, as will be seen in Section 1.4.

The geometric structure described above corresponds to a well known fact in quantum chemistry. In direct optimization implementations, changes in the orbitals are parameterized via a rotation matrix

$$U = e^{\kappa}, \tag{1.2.12}$$

where $\kappa \in \mathbb{R}^{\mathcal{N}_{\mathrm{b}} \times \mathcal{N}_{\mathrm{b}}}$ is a skew-symmetric matrix with the following block structure:

$$\kappa = \begin{pmatrix} 0 & \kappa_{IA} & \kappa_{IE} \\ -\kappa_{IA}^T & 0 & \kappa_{AE} \\ -\kappa_{IE}^T & -\kappa_{AE}^T & 0 \end{pmatrix} \tag{1.2.13}$$

The vanishing diagonal blocks, that would mix orbitals belonging to the same class, are the practical translation of the quotient process mentioned above. We further note that the map $\kappa \mapsto Ce^{\kappa}$, with κ as in Eq. 1.2.13 provides a non-redundant local parametrization of the quotient manifold

$$O(\mathcal{N}_b)/(O(N_I) \times O(N_A) \times O(N_E)).$$

In standard quantum chemistry direct optimization implementations, a sequence of MO coefficients $\{C^{(k)}\}_{k=0}^{N_{it}}$ is generated starting from an initial guess $C^{(0)}$ such that the sequence of energies $\{E^{(k)}\}_{k=0}^{N_{it}}$ is

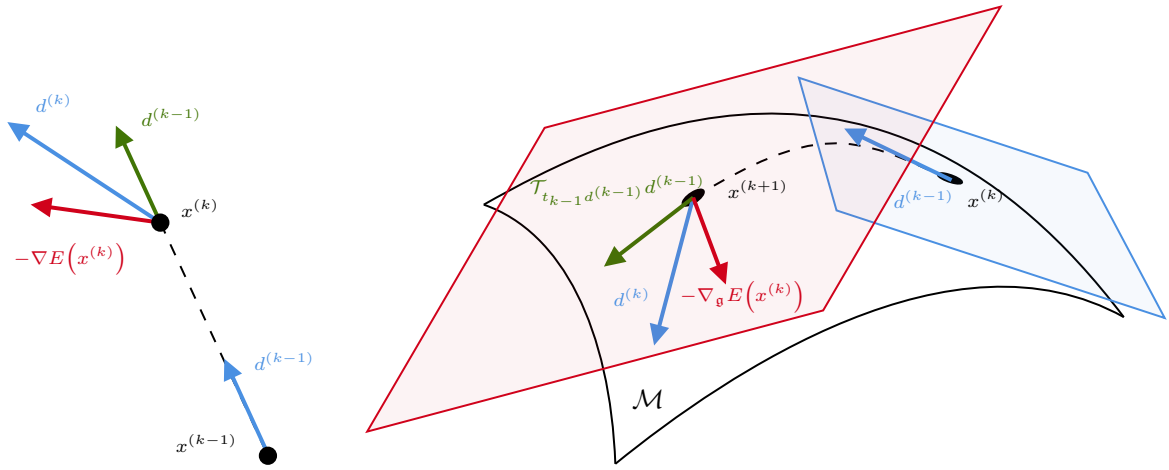


Figure 1.1 – Conjugate gradient algorithm in $\mathbb{R}^n$ (left). Riemannian conjugate gradient (RCG) algorithm on a Riemannian manifold $\mathcal{M}$ (right).

non increasing and hopefully converges to the ground state energy. The passage from $C^{(k)}$ to $C^{(k+1)}$ is obtained by

$$C^{(k+1)} = C^{(k)} e^{\kappa^{(k)}} \tag{1.2.14}$$

where $\kappa^{(k)}$ is the result of some optimization procedure (e.g., steepest descent, or Levenberg-Marquardt second order optimization). Eq. 1.2.14 amounts to changing the center of the local parameterization of the quotient manifold. For the steepest descent and Newton optimization methods, the calculation of $\kappa^{(k)}$ relies only on information relative to the point $C^{(k)}$, i.e., it makes no use of the history. Employing an optimization method that does, e.g., non-linear conjugate gradient (CG) or quasi-Newton methods, comes with a complication. Let us consider non-linear CG as an example. In the (flat) vector space $\mathbb{R}^n$, the CG descent direction at a given iterate is computed by linearly combining the gradient at the current iterate with the descent direct at the previous direction, see Fig. 1.1 (left) and Eq. 1.3.3. On a Riemannian manifold, the gradient and descent direction at a given iterate belong to the tangent space to the manifold at this particular iterate, which changes from iteration to iteration. Therefore, it is not possible to linearly combine tangent vectors at different points, as required by CG, in an obvious way. The operation of correctly transferring a vector quantity from the tangent space at a given point of the manifold to the tangent space at another point of the manifold is called *transport*, see Fig 1.1 (right) and Eq. 1.3.4.

For RHF and RKS, this problem is not apparent because the optimization takes place in a Grassmann manifold, for which the parallel transport map is trivial in the right parameterization (see Eq. 1.4.4). This is not the case for the flag manifolds on which ROHF and CASSCF optimization problems are set (see Eq. 1.4.8).

Given $(\Pi^{\mathrm{I}}, \Pi^{\mathrm{A}}) \in \mathcal{M}_{\mathrm{DM}}$, there exists a unique (up to an irrelevant global phase) normalized ROHF wavefunction $\Phi^{\mathrm{ROHF}}_{(\Pi^{\mathrm{I}},\Pi^{\mathrm{A}})}$ with maximal spin polarization $S = S_z = \frac{N_A}{2}$ associated with $(\Pi^{\mathrm{I}}, \Pi^{\mathrm{A}})$: it is the Slater determinant whose spin-1-RDM in the AO basis (χ_μ) is given by Eq. 1.2.4. The ROHF energy functional

$$E^{\mathrm{ROHF}}(\Pi^{\mathrm{I}}, \Pi^{\mathrm{A}}) := \langle \Phi^{\mathrm{ROHF}}_{(\Pi^{\mathrm{I}},\Pi^{\mathrm{A}})} | \widehat{H}_N | \Phi^{\mathrm{ROHF}}_{(\Pi^{\mathrm{I}},\Pi^{\mathrm{A}})} \rangle \tag{1.2.15}$$

is therefore a well-defined function of $(\Pi^{\mathrm{I}}, \Pi^{\mathrm{A}})$, and in fact a quadratic function in Π^{I} and Π^{A}. From a geometrical point of view, the ROHF problem is therefore a smooth optimization problem on a flag manifold, for which the energy is quadratic in the density-matrix formalism.

CASSCF can be also seen as an optimization problem on a flag manifold. In the spin-collinear approximation, the corresponding CASSCF energy functional can be written as

$$E^{\mathrm{CAS}}(\Pi^{\mathrm{I}}, \Pi^{\mathrm{A}}) = \min_{\Psi \in \mathcal{W}^{\mathrm{CAS}}_{\Pi^{\mathrm{I}},\Pi^{\mathrm{A}}}} \langle \Psi | \widehat{H}_N | \Psi \rangle,$$

with

$$\mathcal{W}^{\rm CAS}_{\Pi^{\rm I},\Pi^{\rm A}} := \left\{\Psi \text{ s.t. } \|\Psi\| = 1,\ \Pi^{\rm I} \le P^{\Psi,\sigma} \le \Pi^{\rm I} + \Pi^{\rm A},\ \mathrm{tr}\big(P^{\Psi,\sigma}\big) = N_I + m_\sigma,\ \sigma = \alpha, \beta\right\}.$$

Recall that for $A, B \in \mathbb{R}^{\mathcal{N}_{\rm b}\times\mathcal{N}_{\rm b}}_{\rm sym}$, $A \le B$ means that $X^T A X \le X^T B X$ for all $X \in \mathbb{R}^{\mathcal{N}_{\rm b}}$.

A very appealing feature of quotient manifolds is that if closed form expressions for parallel transport and geodesics on $\mathcal{M}$ are known, then closed form expressions for parallel transports on $\mathcal{M}/G$ can be derived from the ones on $\mathcal{M}$. The manifold $O(\mathcal{N}_{\rm b})$ is in fact a Lie group. For this reason, closed form expressions for parallel transport and geodesics on $O(\mathcal{N}_{\rm b})$ can be constructed from the exponential map.

1.3 Optimization on Riemannian manifolds

Riemannian optimization (i.e. optimization on manifold endowed with a Riemannian metric) is a major field of computational mathematics with many applications in various areas of science and technology. Several Riemannian optimization libraries are available, in which the most common Riemannian optimization methods are implemented. One of the advantages of using an optimization library is that this allows one to test and compare many different optimization methods with limited development effort. For optimization in the flat space $\mathbb{R}^n$, the user of an optimization library is just asked to provide the code returning the value of the function and its gradient at an input point $x \in \mathbb{R}^n$ (and possibly also, for some methods, a preconditioner and/or the Hessian at x). For optimization on a Riemannian manifold $\mathcal{M}$, the user is asked to provide four pieces of codes returning respectively:

1. the value of the function and its Riemannian gradient at an input point $x \in \mathcal{M}$, (and possibly also, for some methods, a preconditioner and/or the Riemannian Hessian at x);
2. the value of $\mathfrak{g}_x(p_x, q_x) \in \mathbb{R}$, where $\mathfrak{g}_x$ is the Riemannian metric, for an input point $x \in \mathcal{M}$ and two tangent vectors $p_x, q_x \in T_x\mathcal{M}$ at point x;
3. the value of $\mathcal{R}_x(p_x) \in \mathcal{M}$, where $\mathcal{R}$ is the chosen retraction, for an input point $x \in \mathcal{M}$ and a tangent vector $p_x \in T_x\mathcal{M}$ at point x;
4. the value of $\mathcal{T}_{p_x} q_x \in T_{\mathcal{R}_x(p_x)}\mathcal{M}$, where $\mathcal{T}$ is the chosen transport, for an input point $x \in \mathcal{M}$ and two tangent vectors $p_x, q_x \in T_x\mathcal{M}$ at point x.

Let us first recall the role of $\mathfrak{g}$, $\mathcal{R}$ and $\mathcal{T}$ in Riemannian optimization algorithms, and illustrate these concepts on the simple example of optimization on the orthogonal group $O(\mathcal{N}_b)$. It is well-known that the tangent space to $O(\mathcal{N}_b)$ at some $C \in O(\mathcal{N}_b)$ is given by

$$T_C O(\mathcal{N}_b) = \{CA,\ A \in \mathbb{R}^{\mathcal{N}_b\times\mathcal{N}_b}_{\rm antisym}\},$$

where $A \in \mathbb{R}^{\mathcal{N}_b\times\mathcal{N}_b}_{\rm antisym}$ is the vector space of $\mathcal{N}_b \times \mathcal{N}_b$ real antisymmetric matrix. The Frobenius inner product

$$\langle M, N\rangle_{\rm F} := \mathrm{tr}\big(M^T N\big) = \sum_{\mu,\nu=1}^{\mathcal{N}_b} M_{\mu\nu} N_{\mu\nu}$$

on $\mathbb{R}^{\mathcal{N}_b\times\mathcal{N}_b}$ induces a Riemannian metric on $O(\mathcal{N}_b)$ defined by

$$\mathfrak{g}_C(CA, CA') = \mathrm{tr}\big(A^T A'\big) = -\,\mathrm{tr}(AA') \quad \text{for all } C \in O(\mathcal{N}_b) \text{ and } CA, CA' \in T_C O(\mathcal{N}_b).$$

Consider a smooth function $E : \mathbb{R}^{\mathcal{N}_b\times\mathcal{N}_b} \to \mathbb{R}$. The gradient of E at some point $C \in \mathbb{R}^{\mathcal{N}_b\times\mathcal{N}_b}$ is the unique matrix $\nabla E(C) \in \mathbb{R}^{\mathcal{N}_b\times\mathcal{N}_b}$ such that

$$E(C + \delta C) = E(C) + \langle \nabla E(C), \delta C\rangle_{\rm F} + o(\delta C).$$

If $C \in O(\mathcal{N}_b)$, the Riemannian gradient of E at C is the matrix $\nabla_{\mathfrak{g}} E(C) \in T_C O(\mathcal{N}_b)$ obtained by orthogonally projecting $\nabla E(C)$ on $T_C O(\mathcal{N}_b)$ for the Frobenius inner product. Its expression is given by

$$\nabla_{\mathfrak{g}} E(C) = \frac{1}{2} C\left(C^T \nabla E(C) - \nabla E(C)^T C\right).$$

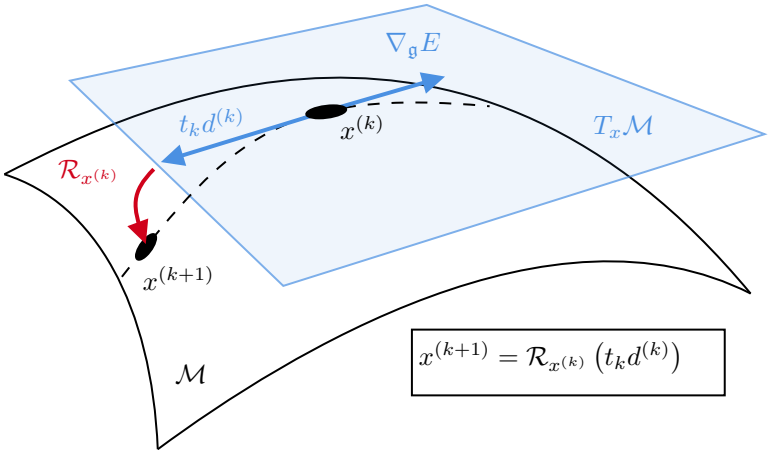


Figure 1.2 – Riemannian steepest descent (RSD) algorithm.

A retraction $\mathcal{R}$ on a manifold $\mathcal{M}$ is a map $\mathcal{R} : T\mathcal{M} \to \mathcal{M}$ such that for all $x \in \mathcal{M}$ the restriction $\mathcal{R}_x : T_x\mathcal{M} \to \mathcal{M}$ of $\mathcal{R}$ to $T_x\mathcal{M}$ satisfies for $p_x \in T_x\mathcal{M}$

$$\mathcal{R}_x(p_x) = x + p_x + o(p_x). \tag{1.3.1}$$

Among other things (see below), retractions are used to map straight lines, or more generally paths, drawn on the vector space $T_x\mathcal{M}$ onto paths drawn on the curved manifold $\mathcal{M}$. As a matter of example, the iterates of the fixed-step gradient (also called steepest descent) algorithm are defined by

$$d^{(k)} = -\nabla_{\mathfrak{g}} E(x^{(k)}), \quad x^{(k+1)} = \mathcal{R}_{x^{(k)}}(td^{(k)}), \qquad \text{(fixed-step steepest descent)} \tag{1.3.2}$$

for a chosen fixed step $t > 0$. In words, starting from a point $x^{(k)} \in \mathcal{M}$, the descent direction $d^{(k)}$ is chosen equal to the opposite of the gradient, which is the steepest descent direction for infinitesimal length steps, a step $td^{(k)}$ is made in this direction, and finally, the vector $td^{(k)} \in T_{x^{(k)}}\mathcal{M}$ is mapped back to a point of the manifold thanks to the retraction (see Fig. 1.2).

It follows from (1.3.1) that in the limit of small step lengths, we have

$$x^{(k+1)} = \mathcal{R}_{x^{(k)}}(td^{(k)}) = x^{(k)} + td^{(k)} + o(t)$$

where the remainder term $o(t)$ can be interpreted as a correction due to the curvature of the manifold $\mathcal{M}$.

Among all possible retractions on a Riemannian manifold $\mathcal{M}$, one is canonical: it is the one defined from the geodesics, called the exponential map, and denoted by Exp. In the case of $O(\mathcal{N}_b)$, the exponential map has a simple closed expression and is related to the usual exponential of matrices:

$$\mathrm{Exp}_C(CA) = Ce^A, \quad A \in \mathbb{R}^{\mathcal{N}_b \times \mathcal{N}_b}_{\text{antisym}} \quad \text{(exponential map on } O(\mathcal{N}_b)).$$

In Riemannian optimization, vector transports are used in particular to combine the descent directions of previous iterates. Let us further elaborate on this point, that was qualitatively discussed in Section 1.2. In the standard conjugate gradient algorithm in $\mathbb{R}^n$, the descent direction $d^{(k)}$ at iterate $x^{(k)}$ is a linear combination of $-\nabla E(x^{(k)})$ and $d^{(k-1)}$, the previous descent direction:

$$\left| \begin{array}{l} d^{(k)} = -g^{(k)} + \beta_k d^{(k-1)}, \\ x^{(k+1)} = x^{(k)} + t_k d^{(k)}, \end{array} \right. \qquad \text{(nonlinear CG algorithm in } \mathbb{R}^n) \tag{1.3.3}$$

with $g^{(k)} = \nabla E(x^{(k)})$ and either

$$\beta_k^{\mathrm{FR}} := \frac{\|g^{(k)}\|^2}{\|g^{(k-1)}\|^2} \qquad \text{(Fletcher-Reeves)},$$

$$\beta_k^{\mathrm{PR}} := \frac{{g^{(k)}}^T (g^{(k)} - g^{(k-1)})}{\|g^{(k-1)}\|^2} \qquad \text{(Polak-Ribière)},$$

$$\beta_k^{\mathrm{HS}} := \frac{{g^{(k)}}^T (g^{(k)} - g^{(k-1)})}{(g^{(k)} - g^{(k-1)})^T d^{(k)}} \qquad \text{(Hestenes-Stiefel)}.$$

The step length $t_k \in \mathbb{R}$ is obtained by a line search technique such as Armijo, Wolfe, or Hager-Zhang [HZ06] linesearch.

This idea cannot be directly used in optimization on manifolds, because $-\nabla E(x^{(k)})$ and $d^{(k-1)}$ belong to different vector spaces, namely $T_{x^{(k)}}\mathcal{M}$ and $T_{x^{(k-1)}}\mathcal{M}$ respectively. Before being combined with $-\nabla E(x^{(k)})$ to form the new descent direction $d^{(k)}$, the vector $d^{(k-1)}$ must be transported from $T_{x^{(k-1)}}\mathcal{M}$ to $T_{x^{(k)}}\mathcal{M}$, using a transport map $\mathcal{T}$. A transport map takes as input two vectors p_x, q_x of the tangent space $T_x\mathcal{M}$ at point x, and returns a vector $\mathcal{T}_{p_x} q_x$ of the tangent space at point $\mathcal{R}_x(p_x)$ (compatibility condition with the retraction $\mathcal{R}$). The map $(p_x, q_x) \mapsto \mathcal{T}_{p_x} q_x$ is linear in the variable q_x, and satisfies the consistency relation $\mathcal{T}_0 q_x = q_x$. We thus have

$$\left| \begin{array}{l} d^{(k)} = -g^{(k)} + \beta_k \mathcal{T}_{t_{k-1} d^{(k-1)}} d^{(k-1)}, \\ x^{(k+1)} = \mathcal{R}_{x^{(k)}}(t_k d^{(k)}), \end{array} \right. \qquad \text{(Riemannian CG algorithm)} \tag{1.3.4}$$

with either

$$\beta_k^{\mathrm{RFR}} := \frac{\mathfrak{g}_{x^{(k)}}(g^{(k)}, g^{(k)})}{\mathfrak{g}_{x^{(k-1)}}(g^{(k-1)}, g^{(k-1)})} \qquad \text{(Riemannian Fletcher-Reeves)},$$

$$\beta_k^{\mathrm{RPR}} := \frac{\mathfrak{g}_{x^{(k)}}(g^{(k)}, g^{(k)} - \mathcal{T}_{t_{k-1} d^{(k-1)}}(g^{(k-1)}))}{\mathfrak{g}_{x^{(k)}}(g^{(k-1)}, g^{(k-1)})} \qquad \text{(Riemannian Polak-Ribière)},$$

$$\beta_k^{\mathrm{RHS}} := \frac{\mathfrak{g}_{x^{(k)}}(g^{(k)}, g^{(k)} - \mathcal{T}_{t_{k-1} d^{(k-1)}}(g^{(k-1)}))}{\mathfrak{g}_{x^{(k)}}(g^{(k)}, \mathcal{T}_{t_{k-1} d^{(k-1)}}(d^{(k-1)})) - g_{x^{(k-1)}}(g^{(k-1)}, d^{(k-1)})} \qquad \text{(Riemannian Hestenes-Stiefel)}.$$

Together with $x^{(k)} = \mathcal{R}_{x^{(k-1)}}(t_{k-1} d^{(k-1)})$, the fact that $\mathcal{T}_{p_x} q_x \in T_{\mathcal{R}_x(p_x)}\mathcal{M}$ (comptabilitly with the retraction) ensures that $\mathcal{T}_{t_{k-1} d_{(k-1)}} d^{(k-1)}$ belongs to $T_{x^{(k)}}\mathcal{M}$.

Among all transports compatible with the exponential map Exp associated with the metric $\mathfrak{g}$, one is canonical: it is the parallel transport associated with the Levi-Civita connection of the metric $\mathfrak{g}$. For the example of $O(\mathcal{N}_b)$, this parallel transport has an extremely simple form

$$\mathcal{T}_{CA}(CB) = Ce^A B \quad \text{(parallel transport on } O(\mathcal{N}_b)).$$

1.4 Optimization on Grassmann and flag manifolds

In RHF and RKS models, the state of the system is described by a point of the Grassmann manifold

$$\mathrm{Gr}(N, \mathcal{N}_b) \cong \underbrace{\left\{ P \in \mathbb{R}^{\mathcal{N}_b \times \mathcal{N}_b}_{\mathrm{sym}} \text{ s.t. } P^2 = P,\ \mathrm{tr}(P) = N \right\}}_{\text{DM formalism}} \cong \underbrace{O(\mathcal{N}_b)/(O(N) \times O(\mathcal{N}_b - N))}_{\text{MO formalism}}.$$

In the DM formalism, the Grassmann manifold is parameterized by the matrix P of the orthogonal projector on the vector space spanned by the doubly-occupied MO. In the MO formalism, it is represented by an orthogonal matrix $C \in O(\mathcal{N}_b)$, the first N columns of C corresponding to the N doubly-occupied orbitals, and the last $\mathcal{N}_b - N$ ones to the virtual orbitals. The gauge invariance in the MO formulation is taken into account by quotienting $O(\mathcal{N}_b)$ by the group $O(N) \times O(\mathcal{N}_b - N)$ (rotations of occupied / virtual orbitals).

Likewise, in the ROHF model and the outer CASSCF minimization problem, the state is represented by a point of the flag manifold

$$\mathrm{Flag}(N_I, N_I + N_A, \mathcal{N}_b) \cong \mathcal{M}_{\mathrm{DM}} \cong \underbrace{O(\mathcal{N}_b)/(O(N_I) \times O(N_A) \times O(N_E))}_{\text{MO formalism}}.$$

In both cases, the MO formalism involves the quotient of the orthogonal group $O(\mathcal{N}_b)$ by a closed subgroup ($O(N) \times O(\mathcal{N}_b - N)$ for RHF/RKS, $O(N) \times O(N_I) \times O(N_A) \times O(N_E)$ for ROHF/CASSCF). As a consequence, the closed form expressions for the canonical retraction and parallel transport on $O(\mathcal{N}_b)$ can be translated into closed form expressions for canonical retraction and parallel transport on the quotient manifold [**Ye2022**; EAS98; AMS08; Bou23], leading to the following formulae:

- RHF/RKS setting:

$$\text{tangent space at } C \quad \cong \left\{ C\kappa, \quad \kappa = \begin{pmatrix} 0 & \kappa_{OV} \\ -\kappa_{OV}^T & 0 \end{pmatrix} \right\}, \tag{1.4.1}$$

$$\text{metric:} \qquad \mathfrak{g}_C(C\kappa, C\kappa') = \mathrm{tr}\big(\kappa^T \kappa'\big) = 2\,\mathrm{tr}\big(\kappa_{OV}^T \kappa'_{OV}\big), \tag{1.4.2}$$

$$\text{exponential map (canonical retraction):} \qquad \mathcal{R}_C(C\kappa) = Ce^{\kappa}, \tag{1.4.3}$$

$$\text{parallel transport:} \qquad \mathcal{T}_{C\kappa}(C\kappa') = Ce^{\kappa}\kappa', \tag{1.4.4}$$

- ROHF/CASSCF setting:

$$\text{tangent space at } C \quad \cong \left\{ C\kappa, \quad \kappa = \begin{pmatrix} 0 & \kappa_{IA} & \kappa_{IE} \\ -\kappa_{IA}^T & 0 & \kappa_{AE} \\ -\kappa_{IE}^T & -\kappa_{AE}^T & 0 \end{pmatrix} \right\}, \tag{1.4.5}$$

$$\text{metric:} \qquad \mathfrak{g}_C(C\kappa, C\kappa') = \mathrm{tr}\big(\kappa^T \kappa'\big) \tag{1.4.6}$$

$$\text{exponential map (canonical retraction):} \qquad \mathcal{R}_C(C\kappa) = Ce^{\kappa}, \tag{1.4.7}$$

$$\text{parallel transport:} \qquad \mathcal{T}_{C\kappa}(C\kappa') = Ce^{\kappa}e^{-\phi_\kappa}(\kappa'), \tag{1.4.8}$$

where $\phi_\kappa : \mathfrak{K} \to \mathfrak{K}$ is the linear operator on the vector space

$$\mathfrak{K} = \left\{ \kappa = \begin{pmatrix} 0 & \kappa_{IA} & \kappa_{IE} \\ -\kappa_{IA}^T & 0 & \kappa_{AE} \\ -\kappa_{IE}^T & -\kappa_{AE}^T & 0 \end{pmatrix} \right\}$$

defined by

$$\begin{aligned} \phi_\kappa(\kappa') &= \frac{1}{2}\mathrm{Proj}_{\mathfrak{K}}\left([\kappa, \kappa']\right) \\ &= \frac{1}{2}\begin{pmatrix} 0 & -\kappa_{IE}[\kappa'_{AE}]^T + \kappa'_{IE}\kappa_{AE}^T & \kappa_{IA}\kappa'_{AE} - \kappa'_{IA}\kappa_{AE} \\ \kappa'_{AE}\kappa_{IE}^T - \kappa_{AE}[\kappa'_{IE}]^T & 0 & -\kappa_{IA}^T\kappa'_{IE} + [\kappa'_{IA}]^T\kappa_{IE} \\ -[\kappa'_{AE}]^T\kappa_{IA}^T + \kappa_{AE}^T[\kappa'_{IA}]^T & [\kappa'_{IE}]^T\kappa_{IA} - \kappa_{IE}^T\kappa'_{IA} & 0 \end{pmatrix}, \end{aligned}$$

and

$$e^{-\phi_\kappa} = \sum_{n=0}^{+\infty} \frac{(-1)^n}{n!} \underbrace{(\phi_\kappa \circ \cdots \circ \phi_\kappa)}_{n \text{ times}}. \tag{1.4.9}$$

In computational codes, it is convenient to represent a tangent vector $C\kappa$ by the block $\kappa_{OV} \in \mathbb{R}^{N \times (\mathcal{N}_b - N)}$ for RHS/RKS, and the blocks $(\kappa_{IA}, \kappa_{IE}, \kappa_{AE}) \in \mathbb{R}^{N_I \times N_A} \times \mathbb{R}^{N_I \times N_E}\mathbb{R}^{N_A \times N_E}$ for ROHF/CASSCF. It follows from Eq. 1.4.4, that in this representation the parallel transport for RHF/RKS is the identity operator. This is not the case for ROHF/CASSCF where the transport of the block matrix κ' is done by the map $e^{-\phi_\kappa}(\kappa')$, which transforms and mixes the IA/IE/AE blocks of κ'. Let us note however that in the special case when the transported vector $C\kappa'$ is collinear to the vector $C\kappa$ along which it is transported, then the transport formula Eq. 1.4.8 dramatically simplifies. Indeed, we then have $[\kappa, \kappa'] = 0$, and therefore $e^{-\phi_\kappa}(\kappa') = \kappa'$. This occurs for the Riemannian conjugate gradient method (see Eq. 1.3.4), but not for quasi-Newton methods such as BFGS.

1.5 Numerical Results

In this section, we analyze the performance of Riemannian optimization algorithms for solving the ROHF and CASSCF minimization problems for a few selected test cases. Let us first provide some implementation details.

General implementation. We focus specifically on Riemannian steepest descent (RSD), nonlinear conjugate gradient (RCG) and low-memory Broyden-Fletcher-Goldfarb-Shanno (R-LBFGS) methods, all endowed with preconditioning. We refer to [AMS08; Bou23] for general introductions to Riemannian optimization methods. Our code is structured as follows: first, we implemented the RSD, RCG, and

R-LBFGS optimization routines in the MO formalism within a Julia [Bez+17] package which is then interfaced with PySCF [Sun+20] for ROHF and CFOUR [Mat+20] for CASSCF calculations. These software handle the operations specific to the ROHF and CASSCF models, including the generation of AO basis sets and initial guess MOs, the computation of electronic integrals, and the evaluation of energies and Frobenius gradients.

In our Julia package, we use for all methods the exponential retraction (1.4.7) and parallel transport (1.4.8) as outlined in the previous section. For parallel transport, the exponential operator is computed by truncating the series (1.4.9) so that the Frobenius norm of the last term falls below numerical precision.

Our implementation of RCG is based on *Algorithm 1* in Boumal et al. [BA15] with Polak-Ribière coefficient β^{RPR} as above. For R-LBFGS, we implemented *Algorithm 2* in Huang et al [HGA15]. For CASSCF, we use the inverse diagonal of the Hessian as preconditioner. In the case of ROHF, we tested two different preconditioners. The first one is the modified inverse diagonal Hessian as discussed in [NGL21b]. The other is detailed in Appendix. Our results are presented for the second choice of preconditioner that showcased the best performance for our test case.

All methods use Hager-Zhang [HZ06] linesearch as implemented in the LineSearches.jl [MR18] Julia package. Computations are considered to have reached convergence when the Frobenius norm of the Riemannian gradient reaches 10^{-5}. Comprehensive implementation details are available in our publicly accessible GitHub repository.[1]

Details specific to the R-LBFGS implementation. At each iteration, the R-LBFGS method constructs an approximation B of the inverse Hessian using a certain number m of vectors stored in memory from previous iterations, through an iterative procedure [HGA15]. In addition to preconditioning, it is important to note that the performance of R-LBFGS is influenced by the selection of the maximum depth $m_{\max}$, the initial guess B_0 for the approximate inverse Hessian in the iterative process and the choice of restart strategies, which determines the iterations at which the history is reset. For both ROHF and CASSCF, we define $B_0 = \gamma\,\mathrm{Id}$ with γ as in [HGA15]. If P is the preconditioner, the preconditioned version of R-LBFGS is obtained by replacing B_0 with $\hat{\gamma}P$, with $\hat{\gamma}$ as in [DSH18].

We experimented two restart strategies which depend on the preconditioning. For the first one, called dynamic R-LBFGS, the diagonal Hessian used for preconditioning is updated at each iteration. The history is reset whenever the direction obtained from the R-LBFGS quasi-newton system is not a descent direction. For the second method called fixed R-LBFGS, we use the same preconditioner $P = \mathrm{diag}(\mathrm{Hess}^{(0)})^{-1}$, corresponding to the inverse diagonal Hessian for the guess orbitals, at each iteration until the inverse diagonal Hessian at current point, $\mathrm{diag}(\mathrm{Hess}^{(k)})^{-1}$, deviates too much from P. When this happens, the history is reset, and the procedure is reinitialized with $P = \mathrm{diag}(\mathrm{Hess}^{(k)})^{-1}$. When using the preconditioner described in appendix for ROHF, we applied the dynamic strategy.

1.5.1 ROHF

For ROHF we tested the three aforementioned Riemannian optimization methods on Ti_2O_4 in its D_{2h} geometry, using Dunning's *cc*-pVTZ basis set [Dun89; KDH92]. This system is employed as a template for addressing SCF convergence issues[2] in the Amsterdam Density Functional (ADF) quantum-chemistry package [TV+01]. In order to compare the performance of Riemannian algorithms in different convergence regimes, calculations were started from both a core initial guess (Fig. 1.3) and a guess closer to a minimum (Fig. 1.4). The second guess is obtained by a standard SCF+DIIS method for ROHF, with Guest and Saunders coefficients [PD14], stopped when the Frobenius norm of the gradient reaches 10^{-1}.

In both cases, all three methods provide stable convergence toward a local minimum of the energy. In the optimal scenario, a finely tuned SCF+DIIS method outperforms the Riemannian optimization methods we have tested. However, the performance and stability of SCF routines for ROHF are notoriously sensitive to the choice of method and acceleration parameters, as illustrated in Fig. 1.5. On the other hand, the direct minimization methods described in this paper have the advantage of offering robust convergence, which is a valuable feature in terms of user's time and effort.

We were unfortunately not able to make a direct comparison to the GDM algorithm [DVHG02] due to our lack of access to the code or to the fine details of the implementation. Nevertheless, a simple-minded

[1] https://github.com/LaurentVidal95/ROHFToolkit

[2] See https://www.scm.com/doc/ADF/Examples/SCF_Ti2O4.html

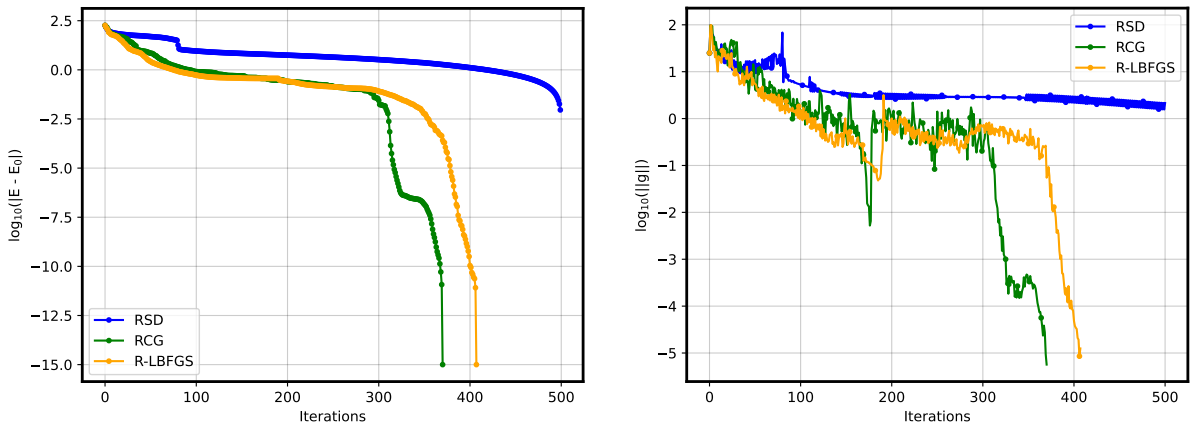


Figure 1.3 – Calculations from core initial guess. On the left, energy difference with respect to the converged energy along the iterations. On the right, Frobenius norm of the Riemannian gradient along the iterations. Only the first 500 iterations of RSD are shown on the graph for readability.

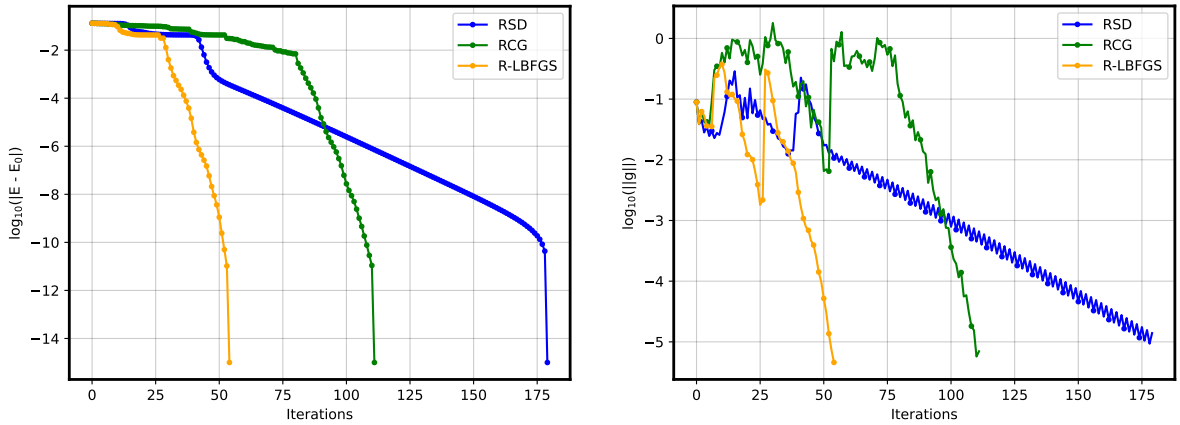


Figure 1.4 – Calculations from a better initial guess, 1 E_h from the expected energy. On the left, energy difference with respect to the converged energy along the iterations. On the right, Frobenius norm of the Riemannian gradient along the iterations.

test performed with the free trial version of Q-Chem [Sha+15] using default parameters showed that GDM and GDM+DIIS exhibit similar performances to our RCG implementation on Ti_2O_4.

An important point that needs to be discussed here is what minimum the various algorithms converged to. Our proof-of-concept implementation does not enforce point-group symmetry, so our calculations were performed in the C_1 group.

Our calculations converged to two different minima, one at -1996.191285 E_h, which was systematically obtained when starting from the core guess, and one at -1996.179398 E_h, obtained when starting from the better guess. We also looked at the lowest triplets for each Irrep enforcing symmetry using a quadratically convergent ROHF implementation [NGL21b], and we determined that the lowest triplet is the B_{2u} state at -1996.142005 E_h. The stability analysis of such solution reveals however that a lower energy, symmetry-broken solution exists. We therefore conclude that the two solutions found with our Riemannian optimization algorithms are two lower-energy symmetry-broken solutions. Whether a lower-energy, symmetry broken-solution is desirable or not depends on the aims of the study, and ultimately on the user; however, we note that our algorithms can be generalized to enforce point group symmetry, which we plan to do in the future.

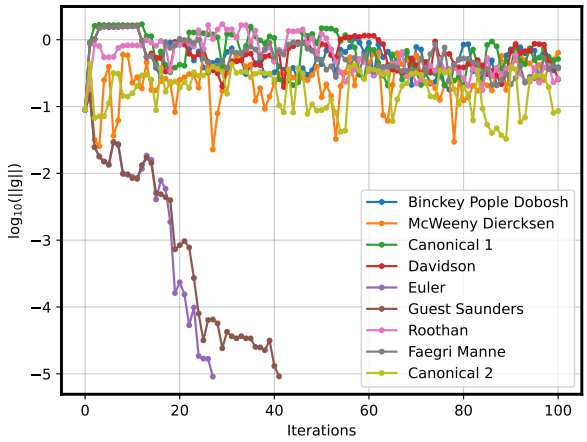


Figure 1.5 – Comparison of standard SCF+DIIS methods for ROHF as listed in [PD14], for Ti_2O_4 in *cc*-pVTZ basis set, starting from our good initial guess about 1 Hartree away from the expected energy. The vertical axis shows the Frobenius norm of the Riemannian gradient along the iterations. Only two methods yield convergence.

1.5.2 CASSCF

The CASSCF method has been tested by running calculations on a subgroup of the benchmark set used in [MKW16] and [NGL21a] using Pople's 6-31G* basis set [HDP72]. Convergence properties of direct minimization algorithms were compared against two well-established CASSCF optimization algorithms namely Super CI (SCI) [Roo80; Sie+81; MRR90; Kol+19; Ang+02] and the norm-extended optimization (NEO) [JJ84; JA86], the latter one being a genuine second-order algorithm. All computations were carried out using two different choices, to simulate, respectively, a troublesome scenario where the calculation starts relatively far away from the converged result, and an ideal starting point that should be close to the final minimum. For the former scenario, we use canonical restricted Hartree-Fock (RHF) orbitals, while as a good starting point we exploit unrestricted natural orbitals (UNO) [PH88; TP20]

We report in Tab. 1.1 for each algorithm the average number of iterations required to converge CASSCF. As expected, the values related to the RHF guess are systematically higher than the ones related to the UNO guess. Moreover, we notice that the numbers related to the RHF guess show a high variability as indicated by the large standard deviation, thus being strongly system dependent. The average number of iterations for all direct minimization methods with the exception of RSD is comparable with and in some cases outperforms the ones of SCI. We conclude this section by looking more in detail

Algorithm	$\langle \text{It.} \rangle^{\text{RHF}}$	σ^{RHF}	$\langle \text{It.} \rangle^{\text{UNO}}$	σ^{UNO}
RSD	115.9	112.9	32.3	12.5
RCG	31.9	11.6	15.4	3.0
R-LBFGS(dyn)	37.9	11.3	19.2	3.5
R-LBFGS(fix)	35.6	15.2	19.1	3.7
SCI(DIIS)	38.4	26.0	14.9	6.2
SCI(no DIIS)	61.5	25.1	21.6	9.9
NEO	12.2	1.8	5.1	0.5

Table 1.1 – Average number of iterations ($\langle \text{It.} \rangle$) and standard deviation (σ) for each tested algorithm starting with two different guess orbitals, namely restricted Hartree-Fock (RHF) and unrestricted natural orbitals (UNO).

at one example, namely, pyridine using a standard CAS(6,6) wavefunction. In Fig. 1.6, we compare the convergence behavior of the R-LBFGS as implemented in the present study with a naive implementation that simply translates the gradient from previous points, without parallel transport. In both cases, we start from canonical orbitals. We note that while both implementations get stuck for a while on a plateau, the R-LBFGS overcomes it in a few iterations and then converges smoothly. On the contrary, the

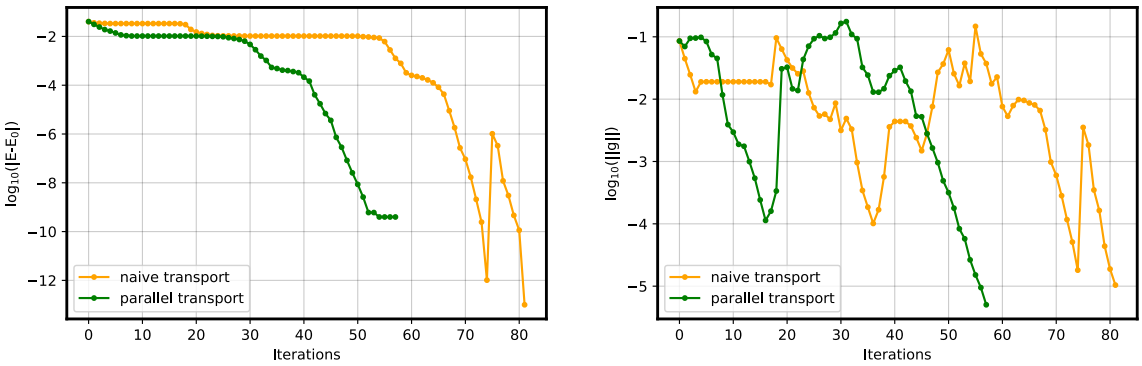


Figure 1.6 – On the left, energy difference with respect to the converged energy along the iterations for a proper R-LBFGS implementation that makes use of the parallel transport (green curve) and a more basic one (orange curve). On the right, Frobenius norm of the gradient along the iterations.

naive LBFGS implementation exhibits worse convergence, with very small and even positive slope steps at the beginning of the optimization. This demonstrates the importance of properly accounting for parallel transport.

Another important point concerns the actual solution obtained. When starting from a poor guess such as the one given by canonical orbitals, without any kind of manual selection for the active space, the optimizer may get stuck in local minima that are ultimately related by orbitals swapping between the inactive and active domains. Using once again pyridine as a test case, we observe that all the algorithms converged to the same local minimum (-246.766857 E_h) with the exception of SCI that converged to a higher minimum (-246.756489 E_h). These two minima are characterized by different converged active orbitals. This difference can be assessed simply by visual inspection or by checking the singular values of the difference between the active part of the one-body density matrix in the AO basis for the two calculations. If the converged active orbitals were (almost) the same, we would observe (almost) vanishing singular values. On the contrary, as depicted in Fig. 1.7, two orbitals are completely different between the two results.

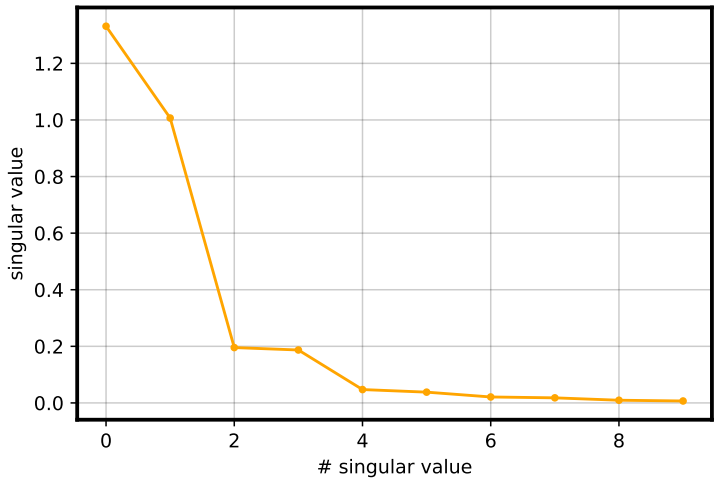


Figure 1.7 – First ten singular values of the difference matrix between the active AO-based one-body density matrices stemming from two calculations that reached different local minima.

We also note that by manually tuning the DIIS parameters, we managed to achieve convergence to the lower-energy solution using SCI. Again, this underlines the robustness of the Riemannian optimization algorithms, which is valuable in terms of user time and effort.

1.6 Conclusions and perspectives

In this contribution, we have explored the use of Riemannian optimization methods on the flag manifold to optimize restricted-open and complete active space self-consistent field wavefunctions. After discussing the geometry of the problem, we have reviewed the general aspects of Riemannian optimization and its application to the aforementioned chemical problem. We have then compared various algorithms to traditional ones. The Riemannian optimization methods illustrated in this work all show robust convergence properties, and do so without requiring the user to finely tune the parameters that control the optimization. Even in the naive implementation presented here, they demonstrate that they can be competitive with other traditional implementations in terms of number of iterations, and thus overall computational cost. Nevertheless, this is just a proof-of-concept study, for which several further developments are required. First, the overall underwhelming performance of the Riemannian Quasi-Newton L-BFGS method for CASSCF, which is expected to outperform conjugate gradient, as observed for ROHF when starting from a good guess (see Fig. 1.4), can be explained by the basic preconditioner and initialization of the inverse Hessian we use. Finding better preconditioners and inverse Hessian approximations is not a straightforward task, and requires further attention. We also have not investigated optimal parameters, including exploring different line-search and restarting strategies, which would require to run extensive numerical tests, but that can greatly improve the overall performance of the methods.

In conclusion, we believe that Riemannian optimization is a valuable addition to the SCF optimization toolbox for ROHF and CASSCF, and that further exploration of the use of such techniques is worthy of attention.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement EMC2 No. 810367). F.L. and T.N. acknowledge financial support from ICSC-Centro Nazionale di Ricerca in High Performance Computing, Big Data, and Quantum Computing, funded by the European Union – Next Generation EU – PNRR, Missione 4 Componente 2 Investimento 1.4.

Appendix: A variant of the diagonal Hessian preconditioner for ROHF

Let $C \in O(\mathcal{N}_b)$ and $\kappa \in \mathfrak{K}$ such that $C\kappa \in T_C O(\mathcal{N}_b)$. The ROHF Hessian applied to $C\kappa$ is given by

$$\mathcal{L}_C(C\kappa) = C \begin{pmatrix} 0 & X & Y \\ -X^T & 0 & Z \\ -Y^T & -Z^T & 0 \end{pmatrix} \tag{1.6.1}$$

where the matrices $X \in \mathbb{R}^{N_I \times N_A}$, $Y \in \mathbb{R}^{N_I \times N_E}$ and $Z \in \mathbb{R}^{N_A \times N_E}$ are defined by

$$\begin{aligned}
X &= 2(\kappa_{IA}(F_I - F_A)_{AA} - (F_I - F_A)_{II}\kappa_{IA}) + \kappa_{IE}(2F_I - F_A)_{EA} + (F_I - 2F_A)_{IE}\kappa_{AE}^T \\
&\qquad + (2J(\lambda_1) - K(\lambda_1))_{IA} + J(\lambda_2)_{IA} \\
Y &= \kappa_{IA}(4F_I - 2F_A)_{AE} + 4(\kappa_{IE}(F_I)_{EE} - (F_I)_{II}\kappa_{IE}) - 2((F_I)_{IA} + (F_A)_{IA})\kappa_{AE} \\
&\qquad + 4(2J(\lambda_1) - K(\lambda_1))_{IE} + 2(2J(\lambda_2) - K(\lambda_2))_{IE} \\
Z &= \kappa_{IA}^T(2F_I - F_A)_{IE} - 2(F_I + F_A)_{AI}\kappa_{IE} + 4(\kappa_{AE}(F_A)_{EE} - (F_A)_{AA}\kappa_{AE}) \\
&\qquad + 2(2J(\lambda_1) - K(\lambda_1))_{AE} + 2(2J(\lambda_2) - K(\lambda_2))_{AE}.
\end{aligned} \tag{1.6.2}$$

In the above expression, we adopted the following conventions: the operators J and K are the standard exchange and Coulomb operators, F_I and F_A are the internal and active Fock matrices, defined for all $\Pi_I = C_I C_I^T$ and $\Pi_A = C_A C_A^T$ by

$$\begin{aligned}
F_I &= h + 2J(\Pi_I) + J(\Pi_A) - K(\Pi_I) - \frac{1}{2}K(\Pi_A) \\
F_A &= \frac{1}{2}(h + 2J(\Pi_I) + J(\Pi_A) - K(\Pi_I) - K(\Pi_A))
\end{aligned}$$

and the matrices $\lambda_1, \lambda_2 \in \mathbb{R}^{\mathcal{N}_b \times \mathcal{N}_b}_{\text{sym}}$ are given by

$$\lambda_1 = \begin{pmatrix} 0 & \kappa_{IA} & \kappa_{IE} \\ \kappa_{IA}^T & 0 & 0 \\ \kappa_{IE}^T & 0 & 0 \end{pmatrix} \quad \text{and} \quad \lambda_2 = \begin{pmatrix} 0 & -\kappa_{IA} & 0 \\ -\kappa_{IA}^T & 0 & \kappa_{AE} \\ 0 & \kappa_{AE}^T & 0 \end{pmatrix}. \tag{1.6.3}$$

Each block X, Y and Z can be decomposed as the sum of two terms, the first one, denoted by $(\widetilde{X}, \widetilde{Y}, \widetilde{Z})$, being simple to compute in terms of internal and active Fock operators, and the second one, denoted by $(\Omega_X, \Omega_Y, \Omega_Z)$, being more costly to compute:

$$\begin{aligned} X &= 2(\kappa_{IA}(F_I - F_A)_{AA} - (F_I - F_A)_{II}\kappa_{IA}) + \Omega_X := \widetilde{X} + \Omega_X \\ Y &= 4(\kappa_{IE}(F_I)_{EE} - (F_I)_{II}\kappa_{IE}) + \Omega_Y := \widetilde{Y} + \Omega_Y \\ Z &= 4(\kappa_{AE}(F_A)_{EE} - (F_A)_{AA}\kappa_{AE}) + \Omega_Z := \widetilde{Z} + \Omega_Z \end{aligned} . \tag{1.6.4}$$

In our implementation, we define a preconditioned direction $C\kappa_{\text{prec}}$ as a solution to the linear system

$$\widetilde{L}_C(C\kappa_{\text{prec}}) = C\kappa. \tag{1.6.5}$$

involving the approximate Hessian

$$\widetilde{\mathcal{L}}_C(C\kappa) = C \begin{pmatrix} 0 & \widetilde{X} & \widetilde{Y} \\ -\widetilde{X}^T & 0 & \widetilde{Z} \\ -\widetilde{Y}^T & -\widetilde{Z}^T & 0 \end{pmatrix}. \tag{1.6.6}$$

The advantage of formulation (1.6.4) is that the lowest eigenvalue of $\widetilde{L}_C$ can be estimated with respect to F_I and F_A, which allows to apply a shift when $\widetilde{\mathcal{L}}$ is not positive definite (which is expected when starting far from a minimum). In addition, the system (1.6.5) reads as three Sylvester matrix equations, that can be solved using standard LAPACK optimized routines.

Chapter 2

Self-consistent Field algorithms in Restricted Open-Shell Hartree-Fock

This chapter resulted in the preprint [LVp2]:

Robert Benda, Eric Cancès, Emmanuel Giner, and Laurent Vidal. "Self-consistent field algorithms in Restricted Open-Shell Hatree-Fock". Submitted

Abstract In this chapter, we propose a simple geometrical derivation of the restricted open-shell Hartree-Fock (ROHF) equations in the density matrix and molecular orbitals formalism. We then introduce a new, parameter-free, basic fixed-point method to solve these equations, that, in contrast with existing self-consistent field (SCF) schemes, is not based on the introduction of a non-physical, parameter-dependent, composite Hamiltonian. We also extend the Optimal Damping Algorithm to the ROHF framework. We finally present numerical results on challenging systems (complexes with transition metals) demonstrating the performance of the new algorithms we propose.

Contents

2.1 Introduction

The ultimate goal of computational chemistry is to propose reliable theoretical tools to describe the chemical properties of any molecular system. The initial step of such a task is always the accurate description of the ground state electronic structure of the system, for which there exist essentially two flavors of approaches: the wave function theory (WFT) and density functional theory (DFT). Although DFT remains certainly the most used theoretical tool for closed-shell systems because of its advantageous ratio between the computational cost and the accuracy of the results, the usual semi-local approximations used in DFT are known to suffer from several issues when open-shell systems need to be considered. For instance, the self-interaction error in open-shell systems is responsible for the over delocalization of electrons in transition metal complexes and has impacts on several chemical properties such as the electronic paramagnetic spectrum, ligand-field excitations or spin-gaps [Rui+98; SMS02; ARS+05; KKN07; Ata+06]. One major issue in DFT is that there is no systematic way to improve the results, which leads to an inflation of different flavors of approximated functionals tailored for a specific class of systems and/or properties [VT20]. The situation of WFT is somehow opposite as there exists many ways of systematically refine the results starting from a mean-field description although it comes to the price of a rapidly growing computational cost. Nevertheless, as remarkable progresses have been obtained in the reduction of the computational cost of correlated WFT methods (see for instance Ref [MW20] and references therein), the latter appear more and more as actual computational tools for the treatment of open-shell systems. Even though WFT-based correlated methods are in active development, they all start with a mean-field Hartree Fock (HF) calculation for which there are many convergence problems in the context of open-shell systems. Therefore, improving the reliability of the HF algorithms becomes an important point in order to popularize the correlated WFT methods.

There exists several avatars of the Hartree-Fock method. The most commonly used are the restricted and unrestricted Hartree-Fock methods (RHF and UHF, respectively), which differ by the constraint imposed in the RHF method to have an unique set of spatial orbitals for both up and down spins. For open-shell systems, the constraint of having the same spatial orbitals for the two spins has an important consequence: while the ROHF Slater determinant is an eigenfunction of the $\hat{\mathcal{S}}^2$ operator, the UHF Slater determinant suffers from spin contamination [TS10]. The latter has a big impact in the post-HF calculations as the correlated wave function built upon a spin-contaminated Slater determinant needs to restore the correct spin symmetry using high-order particle-hole excitations [Boo+13; TS10; DVHG02]. Moreover, the correlated methods using unrestricted orbitals necessary deal with several types of two-electron integrals corresponding to the interaction between electrons of different spins, which also induces several complications in the code structure and memory.

From the mathematical point of view, Hartree-Fock methods give rise to constrained optimization problems, whose first-order optimality conditions are the Hartree-Fock equations. As usual in optimization theory, numerical solutions can be obtained either by solving the Hartree-Fock equations by a fixed-point (self-consistent field - SCF) algorithm, or by a direct minimization of the Hartree-Fock energy functional [DVHG02; VHG02; CKL21].

Many algorithms have been developed for the RHF and UHF frameworks in the past 70 years. Roothaan's [Roo51], level-shifting [SH73], and DIIS algorithms [Pul80; Pul82; HP86; RS11; Chu+21] belong to the class of SCF algorithms. Direct minimization approaches are adopted in *e.g.* Bacskay's quadradic convergent algorithm [Bac81], trust-region methods [Thø+04] and geometric direct minimization (GDM) methods [DVHG02; VHG02]. Let us also mention the second-order SCF (SOSCF) algorithm [CSG97; Nee00], and the DIIS-GDM [DVHG02; VHG02], which combine features from both SCF and direct minimization methods.The optimal damping algorithm (ODA) [CB00] and the EDIIS algorithm [KSC02] solve a relaxed version of the Hartree-Fock optimization problem, whose solutions always coincide with those of the original Hartree-Fock problem for UHF, as well as for the less popular General Hartree-Fock method (GHF) in which each spin-orbital is allowed to have both a spin-up and a spin-down component. For RHF, ODA and EDIIS most often converge to solutions to the RHF problem, but may occasionally converge to one-body density matrices with fractional occupation numbers, which do not correspond to Hartree-Fock states. A robust and efficient method to solve the RHF and UHF problems (which always works for UHF and most of the time for RHF) is to use EDIIS in the first iterations and switch to DIIS to accelerate convergence when the iterates are close enough to the solution [KSC02]. All the above algorithms are relatively well-understood from a mathematical point of view [Can+03].

Roughly speaking, computing RHF and UHF ground states for small and medium-size chemical systems is no longer an issue.

The situation is radically different for ROHF, where existing SCF algorithms fail to converge in many cases, notably for radicals and molecular systems containing transition metals.

In this article, we investigate the SCF algorithms for ROHF. We focus on maximum spin states in order to simplify the presentation, but our approach is valid for any spin state (see Remark 2.2.1). In Section 2.2, we recall the mathematical structure of the ROHF ground state problem in both the density matrix and molecular orbital formalisms. In particular, we point out that the ROHF minimization space has the geometry of a flag manifold, a structure that has been described in the mathematical literature (see e.g. [Lee13; YWL22]). Using this formalism, we derive from a geometric perspective the first-order optimality conditions for the ROHF problem, the ROHF equations.

In contrast with the RHF and UHF settings, the ROHF equations cannot be *naturally* formulated as a nonlinear eigenvalue problem. As a consequence, the simple SCF Roothaan scheme for RHF, *"assemble the Fock matrix for the current iterate, diagonalize it, build the next iterate using the Aufbau principle, that is by selecting the lowest energy orbitals"*, cannot be straightforwardly extended to the ROHF setting. All the existing SCF algorithms we are aware of twist the ROHF equations using coupling operators to transform them into a nonlinear eigenvalue problem. They are based on the construction of a composite, non-physical, effective Hamiltonian obtained by linear combinations of sub-blocks of the Fock matrices F_d and F_s respectively associated to the doubly and singly ROHF orbitals (also refereed to as *internal* and *active* orbitals). These combinations involve six real coefficients A_{tt}, and B_{tt} with t equal to d (doubly occupied), s (singly occupied), or v (virtual), the choice of which characterizes the SCF scheme. For instance, these six coefficients are all equal to 1/2 in the Guest and Saunders algorithm [GS74], but are different and depend on the spin state in the Canonical-I and Canonical-II algorithms introduced by Plakhutin and Davidson [PD14]. From the physical point of view, the choice of A_{tt} and B_{tt} coefficients essentially tries to maintain the *Aufbau* principle in order to avoid numerical instabilities of the SCF algorithm induced by swapping of the singly occupied orbital with doubly occupied or virtual orbitals. It is important to stress that, because of the mathematical restriction imposed by the ROHF Slater determinant, the *Aufbau* principle, inspired by the Koopman theorem, is not guaranteed, and therefore a choice of A_{tt} and B_{tt} which might work for a given system might break down for another, as illustrated for instance in the numerical results reported here (see Section 2.4.2).

In Section 2.3, we present a new SCF scheme, which better respects the essence of the ROHF equations and which is parameter-free. We then briefly describe how the DIIS acceleration algorithms write on the flag manifold of ROHF states. In Section 2.3.2.1, we extend the ODA to the ROHF setting. In Section 2.4, we compare the performance of the new algorithms introduced in this article to the state-of-the-art SCF algorithms for some challenging chemical systems, such as organic ligands chelating – or simply interacting with – transition metals. Although computationally demanding in their current state, our new algorithms showcase robust convergence properties, and give new perspective on the design of black-box SCF algorithms for open-shell systems.

2.2 The ROHF optimization problem

In this section, we first present the ROHF model in density matrices (DM) and molecular orbitals (MO) formalisms (without virtual orbitals). We then introduce the manifold of ROHF states. This manifold has a rich geometrical structure, known as a flag manifold. Although they are equivalent, each formalism DM or MO produces a specific discretization of the flag manifold of ROHF states, the ROHF energy gradient and optimality conditions, each one providing some insight on the ROHF problem.

2.2.1 The ROHF model

In ROHF theory, trial wavefunctions Ψ are not, in general, single Slater determinants, but configuration state functions (CSFs) [PD14; HJO14]. The latter are eigenfunctions of the spin operators $\hat{S}^2$ and $\hat{S}_z$ and of the number operators $\widehat{n}_i = a_{i\uparrow}^\dagger a_{i\uparrow} + a_{i\downarrow}^\dagger a_{i\downarrow}$, for a given orthonormal basis of orbitals $(\varphi_1, \varphi_2, \cdots)$ of $L^2(\mathbb{R}^3; \mathbb{C})$:

$$\hat{S}^2\Psi = s(s+1)\Psi, \quad \hat{S}_z\Psi = m_s\Psi, \quad \widehat{n}_i\Psi = n_i\Psi,$$

for given $s \in \frac{1}{2}\mathbb{N}$, $m_s \in \{-s, -s+1, \cdots, s-1, s\}$, and $n_i \in \{0, 1, 2\}$. Up to reordering the orbitals, we can assume that

$$\begin{aligned} n_i &= 2 \quad \forall i = 1, \cdots, N_d, \\ n_i &= 1 \quad \forall i = N_d + 1, \cdots, N_d + N_s, \\ n_i &= 0 \quad \forall i > N_d + N_s. \end{aligned}$$

Then, Ψ is a finite sum of Slater determinants, each of them made of the N_d doubly occupied orbitals $\varphi_1, \cdots, \varphi_{N_d}$ and N_s spin-orbitals of the form $\varphi_{N_d+1} \otimes \eta_1, \cdots, \varphi_{N_d+N_s} \otimes \eta_{N_s}$, the function η_j being equal to either α (spin-up) or β (spin-down). The numbers N_d, N_s, N (number of electrons in the system), s, and $m_{\rm s}$ are such that

$$2N_d + N_s = N, \quad |m_{\rm s}| \leq s \leq \frac{1}{2}N_s.$$

We also denote by $N_o := N_d + N_s$ the number of (singly or doubly) occupied orbitals.

For maximum spin states ($s = \frac{1}{2}N_s$) and maximum m_s value ($m_{\rm s} = s$), ROHF trial wavefunctions are single Slater determinants built with N_d doubly occupied orbitals $\varphi_1, \cdots, \varphi_{N_d}$ and N_s spin-up-orbitals $\varphi_{N_d+1} \otimes \alpha, \cdots, \varphi_{N_o} \otimes \alpha$, where the φ_i's satisfy $\langle \varphi_i | \varphi_j \rangle = \delta_{ij}$ for all $1 \leq i, j \leq N_o$. The electronic Hamiltonian

$$H_N = -\frac{1}{2}\sum_{i=1}^{N} \Delta_{\mathbf{r}_i} + \sum_{i=1}^{N} V_{\rm nuc}(\mathbf{r}_i) + \sum_{1 \leq i < j \leq N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}$$

being real-valued in the absence of external magnetic field and spin-orbit coupling, we can assume without loss of generality that the orbitals φ_i are real-valued. In order to obtain a computationally tractable model, the φ_i's are expanded in a finite basis set $\mathcal{X} := (\chi_1, \cdots, \chi_{N_b})$ of real-valued functions of the space variable:

$$\varphi_i(\mathbf{r}) = \sum_{\mu=1}^{N_b} [C_o]_{\mu i} \chi_\mu(\mathbf{r}).$$

In practice, the χ_μ's are non-orthogonal atomic orbitals (AO). In order to simplify the presentation, we will however assume here that the basis $\mathcal{X}$ is orthonormal, or equivalently that the overlap matrix is the identity matrix:

$$S_{\mu\nu} := \int_{\mathbb{R}^3} \chi_\mu(\mathbf{r}) \chi_\nu(\mathbf{r}) \, d\mathbf{r} = \delta_{\mu\nu}.$$

Let us emphasize that we make this simplification for pedagogical purposes only; extending our arguments to non-orthogonal basis sets is a simple exercise. In that setting, the orthonormality constraints on the orbitals imply that C_o is a rectangular orthogonal matrix; in other words, a point of the Stiefel manifold

$$C_o \in \mathrm{St}(N_o; \mathbb{R}^{N_b}) := \{C_o \in \mathbb{R}^{N_b \times N_o} \text{ s.t. } C_o^T C_o = I_{N_o}\} \tag{2.2.1}$$

where I_{N_o} denotes the identity matrix of rank N_o. In the following, it will be helpful to decompose C_o as two orthogonal matrices

$$C_o = (C_d | C_s) \quad \text{with} \quad C_d \in \mathbb{R}^{N_b \times N_d} \quad \text{and} \quad C_s \in \mathbb{R}^{N_b \times N_s} \tag{2.2.2}$$

corresponding to the coefficients of the doubly and singly occupied orbitals respectively.

From C_o, one can construct the density matrices (DM) P_d and P_s

$$P_d := C_d C_d^T \quad \text{and} \quad P_s := C_s C_s^T. \tag{2.2.3}$$

The matrices P_d and P_s are the basis representations of the orthogonal projectors on the spaces spanned by the doubly and singly occupied orbitals respectively. Recall that a square matrix P is an orthogonal projector if $P^2 = P = P^T$, and that its rank is the integer $\mathrm{tr}(P)$. These matrices represent the one-body density matrices (projectors)

$$\gamma_d = \sum_{i=1}^{N_d} |\varphi_i\rangle\langle\varphi_i| \quad \text{and} \quad \gamma_s = \sum_{i=N_d+1}^{N_o} |\varphi_i\rangle\langle\varphi_i| \tag{2.2.4}$$

in the basis set $\mathcal{X}$:

$$\gamma_d = \sum_{\mu,\nu=1}^{N_b} [P_d]_{\mu\nu} |\chi_\mu\rangle\langle\chi_\nu| \quad \text{and} \quad \gamma_s = \sum_{\mu,\nu=1}^{N_b} [P_s]_{\mu\nu} |\chi_\mu\rangle\langle\chi_\nu|.$$

We have the following equivalences:

$$\langle\varphi_i|\varphi_j\rangle = \delta_{ij} \text{ for all } 1 \le i,j \le N_o \quad \Leftrightarrow \quad C_d^T C_d = I_{N_d},\ C_s^T C_s = I_{N_s},\ C_d^T C_s = 0 \tag{2.2.5}$$

$$\Leftrightarrow \quad \begin{cases} P_d^2 = P_d = P_d^T,\ \mathrm{tr}(P_d) = N_d, \\ P_s^2 = P_s = P_s^T,\ \mathrm{tr}(P_s) = N_s, \\ P_d P_s = 0. \end{cases} \tag{2.2.6}$$

The maximum spin ROHF wavefunction Ψ generated by orthonormal doubly orbitals $(\varphi_1, \cdots, \varphi_{N_b})$ and singly occupied orbitals $(\varphi_{N_d+1}, \cdots, \varphi_{N_o})$ is completely determined (up to an irrelevant global phase) by the one-body density matrices γ_d and γ_s defined by (2.2.4). Conversely any pair (γ_d, γ_s) of orthogonal projectors satisfying $\mathrm{tr}(\gamma_d) = N_d$, $\mathrm{tr}(\gamma_s) = N_s$, and $\gamma_d\gamma_s = 0$ gives rise to a unique ROHF wavefunction $\Psi^{\mathrm{ROHF}}_{\gamma_d,\gamma_s}$ of maximal spin (up to a global phase), whose energy is a function of (γ_d, γ_s):

$$\mathcal{E}^{\mathrm{ROHF}}(\gamma_d, \gamma_s) := \langle \Psi^{\mathrm{ROHF}}_{\gamma_d,\gamma_s} | H_N | \Psi^{\mathrm{ROHF}}_{\gamma_d,\gamma_s} \rangle.$$

After discretization in the finite basis set $\mathcal{X}$, the ROHF energy functional becomes a function of the matrices P_d and P_s representing γ_d and γ_s in this basis:

$$E(P_d, P_s) := \mathcal{E}^{\mathrm{ROHF}} \left(\sum_{\mu,\nu=1}^{N_b} [P_d]_{\mu\nu} |\chi_\mu\rangle\langle\chi_\nu|,\ \sum_{\mu,\nu=1}^{N_b} [P_s]_{\mu\nu} |\chi_\mu\rangle\langle\chi_\nu| \right).$$

Standard algebraic manipulations lead to

$$\begin{aligned} E(P_d, P_s) = {} & \mathrm{tr}\left(h(2P_d + P_s)\right) + \mathrm{tr}\left((2J(P_d) - K(P_d))(P_d + P_s)\right) \\ & + \frac{1}{2}\,\mathrm{tr}\left((J(P_s) - K(P_s))P_s\right), \end{aligned} \tag{2.2.7}$$

where

$$[h]_{\mu\nu} = \frac{1}{2}\int_{\mathbb{R}^3} \nabla\chi_\mu(\mathbf{r}) \cdot \nabla\chi_\nu(\mathbf{r})\, d\mathbf{r} + \int_{\mathbb{R}^3} V_{\mathrm{nuc}}(\mathbf{r})\chi_\mu(\mathbf{r})\chi_\nu(\mathbf{r})\, d\mathbf{r},$$

$$[J(P)]_{\mu\nu} = \sum_{\kappa,\lambda=1}^{N_b} (\mu\nu|\kappa\lambda) P_{\kappa\lambda}, \quad [K(P)]_{\mu\nu} = \sum_{\kappa,\lambda=1}^{N_b} (\mu\kappa|\nu\lambda) P_{\kappa\lambda},$$

and

$$(\mu\nu|\kappa\lambda) := \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\chi_\mu(\mathbf{r})\chi_\nu(\mathbf{r})\chi_\kappa(\mathbf{r}')\chi_\lambda(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\, d\mathbf{r}\, d\mathbf{r}'.$$

In the following, we will use the fact that the matrix $h \in \mathbb{R}^{N_b \times N_b}_{\mathrm{sym}}$ is symmetric, and that the functions $J, K : \mathbb{R}^{N_b \times N_b}_{\mathrm{sym}} \to \mathbb{R}^{N_b \times N_b}_{\mathrm{sym}}$ are linear and such that

$$\mathrm{tr}(J(P)P') = \mathrm{tr}(J(P')P),\ \mathrm{tr}(K(P)P') = \mathrm{tr}(K(P')P) \text{ for all } P, P' \in \mathbb{R}^{N_b \times N_b}_{\mathrm{sym}}. \tag{2.2.8}$$

Note that the trace of P_d is equal to N_d, the number of doubly-occupied orbitals. The fact that each of these orbitals hosts two electrons is taken into account by the factors 2 in the first two terms of the right-hand side of Eq. (2.2.7). In view of (2.2.6), the density matrix (DM) formulation of the ROHF ground state problem in the basis $\mathcal{X}$ reads

$$\mathcal{E}^{\mathrm{ROHF}}_* := \min\{E(P_d, P_s),\ (P_d, P_s) \in \mathcal{M}_{\mathrm{DM}}(N_d, N_s; \mathbb{R}^{N_b})\}, \tag{2.2.9}$$

where

$$\begin{aligned} \mathcal{M}_{\mathrm{DM}}(N_d, N_s; \mathbb{R}^{N_b}) := \Big\{ & (P_d, P_s) \in \mathbb{R}^{N_b \times N_b}_{\mathrm{sym}} \times \mathbb{R}^{N_b \times N_b}_{\mathrm{sym}} \mid P_d^2 = P_d,\ P_s^2 = P_s,\ P_d P_s = 0, \\ & \mathrm{tr}(P_d) = N_d,\ \mathrm{tr}(P_s) = N_s \Big\}. \end{aligned} \tag{2.2.10}$$

The set $\mathcal{M}_{\mathrm{DM}}$ is the set of admissible pairs of doubly and singly occupied density matrices, that are the pairs of matrices actually representing a maximum spin ROHF state in the basis $\mathcal{X}$.

Remark 2.2.1. *The optimization problem (2.2.9) corresponds to the ROHF model for maximum spin states ($|m_s| = s = \frac{1}{2}N_s$). For other spin states ($|m_s| \leq s < \frac{1}{2}N_s$), the ROHF problem still is of the form (2.2.9). The energy functional E has a different expression (due to the Fock exchange term coupling only spin-orbitals having the same spin), but remains a sum of linear and bilinear forms in (P_d, P_s). See e.g. ref. [HJO14] for the derivation of the non-maximal spin energy expressions using the genealogical coupling scheme. Note that the algorithms presented in this article, although formulated for maximum spin state case, can therefore be straightforwardly extended to any spin state.*

The ROHF energy in MO formalism can be deduced from (2.2.3) and (2.2.7), for all $C_o \in \mathrm{St}(N_o, \mathbb{R}^{N_b})$

$$\mathcal{E}(C_o) = E(C_d C_d^T, C_s C_s^T). \tag{2.2.11}$$

An important difference between the DM and MO formalisms is that an ROHF state is represented by one and only one point of $(P_d, P_s) \in \mathcal{M}_{\mathrm{DM}}(N_d, N_s; \mathbb{R}^{N_b})$ (more precisely, the manifold of ROHF states is diffeomorphic to $\mathcal{M}_{\mathrm{DM}}(N_d, N_s; \mathbb{R}^{N_b})$), while it is represented by an infinity of points in $\mathrm{St}(N_o; \mathbb{R}^{N_b})$, namely the points in the set

$$\left\{ C_o \begin{pmatrix} U_d & 0 \\ 0 & U_s \end{pmatrix} = (C_d U_d | C_s U_s), \text{ where } (U_d, U_s) \in \mathcal{O}_{N_d} \times \mathcal{O}_{N_s} \right\} \subset \mathrm{St}(N_o, \mathbb{R}^{N_b}). \tag{2.2.12}$$

where we denoted $\mathcal{O}_N = \{U \in \mathbb{R}^{N \times N} \text{ s.t. } U^T U = I_N\}$ the orthogonal group of $N \times N$ matrices.

One way to recover the unicity of representation of ROHF states in MO formalism relies on the abstract notion of quotient sets. We introduce the equivalence relation on $\mathrm{St}(N_o; \mathbb{R}^{N_b})$ defined by

$$C_o \sim C_o' \Leftrightarrow \exists (U_d, U_s) \in \mathcal{O}_{N_d} \times \mathcal{O}_{N_s} \text{ such that } C_o' = C_o \begin{pmatrix} U_d & 0 \\ 0 & U_s \end{pmatrix}, \tag{2.2.13}$$

such that the set (2.2.12) is an equivalence class for the equivalence relation (2.2.13). Then the set of all equivalence classes (2.2.12), defined as the quotient

$$\mathcal{M}_{\mathrm{MO}}(N_d, N_s; \mathbb{R}^{N_b}) := \mathrm{St}(N_o; \mathbb{R}^{N_b}) / \sim = \mathrm{St}(N_o; \mathbb{R}^{N_b}) / (\mathcal{O}_{N_d} \times \mathcal{O}_{N_s}) \tag{2.2.14}$$

is diffeomorphic to both $\mathcal{M}_{\mathrm{DM}}(N_d, N_s; \mathbb{R}^{N_b})$ and the set of ROHF states. In particular, a ROHF state is represented by one and only one element of $\mathcal{M}_{\mathrm{MO}}(N_d, N_s; \mathbb{R}^{N_b})$. Let us clarify the meaning of this property. An element of the quotient $\mathcal{M}_{\mathrm{MO}}(N_d, N_s; \mathbb{R}^{N_b})$ is by definition an equivalence class (2.2.13). It can therefore be represented by some $C_o \in \mathrm{St}(N_o; \mathbb{R}^{N_b})$ or by any $C_o' = C_o \begin{pmatrix} U_d & 0 \\ 0 & U_s \end{pmatrix}$, for $(U_d, U_s) \in \mathcal{O}_{N_d} \times \mathcal{O}_{N_s}$. Denoting $[\![C_o]\!]$ the equivalence class containing C_o, ($[\![C_o]\!] \in \mathcal{M}_{\mathrm{MO}}(N_d, N_s; \mathbb{R}^{N_b})$), we have

$$[\![C_o]\!] = [\![C_o \begin{pmatrix} U_d & 0 \\ 0 & U_s \end{pmatrix}]\!], \quad \forall (U_d, U_s) \in \mathcal{O}_{N_d} \times \mathcal{O}_{N_s}.$$

In addition $\mathcal{E}(C_o) = \mathcal{E}(C_o \begin{pmatrix} U_d & 0 \\ 0 & U_s \end{pmatrix})$ (*i.e.* all $C_o \in \mathrm{St}(N_o; \mathbb{R}^{N_b})$ in the same equivalence class have the same ROHF energy), so that $\mathcal{E}$ can be seen as a function from $\mathcal{M}_{\mathrm{MO}}(N_d, N_s; \mathbb{R}^{N_b})$ to $\mathbb{R}$, also denoted $\mathcal{E}$ for simplicity. We can therefore write the ROHF minimization problem in MO formalism as

$$\mathcal{E}_*^{\mathrm{ROHF}} := \min\{\mathcal{E}([\![C_o]\!]),\ [\![C_o]\!] \in \mathcal{M}_{\mathrm{MO}}(N_d, N_s; \mathbb{R}^{N_b})\}. \tag{2.2.15}$$

The quotient nature of $\mathcal{M}_{\mathrm{MO}}(N_d, N_s; \mathbb{R}^{N_b})$ is not a mere theoretical tool, but is crucial to build efficient implementations of optimization algorithms in MO representation. Yet, taking into account this specificity of MO formalism would require to introduce additional mathematical objects, which could obscure the main subject of our discussion. For that reason, we will mainly focus in the following on the DM formalism, for which $\mathcal{M}_{\mathrm{DM}}(N_d, N_s; \mathbb{R}^{N_b})$ can be seen as a simple subset of $\mathbb{R}^{N_b \times N_b}_{\mathrm{sym}} \times \mathbb{R}^{N_b \times N_b}_{\mathrm{sym}}$. Additionally, the ROHF energy functional has a simple form in DM representation, which makes the DM formalism well-suited for methodological developments.

From a mathematical point of view, $\mathcal{M}_{\mathrm{DM}}(N_d, N_s; \mathbb{R}^{N_b})$ and $\mathcal{M}_{\mathrm{MO}}(N_d, N_s; \mathbb{R}^{N_b})$ are smooth (*i.e.* infinitely differentiable, C^∞) compact manifolds. While the DM and MO parametrizations of ROHF states seem quite different, they are in fact two representations of a same geometric object, as we now discuss bellow.

2.2.2 The manifold of ROHF states

The purpose of this section is to give some insights on the manifolds of ROHF states $\mathcal{M}_{\mathrm{DM}}(N_d, N_s; \mathbb{R}^{N_b})$ and $\mathcal{M}_{\mathrm{MO}}(N_d, N_s; \mathbb{R}^{N_b})$. In order to simplify the notations, we will abbreviate the DM and MO sets as $\mathcal{M}_{\mathrm{DM}}$ and $\mathcal{M}_{\mathrm{MO}}$, and denote by x the points in $\mathcal{M}_{\mathrm{DM}}$ and $[\![y]\!]$ the points in $\mathcal{M}_{\mathrm{MO}}$.

Let us start with a point $x = (P_d, P_s) \in \mathcal{M}_{\mathrm{DM}}$. Since P_d is a rank-N_d orthogonal projector (*i.e.* a symmetric matrix fulfilling $P_d^2 = P_d$ and $\mathrm{Tr}(P_d) = N_d$), it can be diagonalized in an orthonormal basis of $\mathbb{R}^{N_b}$ and its only eigenvalues are 1 (multiplicity N_d) and 0 (multiplicity $N_s + N_v$). Likewise, P_s is a rank-N_s orthogonal projector. In addition, as $P_d P_s = 0$, we also have $P_s P_d = (P_d P_s)^T = 0$, which implies that P_d and P_s commute and can therefore be co-diagonalized in the same orthonormal basis. Introducing the projector

$$P_v := I_{N_b} - P_d - P_s$$

on the virtual space (the space spanned by the virtual orbitals), which satisfies $P_v^2 = P_v = P_v^T$, $\mathrm{tr}(P_v) = N_v$, and $P_d P_v = P_s P_v = 0$, we obtain that there exists a unitary matrix $C \in O_{N_b}$ such that

$$P_d = C\mathcal{I}_d C^T, \quad P_s = C\mathcal{I}_s C^T, \quad P_v = C\mathcal{I}_v C^T, \quad CC^T = I_{N_b}, \tag{2.2.16}$$

where

$$\mathcal{I}_{N_d} = \begin{pmatrix} I_{N_d} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathcal{I}_{N_s} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & I_{N_s} & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathcal{I}_{N_v} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I_{N_v} \end{pmatrix}. \tag{2.2.17}$$

The equations (2.2.16) and (2.2.17) are equivalent to finding an orthonormal basis of eigenvectors (which form the unitary matrix C) of the projectors and selecting the ones corresponding to the eigenvalue 1. Decomposing C as $(C_d|C_s|C_v)$ we have

$$P_d = C_d C_d^T, \quad P_s = C_s C_s^T \quad P_v = C_v C_v^T. \tag{2.2.18}$$

In other words, the set C_d (respectively C_s) is the set of N_d (respectively N_s) natural orbitals associated to the density matrix P_d (respectively P_s). The orbitals in C_v are then the orthogonal complement to C_d and C_s. The equations (2.2.18) provide a one-to-one correspondence between $(P_d, P_s) \in \mathcal{M}_{\mathrm{DM}}$ and the set of occupied natural orbitals $[\![C_o = (C_d|C_s)]\!] \in \mathcal{M}_{\mathrm{MO}}$.

This relation between MO and DM formalism can be seen in a geometrical setting, by considering the spaces

$$\mathcal{V}_d = \mathrm{Span}\big(\varphi_i,\, i \in \{1, \dots, N_d\}\big) \quad \text{and} \quad \mathcal{V}_s = \mathrm{Span}\big(\varphi_i,\, i \in \{N_d + 1, \dots, N_d + N_s\}\big) \tag{2.2.19}$$

spanned by the doubly and singly occupied orbitals respectively. In a discretization basis $\mathcal{X}$, the pair of spaces $(\mathcal{V}_d, \mathcal{V}_s)$ can be parametrized by the pair $(P_d, P_s) \in \mathcal{M}_{\mathrm{DM}}$ of respective orthogonal projectors onto $\mathcal{V}_d$ and $\mathcal{V}_s$. It is also parametrized by the coefficients C_d and C_s in the discretization basis $\mathcal{X}$ of an orthonormal basis of $\mathcal{V}_d$ and $\mathcal{V}_s$. Now, all basis sets represented by a matrix in $[\![C_o = (C_d|C_s)]\!]$ span the same spaces. Hence the couple $(\mathcal{V}_d, \mathcal{V}_s)$ is parametrized by a single point $[\![C_o]\!] \in \mathcal{M}_{\mathrm{MO}}$. Because of the orthonormality constraints (2.2.6), $\mathcal{V}_d$ and $\mathcal{V}_s$ verify

$$\begin{cases} \{0_{L^2(\mathbb{R}^3)}\} \subsetneq \mathcal{V}_d \subsetneq \mathcal{V}_d \oplus \mathcal{V}_s \subsetneq \mathrm{Span}\,(\mathcal{X}), \\ \dim(\mathcal{V}_d) = N_d, \quad \dim(\mathcal{V}_d \oplus \mathcal{V}_s) = N_d + N_s. \end{cases} \tag{2.2.20}$$

Mathematically, the pair of spaces $(\mathcal{V}_d, \mathcal{V}_d \oplus \mathcal{V}_s)$ with property (2.2.20) is called a flag with dimensions N_d and $N_d + N_s$. The set of all such pair of spaces has been studied in the mathematical literature (see e.g. [Lee13, Example 21.22]). It is a smooth manifold called a *flag manifold* and denoted $\mathrm{Flag}(N_d, N_d + N_s; \mathbb{R}^{N_b})$.

From the above reasoning there is a one-to-one correspondence between ROHF states and points on $\mathrm{Flag}(N_1, N_1 + N_2; \mathbb{R}^{N_b})$. In other words, the DM and MO sets are two discretizations of the flag manifold $\mathrm{Flag}(N_1, N_1 + N_2; \mathbb{R}^{N_b})$, which writes as the diffeomorphisms

$$\mathcal{M}_{\mathrm{MO}}(N_1, N_2; \mathbb{R}^{N_b}) \simeq \mathrm{Flag}(N_1, N_1 + N_2; \mathbb{R}^{N_b}) \simeq \mathcal{M}_{\mathrm{DM}}(N_1, N_2; \mathbb{R}^{N_b}). \tag{2.2.21}$$

In order to derive the first-order optimality conditions associated to the minimization problem (2.2.9) (a.k.a. the ROHF equations in DM formalism) from a simple geometrical argument, we have to identify the space $T_x\mathcal{M}_{\mathrm{DM}}$ to a point $x \in \mathcal{M}_{\mathrm{DM}}$ of the manifold, that is the vector space of velocities $q = (Q_d, Q_s) = \dot{p}(0)$ at $t = 0$ for all paths

$$p : [-1, 1] \ni t \mapsto p(t) \in \mathcal{M}_{\mathrm{DM}}, \quad \text{such that} \quad p(0) = x \tag{2.2.22}$$

drawn on $\mathcal{M}_{\mathrm{DM}}$ (as shown in Fig. 2.1). Similarly, the ROHF equations in MO formalism are found by identifying the tangent spaces $T_{[\![y]\!]}\mathcal{M}_{\mathrm{MO}}$.

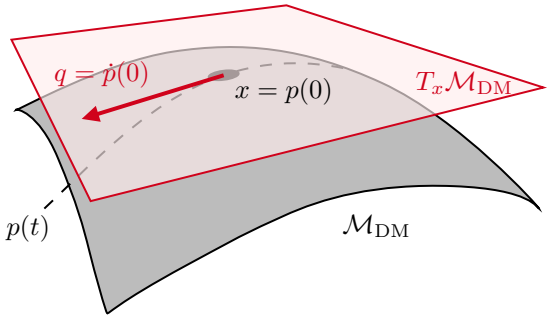


Figure 2.1 – Representation of the tangent space $T_x\mathcal{M}_{\mathrm{DM}}$ at x to the manifold $\mathcal{M}_{\mathrm{DM}}$, and a smooth path $p : [-1, 1] \ni t \mapsto p(t) \in \mathcal{M}_{\mathrm{DM}}$ drawn on $\mathcal{M}_{\mathrm{DM}}$ such that $p(0) = x$ and $\dot{p}(0) = q \in T_x\mathcal{M}_{\mathrm{DM}}$.

Flag manifolds, such as $\mathcal{M}_{\mathrm{DM}}$ and $\mathcal{M}_{\mathrm{MO}}$ have been studied in the context of optimization in the recent work [YWL22], where the authors derive in particular the formulations for the tangent spaces $T_x\mathcal{M}_{\mathrm{DM}}$ and $T_{[\![y]\!]}\mathcal{M}_{\mathrm{MO}}$. To keep this article as self contained as possible, and to make it understandable to readers with limited background in differential geometry, we will adopt in the following section a pedestrian approach, and re-derive in a few lines the tangent spaces and first order optimality conditions in DM formalism. As mentioned above, details concerning the MO formalism are reported in appendix.

Remark 2.2.2. *In general, a flag of length d in a vector space $\mathbb{V}$ of dimension N_b is a sequence of subspaces $\{\mathcal{V}_i\}_{1\leqslant i\leqslant d}$ of $\mathbb{V}$ that is strictly increasing for the inclusion. This is to be understood as $\mathcal{V}_1 \subsetneq \cdots \subsetneq \mathcal{V}_d \subsetneq \mathbb{V}$. A standard example of a flag in $\mathbb{V}$ is given by $\{\mathcal{V}_i = \mathrm{Span}(e_1, \cdots, e_i)\}_{1\leqslant i\leqslant N_b}$ where $(e_1, \cdots, e_{N_b})$ is the canonical basis of $\mathbb{V}$. The set of all flags in $\mathbb{V}$ with fixed respective dimensions $\dim(V_i) = n_i$ is also a smooth manifold denoted $\mathrm{Flag}(n_1, \cdots, n_d; \mathbb{V})$ (see e.g. [Lee13, Example 21.22]).*

2.2.3 First-order optimality conditions

2.2.3.1 General considerations on optimization in the DM framework

Finding a point $x_* = (P_{d*}, P_{s*})$ in $\mathcal{M}_{\mathrm{DM}}$ which minimizes the energy functional defined in (2.2.7) requires the definition of the derivative of E with respect to the pair of density matrices $x = (P_d, P_s)$. The ROHF energy functional $E(P_d, P_s)$ is not only defined for density matrices, but for any pair of real-valued symmetric matrix $z = (W_d, W_s) \in \mathbb{R}^{N_b\times N_b}_{\mathrm{sym}} \times \mathbb{R}^{N_b\times N_b}_{\mathrm{sym}}$, which might not be admissible density matrices. Therefore, although the energy gradient $\nabla E(z)$ with respect to $z = (W_d, W_s)$ can be easily computed once a topology, allowing to define the later, has been chosen, imposing $\nabla E(z) = 0$ is not enough to find the optimal ROHF density matrices because of the constraints imposed by the properties of density matrices (see Eq. (2.2.10)). The reason for this is that the gradient $\nabla E(x)$ has a component outside the manifold $\mathcal{M}_{\mathrm{DM}}$ of density matrices, and following that component of the gradient will necessary lead outside the manifold $\mathcal{M}_{\mathrm{DM}}$. As illustrated in Fig. 2.2, the correct ROHF condition is therefore to find the point $x_* \in \mathcal{M}_{\mathrm{DM}}$ such that the projection of $\nabla E(x_*)$ onto the tangent space $T_{x_*}\mathcal{M}$ is zero.

2.2.3.2 Characterization of the DM tangent spaces

Let p be a path as in (2.2.22). We have for all $t \in [-1, 1]$,

$$p(t) \in \mathcal{M}_{\mathrm{DM}} \quad \text{and} \quad p(t) = x + tq + O(t^2) = (P_d + tQ_d + o(t), P_s + tQ_s + o(t)), \tag{2.2.23}$$

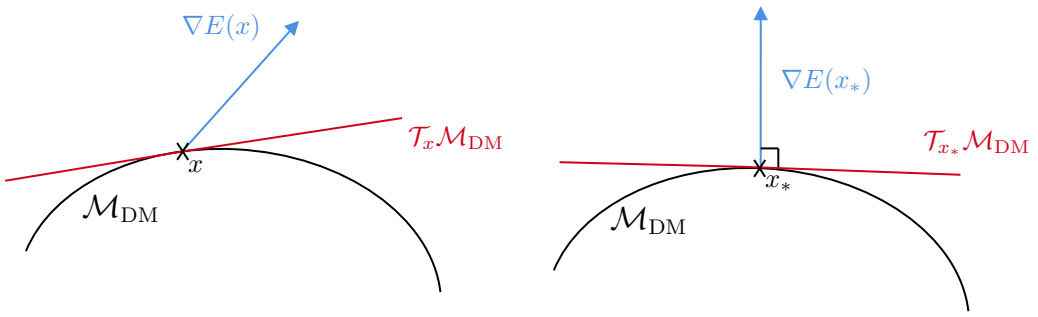


Figure 2.2 – Side view of the manifold $\mathcal{M}_{\rm DM}$ with tangent space and ambient ROHF energy gradient at (left) arbitrary point x (right) at optimal point x_*. The gradient $\nabla E(x_*)$ is orthogonal to the tangent space $T_{x_*}\mathcal{M}_{\rm DM}$ (first order optimality conditions).

where the $O(.)$ and $o(.)$ notations are relative to the usual Euclidean topology. In other words, the conditions (2.2.23) are equivalent to defining the tangent space $T_x\mathcal{M}_{\rm DM}$ to $x = (P_d, P_s)$ as the vector space of pairs of symmetric real matrices $q = (Q_d, Q_s)$ which allow to locally approximate the manifold of density matrices $\mathcal{M}_{\rm DM}$ by an affine space, as pictorially represented in Fig. 2.1. The constraints defining the manifold $\mathcal{M}_{\rm DM}$ (see Eq. (2.2.10)) are equivalent to the following at first order:

$$p_d(t)^2 = p_d(t), \quad \mathrm{tr}(p_d(t)) = N_d \quad \Leftrightarrow \quad P_dQ_d + Q_dP_d = Q_d, \quad \mathrm{tr}(Q_d) = 0, \tag{2.2.24}$$

$$p_s(t)^2 = p_s(t), \quad \mathrm{tr}(p_s(t)) = N_d \quad \Leftrightarrow \quad P_sQ_s + Q_sP_s = Q_s, \quad \mathrm{tr}(Q_s) = 0, \tag{2.2.25}$$

$$p_d(t)p_s(t) = 0 \quad \Leftrightarrow \quad P_dQ_s + Q_dP_s = 0. \tag{2.2.26}$$

In the representation (2.2.16)-(2.2.17), the constraints (2.2.24)-(2.2.26) are equivalent to

$$Q_d = C \begin{pmatrix} 0 & X & Y \\ X^T & 0 & 0 \\ Y^T & 0 & 0 \end{pmatrix} C^T \quad \text{and} \quad Q_s = C \begin{pmatrix} 0 & -X & 0 \\ -X^T & 0 & Z \\ 0 & Z^T & 0 \end{pmatrix} C^T, \tag{2.2.27}$$

where $X \in \mathbb{R}^{N_d \times N_s}$, $Y \in \mathbb{R}^{N_d \times N_v}$, $Z \in \mathbb{R}^{N_s \times N_v}$ are generic matrices. It follows that for all $x = (P_d, P_s)$ in $\mathcal{M}_{\rm DM}$:

$$\begin{aligned} T_x\mathcal{M}_{\rm DM} &= \{(Q_d, Q_s) \in \mathcal{V}_{\rm sym} \text{ of the form (2.2.27)}\} \\ &= \{(Q_d, Q_s) \in \mathcal{V}_{\rm sym} \mid P_dQ_dP_d = P_sQ_dP_s = P_vQ_dP_v = P_sQ_dP_v = 0, \\ &\qquad P_dQ_sP_d = P_sQ_sP_s = P_vQ_sP_v = P_dQ_sP_v = 0,\ P_d(Q_d + Q_s)P_s = 0\}. \end{aligned}$$

2.2.3.3 ROHF-Brillouin condition in the MO and DM framework

We denote the ambient DM space

$$\mathcal{V}_{\rm DM} = \mathbb{R}^{N_b \times N_b}_{\rm sym} \times \mathbb{R}^{N_b \times N_b}_{\rm sym} \tag{2.2.28}$$

endowed with the Frobenius-like scalar product

$$\langle (M_1, N_1), (M_2, N_2) \rangle_{\rm DM} := \frac{1}{2} \left(\mathrm{tr}(M_1M_2) + \mathrm{tr}(N_1N_2) \right). \tag{2.2.29}$$

Thanks to this inner product, the critical points of E on $\mathcal{M}_{\rm DM}$ can be characterized in a simple geometric way (see Fig. 2.2):

$$x_* \text{ critical point of } E \text{ on } \mathcal{M}_{\rm DM} \quad \Leftrightarrow \quad \nabla E(x_*) \in T_{x_*}\mathcal{M}^{\perp}_{\rm DM}, \tag{2.2.30}$$

where $\nabla E(x_*)$ is the gradient of E for the inner product $\langle \cdot, \cdot \rangle_{\rm DM}$, and $T_{x_*}\mathcal{M}^{\perp}_{\rm DM}$ the orthogonal subspace to $T_{x_*}\mathcal{M}_{\rm DM}$, still for the inner product $\langle \cdot, \cdot \rangle_{\rm DM}$. The condition of Eq. (2.2.30) is equivalent to state that, taken at the optimal point x_*, the component of $\nabla E(x_*)$ on the tangent plane $T_{x_*}\mathcal{M}$ is zero. Recall that for any $x \in \mathcal{V}_{\rm DM}$, $\nabla E(x)$ is the vector of $\mathcal{V}_{\rm DM}$ characterized by

$$E(x + \delta x) = E(x) + \langle \nabla E(x), \delta x \rangle_{\rm DM} + o(\delta x),$$

which implies that the gradient depends on the choice of inner product. Also, for any $x \in \mathcal{M}_{\rm DM}$, the vector space $T_x\mathcal{M}_{\rm DM}^\perp$ is defined by

$$T_x\mathcal{M}_{\rm DM}^\perp = \{q' \in \mathcal{V}_{\rm DM} \mid \forall q \in T_x\mathcal{M},\ \langle q, q'\rangle_{\rm DM} = 0\}\,.$$

Gradient of E. Let us first detail the computation of $\nabla E(x)$ for any ROHF state $x = (P_d, P_s) \in \mathcal{V}_{\rm DM}$. Introducing the Fock operators

$$F_d(P_d, P_s) := h + 2J(P_d) + J(P_s) - K(P_d) - \frac{1}{2}K(P_s), \tag{2.2.31}$$

$$F_s(P_d, P_s) := \frac{1}{2}\left(h + 2J(P_d) + J(P_s) - K(P_d) - K(P_s)\right), \tag{2.2.32}$$

we have for all $(M_d, M_s) \in \mathcal{V}_{\rm DM}$

$$\begin{aligned}
E(P_d + M_d, P_s + M_s) &= \mathrm{tr}\left(h(2P_d + 2M_d + P_s + M_s)\right) \\
&\quad + \mathrm{tr}\left((2J(P_d + M_d) - K(P_d + M_d))(P_d + M_d + P_s + M_s)\right) \\
&\quad + \frac{1}{2}\,\mathrm{tr}\left((J(P_s + M_s) - K(P_s + M_s))(P_s + M_s)\right) \\
&= E(P_d, P_s) + \mathrm{tr}\left(2F_d(P_d, P_s)M_d\right) + \mathrm{tr}\left(2F_s(P_d, P_s)M_s\right) \\
&\quad + \mathrm{tr}\left((2J(M_d) - K(M_d))(M_d + M_s)\right) + \frac{1}{2}\,\mathrm{tr}\left((J(M_s) - K(M_s))M_s\right) \\
&= E(P_d, P_s) + \langle (4F_d(P_d, P_s), 4F_s(P_d, P_s)), (M_d, M_s)\rangle_{\rm DM} \\
&\quad + \mathrm{tr}\left((2J(M_d) - K(M_d))(M_d + M_s)\right) + \frac{1}{2}\,\mathrm{tr}\left((J(M_s) - K(M_s))M_s\right).
\end{aligned}$$

The gradient of E at $x = (P_d, P_s)$ for the inner product $\langle\cdot,\cdot\rangle_{\rm DM}$ is therefore

$$\nabla E(x) = (4F_d(P_d, P_s), 4F_s(P_d, P_s)) \text{ with } F_d(P_d, P_s) \text{ and } F_s(P_d, P_s) \text{ given by (2.2.31)-(2.2.32).} \tag{2.2.33}$$

Characterization of $T_x\mathcal{M}_{\rm DM}^\perp$. Let $q' = (M_d, M_s) \in \mathcal{V}_{\rm DM}$. Using the decomposition

$$M_d = U \begin{pmatrix} M_d^{dd} & M_d^{ds} & M_d^{dv} \\ M_d^{sd} & M_d^{ss} & M_d^{sv} \\ M_d^{vd} & M_d^{vs} & M_d^{vv} \end{pmatrix} U^T \quad \text{and} \quad M_s = U \begin{pmatrix} M_s^{dd} & M_s^{ds} & M_s^{dv} \\ M_s^{sd} & M_s^{ss} & M_s^{sv} \\ M_s^{vd} & M_s^{vs} & M_s^{vv} \end{pmatrix} U^T, \tag{2.2.34}$$

and the fact that M_d and M_s are symmetric matrices, we obtain that for all $q = (Q_d, Q_s) \in T_x\mathcal{M}_{\rm DM}$ of the form (2.2.27),

$$\begin{aligned}
\langle q, q'\rangle_{\rm DM} &= \frac{1}{2}\,\mathrm{tr}\left(U \begin{pmatrix} 0 & X & Y \\ X^T & 0 & 0 \\ Y^T & 0 & 0 \end{pmatrix} U^T U \begin{pmatrix} M_d^{dd} & M_d^{ds} & M_d^{dv} \\ M_d^{sd} & M_d^{ss} & M_d^{sv} \\ M_d^{vd} & M_d^{vs} & M_d^{vv} \end{pmatrix} U^T\right) \\
&\quad + \frac{1}{2}\,\mathrm{tr}\left(U \begin{pmatrix} 0 & -X & 0 \\ -X^T & 0 & Z \\ 0 & Z^T & 0 \end{pmatrix} U^T U \begin{pmatrix} M_s^{dd} & M_s^{ds} & M_s^{dv} \\ M_s^{sd} & M_s^{ss} & M_s^{sv} \\ M_s^{vd} & M_s^{vs} & M_s^{vv} \end{pmatrix} U^T\right) \\
\Leftrightarrow \langle q, q'\rangle_{\rm DM} &= \mathrm{tr}\left(X^T(M_d^{ds} - M_s^{ds})\right) + \mathrm{tr}\left(Y^T M_d^{dv}\right) + \mathrm{tr}\left(Z^T M_s^{sv}\right).
\end{aligned} \tag{2.2.35}$$

Now, q' belongs to the orthogonal subspace $T_x\mathcal{M}_{\rm DM}^\perp$ if $\langle q, q'\rangle_{\rm DM} = 0$ for all $q \in T_x\mathcal{M}_{\rm DM}$. Therefore, according to Eq. (2.2.35)

$$q' \in T_x\mathcal{M}_{\rm DM}^\perp \quad \Leftrightarrow \quad \left(M_d^{ds} - M_s^{ds} = 0,\ M_d^{dv} = 0,\ M_s^{sv} = 0\right). \tag{2.2.36}$$

The critical points $x_* = (P_{d*}, P_{s*})$ of E on $\mathcal{M}_{\rm DM}$ are then characterized by the first-order optimality condition of Eq. (2.2.30), which according to Eqs. (2.2.33) and (2.2.36), leads to

$$(F_{d*} - F_{s*})^{ds} = 0,\ F_{d*}^{dv} = 0,\ F_{s*}^{sv} = 0, \quad \text{with} \quad F_{d*} := F_d(P_{d*}, P_{s*}) \text{ and } F_{s*} := F_s(P_{d*}, P_{s*}).$$

We recover the well-known ROHF optimality conditions (see e.g. [PD14]), which can also be written as

$$\left\{\begin{array}{l} P_{d*}(F_{d*}-F_{s*})P_{s*}=0, \quad P_{d*}F_{d*}P_{v*}=0, \quad P_{s*}F_{s*}P_{v*}=0, \\ \text{with } F_{d*}:=F_d(P_{d*},P_{s*}) \text{ and } F_{s*}:=F_s(P_{d*},P_{s*}). \end{array}\right. \tag{2.2.37}$$

We can similarly derive the optimality conditions in the MO representation, by endowing

$$\mathcal{V}_{\rm MO} := \mathbb{R}^{N_b\times N_o}$$

with the Frobenius inner product

$$\langle C_o|C_o'\rangle_{\rm MO} = {\rm tr}\left(C_o^TC_o'\right) = {\rm tr}\left(C_d^TC_d'\right) + {\rm tr}\left(C_s^TC_s'\right). \tag{2.2.38}$$

This inner product is natural since it reproduces the L^2-inner product. A calculation reported in appendix shows that for all $y=(C_d,C_s)\in\mathcal{V}_{\rm MO}$

$$\nabla\mathcal{E}(y) = (4F_d(C_dC_d^T,C_sC_s^T)C_d, 4F_s(C_dC_d^T,C_sC_s^T)C_s) \tag{2.2.39}$$

and that $y_*=(C_{d*},C_{s*})\in\mathcal{M}_{\rm MO}$ is a critical point of $\mathcal{E}$ on $\mathcal{M}_{\rm MO}$ if and only if

$$\left\{\begin{array}{l} F_{d*}C_{d*} = C_{d*}(C_{d*}^TF_{d*}C_{d*})\dfrac{1}{2}(C_{s*}(C_{s*}^T(F_{d*}+F_{s*})C_{d*}), \\ \\ F_{s*}C_{s*} = C_{s*}(C_{s*}^TF_{s*}C_{s*}) + \dfrac{1}{2}C_{d*}(C_{d*}^T(F_{s*}+F_{d*})C_{s*}), \\ \\ \text{with } F_{d*}:=F_d(C_{d*}C_{d*}^T,C_{s*}C_{s*}^T) \text{ and } F_{s*}:=F_s(C_{d*}C_{d*}^T,C_{s*}C_{s*}^T). \end{array}\right. \tag{2.2.40}$$

It can be checked that $C_*=(C_{d*},C_{s*})\in\mathcal{M}_{\rm MO}$ is solution to (2.2.40) if and only if $(P_{d*},P_{s*})\in\mathcal{M}_{\rm DM}$ is solution to (2.2.37), where $P_{d_*}:=C_{d*}C_{d*}^T$, $P_{s_*}:=C_{s*}C_{s*}^T$. An important implication of Eqs. (2.2.40) is that, unlike in the RHF and UHF frameworks, the optimal ROHF orbitals in C_{d*} and C_{s*} are not eigenfunctions of the Fock operators F_{d*} and F_{s*}, because of the second term in the right hand side of the first two equations in (2.2.40). As a consequence, SCF algorithms based on Fock-like operators involve *ad-hoc* effective Hamiltonians for which the *Aufbau* principal is not always satisfied (see for instance Ref. [PD14]).

2.3 Self-consistent field (SCF) algorithms

In this section, we first present the various basic SCF iterations proposed in the literature, and introduce a new one, which better respects the mathematical structure of the ROHF equations (2.2.37) and (2.2.40). We then discuss the stabilization and acceleration of basic SCF iterations using Anderson-Pulay (DIIS-type) algorithms.

2.3.1 Basic SCF iterations

The basic SCF algorithm for RHF was introduced by Roothaan [Roo60]. It consists in assembling the Fock matrix for the current iterate (molecular orbitals or density matrix), diagonalize it (we still assume orthonormality of the basis set for simplicity), and select the lowest energy eigenvectors to form the next iterate (*Aufbau* principle). This idea can be straightforwardly extended to the UHF model, but not to the ROHF model since the ROHF equations (2.2.40) cannot be formulated as a nonlinear eigenvalue problem.

Let $x^{(k)}=(P_d^{(k)},P_s^{(k)})\in\mathcal{M}_{\rm DM}$ be the current iterate and

$$P_d^{(k)} = C^{(k)}\mathcal{I}_dC^{(k)^T}, \quad P_s^{(k)} = C^{(k)}\mathcal{I}_sC^{(k)^T}, \quad C^{(k)}C^{(k)^T} = I_{N_b},$$

with $C^{(k)}=(C_d^{(k)}|C_s^{(k)}|C_v^{(k)})\in\mathcal{O}(N_b)$ the associated matrix of natural orbitals via (2.2.16). Let also $F_d^{(k)}:=F_d(P_d^{(k)},P_s^{(k)})$ and $F_s^{(k)}:=F_s(P_d^{(k)},P_s^{(k)})$ be the associated Fock matrices:

$$F_d^{(k)} = C^{(k)}\begin{pmatrix} {F_d^{(k)}}^{dd} & {F_d^{(k)}}^{ds} & {F_d^{(k)}}^{dv} \\ {F_d^{(k)}}^{sd} & {F_d^{(k)}}^{ss} & {F_d^{(k)}}^{sv} \\ {F_d^{(k)}}^{vd} & {F_d^{(k)}}^{vs} & {F_d^{(k)}}^{vv} \end{pmatrix}C^{(k)^T}, \quad F_s^{(k)} = C^{(k)}\begin{pmatrix} {F_s^{(k)}}^{dd} & {F_s^{(k)}}^{ds} & {F_s^{(k)}}^{dv} \\ {F_s^{(k)}}^{sd} & {F_s^{(k)}}^{ss} & {F_s^{(k)}}^{sv} \\ {F_s^{(k)}}^{vd} & {F_s^{(k)}}^{vs} & {F_s^{(k)}}^{vv} \end{pmatrix}C^{(k)^T}.$$

2.3.1.1 Standard approaches

The most popular simple SCF for ROHF consists in assembling and diagonalizing a composite effective Hamiltonian of the form

$$H_{A,B}^{(k)} := C^{(k)} \begin{pmatrix} R_{dd}^{(k)} & (F_d^{(k)} - F_s^{(k)})^{ds} & F_d^{(k)^{dv}} \\ (F_d^{(k)} - F_s^{(k)})^{sd} & R_{ss}^{(k)} & F_s^{(k)^{sv}} \\ F_d^{(k)^{vd}} & F_s^{(k)^{vs}} & R_{vv}^{(k)} \end{pmatrix} C^{(k)^T}, \tag{2.3.1}$$

where $R_{dd}^{(k)}$, $R_{ss}^{(k)}$, and $R_{vv}^{(k)}$ are symmetric matrices. The matrices $R_{tt}^{(k)}$ are of the form

$$R_{tt}^{(k)} = 2A_{tt}\left(F_s^{(k)}\right)^{tt} + 2B_{tt}\left(F_d^{(k)} - F_s^{(k)}\right)^{tt}, \quad t \in \{d, s, v\},$$

where $A = (A_{dd}, A_{ss}, A_{vv}) \in \mathbb{R}^3$ and $B = (B_{dd}, B_{ss}, B_{vv}) \in \mathbb{R}^3$ are coefficients characterizing the SCF algorithm (see Table I in [PD14]). For instance, they are all equal to 1/2 in Guest and Saunders algorithm [GS74], but are different and depend on the spin state in the Canonical-I and Canonical-II algorithms introduced by Plakhutin and Davidson [PD14]. The next iterate $(P_d^{(k+1)}, P_s^{(k+1)})$ is obtained by filling up first the doubly occupied orbitals, then the singly occupied orbitals, using the *Aufbau* principle. The meta-algorithm for the basic SCF iteration is summarized in the algorithm 1. The iterates are uniquely

Algorithm 1: Standard SCF iteration for ROHF

Given: $x^{(k)} = (P_d^{(k)}, P_s^{(k)}) \in \mathcal{M}_{\mathrm{DM}}$, $A = (A_{dd}, A_{ss}, A_{vv})$ and $B = (B_{dd}, B_{ss}, B_{vv})$.

1. Assemble $H_{A,B}^{(k)}$ and diagonalize in an orthonormal basis

$$H_{A,B}^{(k)} C_i^{(k+1)} = \varepsilon_i^{(k+1)} C_i^{(k+1)}, \quad (C_i^{(k+1)})^T C_j^{(k+1)} = \delta_{ij}, \quad \varepsilon_1^{(k+1)} \leqslant \cdots \leqslant \varepsilon_{N_b}^{(k+1)}.$$

2. Select the N_o first orbitals via the *Aufbau principle*

$$C_d^{(k+1)} = (C_1^{(k+1)}|\cdots|C_{N_d}^{(k+1)}), \quad C_s^{(k+1)} = (C_{N_d+1}^{(k+1)}|\cdots|C_{N_d+N_s}^{(k+1)}).$$

3. Construct the new iterate via (2.2.3)

$$P_d^{(k+1)} = C_d^{(k+1)} C_d^{(k+1)^T}, \quad P_s^{(k+1)} = C_s^{(k+1)} C_s^{(k+1)^T}, \quad x^{(k+1)} = (P_d^{(k+1)}, P_d^{(k+1)}).$$

defined provided

$$\varepsilon_{N_d}^{(k+1)} < \varepsilon_{N_d+1}^{(k+1)} \quad \text{and} \quad \varepsilon_{N_o}^{(k+1)} < \varepsilon_{N_o+1}^{(k+1)} \tag{2.3.2}$$

(energy gaps between doubly and single-occupied orbitals on the one-hand, occupied and virtual orbitals on the other hand). If the conditions (2.3.2) are not satisfied, iterates are defined by choosing randomly the orbitals among those satisfying the *Aufbau* principle, or by selecting the ones minimizing the ROHF energy functional. The SCF procedure interprets as a fix point method on the function $g_{A,B} : \mathcal{V}_{\mathrm{DM}} \longrightarrow \mathcal{M}_{\mathrm{DM}}$ defined by

$$g_{A,B}(x^{(k)}) := x^{(k+1)}, \text{ with } x^{(k+1)} = (P_d^{(k+1)}, P_s^{(k+1)}) \text{ as in algorithm 1.} \tag{2.3.3}$$

The basic SCF iterations (2.3.3) being extremely unstable (see section 2.4), they are generally stabilized by *direct inversion of the iterative subspace* (DIIS) schemes [Pul80; Pul82; HP86; RS11; Chu+21].

A necessary and sufficient condition for $(P_{d*}, P_{s*}) \in \mathcal{M}_{\mathrm{DM}}$ to be a fixed point of $g_{A,B}$ is

$$H_{A,B*} C_{i*} = \varepsilon_{i*} C_{i*}, \quad C_{i*}^T C_{j*} = \delta_{ij}, \quad \varepsilon_{1*} \leqslant \cdots \leqslant \varepsilon_{N_o*}. \tag{2.3.4}$$

Let $x_* = (P_{d*}, P_{s*})$ be such a fixed point and C_* the associated matrix of natural orbitals via (2.2.3). Then

$$P_{d*} H_{A,B*} P_{s*} = P_{d*}(F_{d*} - F_{s*}) P_{s*} = \sum_{i=Nd+1}^{N_o} P_{d*} H_{A,B*} C_{i*} C_{i*}^T = \sum_{i=N_d+1}^{N_o} \varepsilon_{i*} \underbrace{P_{d*} C_{i*}}_{=0} C_{i*}^T = 0.$$

A similar argument leads to $P_{d*}F_{d*}P_{v*} = 0$ and $P_{s*}F_{s*}P_{v*} = 0$ so that x_* satisfy the optimality conditions (2.2.37). Conversely, if x_* satisfies (2.2.37), then $H_{A,B*} = \mathrm{diag}\,(R_{dd}, R_{ss}, R_{vv})$ is bloc diagonal in the orthogonal decomposition $\mathrm{Ran}(P_{d*}) \oplus \mathrm{Ran}(P_{s*}) \oplus \mathrm{Ran}(P_{v*})$ of $\mathbb{R}^{N_b}$. Therefore, we have

$$\begin{cases} H_{A,B*}C_{i*} = \varepsilon_{i*}C_{i*}, \quad C_{i*}^T C_{j*} = \delta_{ij}, \\ P_{d*} = \displaystyle\sum_{i=1}^{N_d} C_{i*}C_{i*}^T, \quad P_{s*} = \sum_{i=N_d+1}^{N_o} C_{i*}C_{i*}^T, \end{cases}$$

for some orthonormal basis $(C_{i*})_{1\le i\le N_b}$ of $\mathbb{R}^{N_b}$ diagonalizing $H_{A,B*}$. It follows that a point $x_* \in \mathcal{M}_{\mathrm{DM}}$ is a critical point of E if and only if x_* satisfies the conditions (2.3.4) *except possibly the fact that the doubly-occupied orbitals do not necessarily correspond to the lowest N_d eigenvalues of $H_{A,B*}$, or the singly-occupied orbitals to the next N_s ones*, which is equivalent to saying that the *Aufbau* principle does not need to be satisfied *a priori*. As discussed in [PD14], there are indeed local minima of the ROHF problem for which the *Aufbau* principle is not satisfied for any of the usual choices of A and B. We are therefore facing a dilemma. Either the *Aufbau* principle can be kept in the definition of the SCF procedure, leading to a simple iterative scheme, which is however unable to find the ROHF ground state in some cases. Or the *Aufbau* principle can be discarded and replaced by a more complicated construction procedure, to be specified.

2.3.1.2 A new strategy not based on the *Aufbau* principle

A way out of this dilemma is to attack the problem from a different perspective, using another interpretation of the Roothaan scheme in DM formalism: in the RHF setting, the next iterate $P^{(k+1)}$ obtained by an SCF iteration is the point P of the RHF manifold

$$\mathcal{M}_{\mathrm{DM}}^{\mathrm{RHF}} := \left\{P \in \mathbb{R}^{N_b\times N_b}_{\mathrm{sym}} \mid P^2 = P,\ \mathrm{tr}(P) = N_d\right\}$$

in the direction along which the slope of the function $t \mapsto E^{\mathrm{RHF}}(P^{(k)} + t(P - P^{(k)}))$ is minimum [Can+03], *i.e.*

$$P^{(k+1)} \in \operatorname*{argmin}_{P\in\mathcal{M}_{\mathrm{DM}}^{\mathrm{RHF}}} \left\langle \nabla E^{\mathrm{RHF}}(P^{(k)}) \middle| P \right\rangle_{\mathcal{V}_{\mathrm{DM}}^{\mathrm{RHF}}} = \operatorname*{argmin}_{P\in\mathcal{M}_{\mathrm{DM}}^{\mathrm{RHF}}} \mathrm{Tr}\left[F^{\mathrm{RHF}}(P^{(k)})P\right] \tag{2.3.5}$$

where $F^{\mathrm{RHF}}(P) = \frac{1}{2}\nabla E^{\mathrm{RHF}}(P)$ is the Fock matrix associated with the density matrix P, and where $\mathcal{V}_{\mathrm{DM}}^{\mathrm{RHF}} = \mathbb{R}^{N_b\times N_b}_{\mathrm{sym}}$ is the ambient vector space for the RHF problem. In (2.3.5), argmin refers to the set of minimizers of the linear form $P \longrightarrow \left\langle \nabla E^{\mathrm{RHF}}(P^{(k)}), P\right\rangle_{\mathcal{V}_{\mathrm{DM}}^{\mathrm{RHF}}}$ on $\mathcal{M}_{\mathrm{DM}}^{\mathrm{RHF}}$, to which $P^{(k+1)}$ belongs. This set is always non empty, but may contain several elements. Transposing this characterization to the ROHF setting, we can define a new basic SCF scheme on the manifold $\mathcal{M}_{\mathrm{DM}}$: $x^{(k+1)} := (P_d^{(k+1)}, P_s^{(k+1)})$ is the point $x \in \mathcal{M}_{\mathrm{DM}}$ in the direction along which the slope of the function $t \mapsto E(x^{(k)} + t(x - x^{(k)}))$ is minimum. It is therefore obtained from $x^{(k)} = (P_d^{(k)}, P_s^{(k)})$ as

$$x^{(k+1)} \in \operatorname*{argmin}_{x\in\mathcal{M}_{\mathrm{DM}}} \langle \nabla E(x^{(k)}), x\rangle_{\mathcal{V}_{\mathrm{DM}}} = \operatorname*{argmin}_{x=(P_d,P_s)\in\mathcal{M}_{\mathrm{DM}}} \mathrm{tr}\Big(F_d^{(k)}P_d\Big) + \mathrm{tr}\Big(F_s^{(k)}P_s\Big), \tag{2.3.6}$$

where $F_d^{(k)} := F_d(P_d^{(k)}, P_s^{(k)})$ and $F_s^{(k)} := F_s(P_d^{(k)}, P_s^{(k)})$. This motivates the introduction of the new basic SCF scheme in algorithm 2 . The fixed points (P_{d*}, P_{s*}) of this SCF scheme verifies

Algorithm 2: New SCF iteration on $\mathcal{M}_{\mathrm{DM}}$

Given: $(P_d^{(k)}, P_s^{(k)}) \in \mathcal{M}_{\mathrm{DM}}$.

1. Compute the Fock matrices $F_d^{(k)} = F_d(P_d^{(k)}, P_s^{(k)})$ and $F_s^{(k)} = F_s(P_d^{(k)}, P_s^{(k)})$
2. Choose next iterate $(P_d^{(k+1)}, P_s^{(k+1)})$ in

$$\operatorname{argmin}\left\{\mathrm{Tr}\left[F_d^{(k)}P_d + F_s^{(k)}P_s\right],\ (P_d, P_s) \in \mathcal{M}_{\mathrm{DM}}\right\}$$

$$\begin{cases} (P_{d*}, P_{s*}) \in \operatorname{argmin} \{E_*(P_d, P_s),\ (P_d, P_s) \in \mathcal{M}_{\mathrm{DM}}\}, \\ \quad \text{with} \quad E_*(P_d, P_s) = \operatorname{tr}(F_d(P_{d*}, P_{s*})P_d) + \operatorname{tr}(F_s(P_{d*}, P_{s*})P_s). \end{cases} \tag{2.3.7}$$

Again this SCF procedure can be interpreted as a fix-point method on the function

$$g_{\mathrm{new}}(x^{(k)}) := x^{(k+1)}, \quad \text{with } x^{(k+1)} = (P_d^{(k+1)}, P_s^{(k+1)}) \text{ as in algorithm 2.} \tag{2.3.8}$$

As E_* is a linear form, its gradient is constant and equal for the inner product $\langle \cdot, \cdot \rangle_{\mathcal{V}_{\mathrm{DM}}}$ to $(4F_{d*}, 4F_{s*})$. Replacing E with E_* in the arguments in Section 2.2.3.3, we obtain that (2.3.7) implies (2.2.37), hence that any fixed point (P_{d*}, P_{s*}) of the function g_{new} is a critical point of E on $\mathcal{M}_{\mathrm{DM}}$.

The inner optimization problem

$$\operatorname{argmin} \left\{ \operatorname{Tr}\left(F_d^{(k)} P_d + F_s^{(k)} P_s\right),\ (P_d, P_s) \in \mathcal{M}_{\mathrm{DM}} \right\} \tag{2.3.9}$$

on $\mathcal{M}_{\mathrm{DM}}$ solved at each step is easier and much cheaper to solve numerically than the original problem (2.2.9) since the function $(P_d, P_s) \mapsto \operatorname{Tr}(F_d^{(k)} P_d + F_s^{(k)} P_s)$ is *linear* while the ROHF energy function $E(P_d, P_s)$ is nonlinear (see Eq. (2.2.7)). In particular, the Coulomb and Fock terms are not recomputed at each iteration. To solve it, we can use a direct minimization algorithm with initial guess in

$$\operatorname{argmin} \left\{ \operatorname{Tr}(H^{(k)} P_d + \frac{1}{2} H^{(k)} P_s),\ (P_d, P_s) \in \mathcal{M}_{\mathrm{DM}} \right\}, \tag{2.3.10}$$

where $H^{(k)} = F_d^{(k)}$, or $H^{(k)} = H_{A,B}^{(k)}$, with $H_{A,B}^{(k)}$ given by (2.3.1). The solutions to (2.3.10) are easily obtained by diagonalizing $H^{(k)}$ and applying the *Aufbau* principle. For $H^{(k)} = H_{A,B}^{(k)}$, the iterate of the new basic SCF scheme (2) is obtained using $g_{A,B}(P_d^{(k)}, P_s^{(k)})$ as initial guess for the minimization problem (2.3.9). It is also possible and more efficient in some cases to use, as an initial guess for the minimization problem (2.3.9), the previous iterate $(P_d^{(k-1)}, P_s^{(k-1)})$. Let us mention however that this approach only provides *local* (non-necessarily global) minima of (2.3.9). In practice, we choose for $(P_d^{(k+1)}, P_s^{(k+1)})$ the approximation of the local minimum of $(P_d, P_s) \mapsto \operatorname{Tr}(F_d^{(k)} P_d + F_s^{(k)} P_s)$ on $\mathcal{M}_{\mathrm{DM}}$ obtained by a few iterations of a preconditioned steepest-descent algorithm.

2.3.2 Anderson-Pulay (DIIS-type) acceleration

Anderson-Pulay acceleration (APA) is a terminology recently coined in [Chu+21] to gather various acceleration schemes into a general framework, including the *Anderson acceleration* scheme [And65] and the DIIS scheme. Anderson-Pulay acceleration methods can be applied to any fixed-point problems of the form

$$\text{find } x_* \in \mathcal{W} \text{ such that } g(x_*) = x_* \tag{2.3.11}$$

where $g : \mathcal{W} \to \mathcal{M}$ is a C^2 function from an open subset $\mathcal{W}$ of $\mathbb{R}^n$ into a smooth submanifold $\mathcal{M}$ of $\mathbb{R}^n$. In addition to the fix point map g, APA schemes require a residual function $f : \mathcal{W} \to \mathbb{R}^p$ of class C^2 with $p \leqslant n$, such that for any $x \in \mathcal{W}$, $g(x) = x$ if and only if $f(x) = 0$ (the residual vanishes at solutions to the fixed point problem and only at those points). A possible choice is $f(x) = x - g(x)$ (in which case $p = d$), but the performance of the algorithm can usually be dramatically improved by resorting to well suited residual functions. The APA schemes are based on linear combinations of the current iterate with the previous ones, up to a certain depth $0 \leqslant m \leqslant m_{\max}$. As an example, the standard DIIS acceleration scheme writes for a given depth m, and fix-point map g

$$x^{(k+1)} = g(\mathcal{A}_{\mathrm{DIIS}}(x^{(k)}, \dots x^{(k-m)})) \tag{2.3.12}$$

where the map $\mathcal{A}_{\mathrm{DIIS}}$ is defined as follows. Let $r^{(k)} := f(x^{(k)})$ and define

$$\mathscr{Y}^{(k)} = \left[x^{(k-m+1)} - x^{(k-m)}, \cdots, x^{(k)} - x^{(k-1)}\right], \quad \mathscr{S}^{(k)} = \left[r^{(k-m+1)} - r^{(k-m)}, \cdots, r^{(k)} - r^{(k-1)}\right].$$

Then

$$\mathcal{A}_{\mathrm{DIIS}}(x^{(k)}, \cdots, x^{(k-m)}) := x^{(k)} + r^{(k)} - (\mathscr{Y}^{(k)} + \mathscr{S}^{(k)})\alpha^{(k)}, \tag{2.3.13}$$

where the coefficients $\alpha^{(k)} \in \mathbb{R}^m$ are solution to the least square problem

$$\alpha^{(k)} \in \operatorname*{argmin}_{\alpha \in \mathbb{R}^m} \left\| r^{(k)} - \mathscr{J}^{(k)} \alpha \right\|_{\mathbb{R}^p}^2 .$$

Mathematical studies on the convergence of DIIS algorithms can be found in [RS11; CKL21; Chu+21]. The parameter $m_{\max}$ must be chosen large enough (typically $m_{\max} = 10$ or 20 in quantum chemistry packages) to ensure fast convergence, using sufficient information from previous iterations. One of the limitations of DIIS is that iterates with large residuals (far away from the minimizer) are considered as well, whereas they should be discarded. To cure this deficiency, an adaptive depth approach is proposed in [Chu+21], which should be investigated.

Choice of g. In order to be applied to SCF iterations, we need an iteration function defined in an open neighborhood $\mathcal{W}$ of $\mathcal{M}_{\mathrm{DM}}$ since the points $\mathcal{A}(x^{(k)}, \ldots, x^{(k-m)})$, which are linear combinations of points of $\mathcal{M}_{\mathrm{DM}}$, do not belong to $\mathcal{M}_{\mathrm{DM}}$ in general. We can directly use one of the basic SCF iteration functions $g_{A,B}$ or g_{new} corresponding to the respective algorithms 1 and 2, since they are defined for any point of $\mathcal{V}_{\mathrm{DM}}$.

Choice of f. From (2.2.37), a natural choice for the residual function is to take for all $x = (P_d, P_s)$

$$f(P_d, P_s) := ((F_d(P_d, P_s) - F_s(P_d, P_s))^{ds},\ (F_d(P_d, P_s))^{dv},\ (F_s(P_d, P_s))^{sv}) \tag{2.3.14}$$

which is the projection on $T_x \mathcal{M}_{\mathrm{DM}}^{\perp}$ of the gradient $\nabla E(x)$. Remark that this is but a geometrical derivation of the standard commutator based residual used e.g. in GAMESS. In DIIS algorithms, the residual function f is only evaluated at points of the manifold $\mathcal{M}_{\mathrm{DM}}$, but must have a C^2 extension to $\mathcal{W}$ for local convergence to be mathematically guaranteed [Chu+21]. This is obviously the case for the function f defined by (2.3.14) on $\mathcal{M}_{\mathrm{DM}}$.

2.3.2.1 Relaxed constrained algorithms for ROHF

Relaxed constrained algorithms for the Unrestricted and General Hartree-Fock setting were introduced in [CB00]. They consist in optimizing the energy functional in the DM formulation on the convex hull of the admissible set. For the UHF and GHF problems, it can be shown that the relaxed constrained problem has the same global minimizers as the original one [Can00; CKL21]. The advantage of the relaxed constrained problems is that convex combinations of admissible solutions are admissible solutions as well.

Algorithm 3: ODA iteration for ROHF

Given: current Fock-like matrices $(\widetilde{F}_d^{(k)}, \widetilde{F}_s^{(k)})$

1. Pick $(P_d^{(k+1)}, P_s^{(k+1)}) \in \operatorname{argmin} \left\{ \operatorname{Tr}\left(\widetilde{F}_d^{(k)} P_d + \widetilde{F}_s^{(k)} P_s \right),\ (P_d, P_s) \in \mathcal{M}_{\mathrm{DM}} \right\}$
2. Compute the Fock matrices $F_d^{(k+1)} := F_d(P_d^{(k+1)}, P_s^{(k+1)})$, $F_s^{(k+1)} := F_s(P_d^{(k+1)}, P_s^{(k+1)})$ and set

$$(\widetilde{P}_d^{(k+1)}, \widetilde{P}_s^{(k+1)}) = (1 - t_k)(\widetilde{P}_d^{(k)}, \widetilde{P}_s^{(k)}), + t_k (P_d^{(k+1)}, P_s^{(k+1)})$$
$$(\widetilde{F}_d^{(k+1)}, \widetilde{F}_s^{(k+1)}) = (1 - t_k)(\widetilde{F}_d^{(k)}, \widetilde{F}_s^{(k)}), + t_k (F_d^{(k+1)}, F_s^{(k+1)})$$

where t_k is the minimizer of the quadratic function

$$[0, 1] \ni t \mapsto E((1 - t)(\widetilde{P}_d^{(k)}, \widetilde{P}_s^{(k)}) + t(P_d^{(k+1)}, P_s^{(k+1)})).$$

The simplest relaxed constrained algorithm is the optimal damping algorithm (ODA). It generates two sequences of iterates:

- a sequence $(x^{(k)})$ of points on the admissible manifold $\mathcal{M}_{\mathrm{DM}}$;
- a sequence $(\widetilde{x}^{(k)})$ of points in the convex hull of $\mathcal{M}_{\mathrm{DM}}$.

The point $\widetilde{x}^{(k+1)}$ is obtained by doing an optimal convex combination of $\widetilde{x}^{(k)}$ and $x^{(k+1)}$:

$$t_k = \underset{t\in[0,1]}{\operatorname{argmin}} E(tx^{(k+1)} + (1-t)\widetilde{x}^{(k)}), \quad \widetilde{x}^{(k+1)} = t_k x^{(k+1)} + (1-t_k)\widetilde{x}^{(k)}.$$

The function $p_k(t) := E(tx^{(k+1)} + (1-t)\widetilde{x}^{(k)})$ is a second degree polynomial and we have

$$p_k(0) = E(\widetilde{x}^{(k)}) \quad \text{and} \quad p_k'(0) = \langle \nabla E(\widetilde{x}^{(k)}), x^{(k+1)} - \widetilde{x}^{(k)} \rangle_{\mathcal{V}_{\mathrm{DM}}}.$$

Computing $p_k(1) = E(x^{(k+1)})$, we obtain the value of t_k explicitly. The point $x^{(k+1)}$ is chosen so as to minimize the slope $p_k'(0)$; it is therefore obtained from $\widetilde{x}^{(k)}$ as

$$x^{(k+1)} \in \underset{x\in\mathcal{M}_{\mathrm{DM}}}{\operatorname{argmin}} \langle \nabla E(\widetilde{x}^{(k)}), x \rangle_{\mathcal{V}_{\mathrm{DM}}} = g_{\mathrm{new}}(\widetilde{x}^{(k)}),$$

where g_{new} is defined in (2.3.8). The ODA is initialized by choosing an initial guess $x^{(0)} = (P_d^{(0)}, P_s^{(0)})$ in $\mathcal{M}_{\mathrm{DM}}$, by setting $\widetilde{x}^{(0)} = x^{(0)}$, and by computing $(\widetilde{F}_d^{(0)}, \widetilde{F}_s^{(0)}) = (F_d(P_d^{(0)}, P_s^{(0)}), F_s(P_d^{(0)}, P_s^{(0)}))$. One then performs ODA iteration as written in algorithm 3.

2.4 Numerical results

2.4.1 Methodology and summary of the results

We now analyze the performance of the algorithms introduced in this article which are

- the standard SCF (algorithm 1) and new SCF (algorithm 2), with respective fix point map $g_{A,B}$ and g_{new}, endowed with a DIIS acceleration with residual f given by (2.3.14);
- the ODA scheme as described in algorithm 3.

Convergence behaviors are investigated in two distinct regimes:

- the global convergence regime. The goal here is to reach the vicinity of a minimizer, starting from a bad initial guess obtained in practice by diagonalizing the core Hamiltonian;
- the local convergence regime, when the initial guess is close to a minimizer. We choose in this study the extended Hückel initial guess derived from the Wolfsberg-Helmholtz approximation [WH52; Hof63; Amm+78].

Our implementation. The application of the g_{new} map requires to solve the inner optimization problem (2.3.9). In our implementation, we use the initial guess (2.3.10) with $H^{(k)} = F_d^{(k)}$. We then apply a maximum of 10 iterations of preconditioned steepest descent on the DM manifold.

For the ODA method, it happens in some cases that the coefficient t_k of the ODA convex combination becomes zero, which results in the algorithm getting stuck on the iterate $x^{(k)}$. In that case, we automatically try a different guess for the inner problem (2.3.9). Using a guess generated with $g_{A,B}$ and Euler or Guest and Saunders coefficients whenever $t_k = 0$ proved effective in all the cases we encountered.

The new algorithms we introduce, along with the classical SCF schemes, have been implemented in a `Julia` [Bez+17] package as a proof of concept. This package is built as an overlay to the `PySCF` [Sun+20] python library, which handles the core computations for ROHF (generation of the AO basis and initial MOs, computation of the electronic integrals). Comprehensive details of implementation can be found in our open-source research code `https://github.com/LaurentVidal95/ROHFToolkit`. The best performing algorithms will be added as a plugin within the Quantum Package [Sce+16; Gar+19] and made freely available to the community.

Comparison to external code. In order to assert the validity of our code, we compare the performances of our algorithms with the SCF algorithms for ROHF available in `GAMESS` [Sch+93]. We have chosen this popular software because all the classical functions $g_{A,B}$ are implemented, as well as the residual (2.3.14) for DIIS and the SOSCF algorithm. We have also run tests with `PySCF` and `Psi4` [Tur+12] (which respectively implement Roothan and Guest and Saunders' $g_{A,B}$). The DIIS residual functions implemented in these codes can be slightly different but all also use commutator-based residual functions *à la* Pulay [Pul82], involving the effective Hamiltonian $H_{A,B}$.

The initial guesses for the SCF problem, the one generated by `GAMESS` and those employed in our implementation (generated by `PySCF`), can differ significantly. Specifically, the extended Hückel guess in `GAMESS` tends to yield energies approximately 1 to 2 Ha above the ground state energy in our test cases, while the `PySCF` Hückel guesses produce initial energies ranging from 20 to 60 Ha above the ground state. To ensure methodological consistency, and facilitate the direct comparison between the two codes, we manually imported the `GAMESS` Hückel guess in our code for the 6-31G basis set [HP74; Fra+82; Bla+97] for some of our test cases. We observed no qualitative difference for this choice of basis set. Unfortunately, the two quantum chemistry packages employ different conventions in generating atomic basis sets, particularly concerning the number and order of the atomic orbitals, which makes the systematic import of `GAMESS` guesses in our code a laborious task. The comparison with `GAMESS` should therefore only serve as a qualitative evaluation of our implementation.

Global convergence regime. First, the algorithms are tested by starting *very far* from an expected minimum, *i.e.* starting from a core Hamiltonian diagonalization guess, obtained with `GAMESS` and `PySCF` respectively. Poor quality guesses do not usually verify the *Aufbau* principle on which the classical SCF methods, built with function $g_{A,B}$ (2.3.3), rely (as recalled section 2.3). Numerical results presented in Section 2.4.3.1 confirm that, unlike the classical SCF methods for ROHF, which mostly fail to converge in this regime (in all the tested cases but the simplest one), our methods built on g_{new}, which are free of *Aufbau* principle requirement, exhibit a strong robustness with respect to the initial guess.

Local convergence regime. As detailed in section 2.4.3.2, existing methods built on the classical $g_{A,B}$ barely benefit from the use of an extended Hückel guess, which is more commonly used in practice. Only two or four choices of A_{tt} and B_{tt} coefficients, depending on the test case, yield convergence for these so-called $g_{A,B}$-based methods (see Table 2.4), with the Guest and Saunders choice being the most successful. Our g_{new}–based methods, that are free of the choice of such coefficients, manage to converge in all cases from this starting guess.

Local minima. The respective $g_{A,B}$-based methods, as well as our g_{new}-based methods, converge toward a variety of local minima. The list of all minima have been reported in appendix. Note that the variation in the implementation of basis sets between GAMESS and our code results in a minor difference in energies. A detailed analysis of the encountered local minima, reached from the core guess and from the Hückel guess, would be needed to assess their quality. It appears that in some cases, the local minima found by starting from the core initial guess, are lower in energy than other minima reached from extended Hückel initial guesses. One should elaborate further on this point in another study.

Our best performing method. When focusing on the energy only, the ODA algorithm seems to target a low minima, independently of the initial guess, while being very slow to converge to chemical accuracy. Applying a few iterations of ODA, followed by g_{new}+DIIS to help convergence is a good candidate for an efficient black-box SCF less sensitive to the initial starting point (see Table 2.5).

Throughout the next sections, qualitative convergence results are tagged with the following convention:

- ⧸ non-convergence: the energies of the iterates oscillate above the ground state energy by at least 10^{-2} Ha and the residual does not go to zero. In many cases, the oscillations occur between 1 and 100 Ha above the ground state energy;
- ⧸ stagnation or small-amplitude oscillations : the algorithm stalls or the iterates display small-amplitude oscillations while the residual is small but not small enough in the sense that the limit

values of the energy are 10^{-4} to 10^{-2} Ha higher than the ground state energy (or another local minimum)

- ☑ convergence to a local minimizer.

2.4.2 Basic SCF iterations

We first illustrate the limitations of the classical iteration functions $g_{\mathrm{A,B}}$, as defined in (2.3.3), and the relevance of the new iteration function g_{new} defined in (2.3.8), by analyzing the behavior of the corresponding basic SCF algorithms $x^{(k+1)} = g(x^{(k)})$ (without any stabilization/acceleration technique) on simple mono-atomic systems: an oxygen atom in the triplet state, Fe^{2+} and Fe^{3+} ions in high-spin configurations (respectively quintet and sextet states).

Recall that the function $g_{A,B}$ is computed by diagonalizing an effective Hamiltonian depending on the input ROHF state and *ad hoc* coefficients A_{tt} and B_{tt}, and constructing the output ROHF state using the *Aufbau* principle (see Section 2.3). The performance of the basic SCF algorithm $x^{(k+1)} = g_{A,B}(x^{(k)})$ is found to be very sensitive to the choice of the A_{tt} and B_{tt} coefficients; besides, no choice of coefficients provides consistent convergence for the three simple systems. In contrast, the basic fixed-point algorithm built upon the parameter-free iteration function g_{new} has been able to converge for the three systems. The results reported in Table 2.1 have been obtained with the double-zeta correlation-consistent Dunning's type basis set (cc-pVDZ) [Dun89] and the Hückel initial guess from `PySCF`. Qualitatively similar results have been obtained with the core initial guess and/or other basis sets (*e.g* 6-31G, pc-1).

Method	A_{tt}	B_{tt}	O (triplet)	Fe^{2+} (quintet)	Fe^{3+} (sextet)
Roothan	$(-\frac{1}{2}, \frac{1}{2}, \frac{3}{2})$	$(\frac{3}{2}, \frac{1}{2}, -\frac{1}{2})$	✓(17)		✓(45)
McWeeny and Diercksen	$(\frac{1}{3}, \frac{1}{3}, \frac{2}{3})$	$(\frac{2}{3}, \frac{1}{3}, \frac{1}{3})$	✓(13)		
Davidson	$(\frac{1}{2}, 1, 1)$	$(\frac{1}{2}, 0, 0)$			✓(12)
Guest and Saunders	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$	✓(11)		✓(22)
Binkley, Pople and Dobosh	$(\frac{1}{2}, 1, 0)$	$(\frac{1}{2}, 0, 1)$			✓(10)
Faegri and Manne	$(\frac{1}{2}, 1, \frac{1}{2})$	$(\frac{1}{2}, 0, \frac{1}{2})$			✓(11)
Euler equations	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$	$(\frac{1}{2}, 0, \frac{1}{2})$	✓(10)		
Canonical-ROHF I	$(\frac{2S+1}{2S}, 1, 1)$	$(-\frac{1}{2S}, 0, 0)$			✓(11)
Canonical-ROHF II	$(0, 0, -\frac{1}{2S})$	$(1, 1, \frac{2S+1}{2S})$	✓(20)		
g_{new} (2.3.8)	parameter free		✓(10)	✓(21)	✓(12)

Table 2.1 – Convergence of the basic fixed-point algorithm $x^{(k+1)} = g(x^{(k)})$ for the atomic systems O, Fe^{2+}, and Fe^{3+} (cc-pVDZ basis set, `PySCF` Hückel initial guess), for (i) the classical $g_{A,B}$ iteration functions (see Table I in Ref. [PD14]), and (ii) the g_{new} iteration function (this work). The table follows the conventions detailed in the introduction to Section 2.4. The number of iterations needed to reach convergence is specified when the algorithm happens to converge (chosen convergence criterion: the energy of the current iterate is at most 10^{-6} Ha above the ROHF ground state).

2.4.3 Stabilized and accelerated iteration schemes

Table 2.2 summarizes the benchmark systems considered in this section. They consist of organic molecules bearing aromatic moieties (such as pyridine or porphyrin), interacting with open-shell metallic ions (see Figure 2.3). These systems are representative of the complexity of open-shell calculations in quantum

chemistry as they contain transition metal ions with high spin in interaction with non trivial aromatic organic ligands [LMA18]. The combination of strong repulsion in the 3d shell of the metals together with the very delocalized character of the π system in these organic ligands can lead to SCF instabilities precisely because, according to the choice of the flavor of effective Hamiltonian used in the g_{AB} function, the *Aufbau* principle is not fulfilled in these systems. We have picked up both systems having space symmetries, such as pyridine–Cu^{2+} (C_s symmetry) and the Porphyrin model–Fe^{2+} (D_{4h} symmetry), and systems with slightly broken symmetry, such as Pyridine–Fe^{n+}. We infer the spin multiplicities $M = 2S+1$ of these systems (where S is the total spin) from the corresponding spin multiplicities of the metallic ions, following Hund's rule. In some cases, it is actually challenging to determine the spin multiplicity of the ground state (*e.g.* triplet or quintet), such as for the iron–porphyrin model system [LMA18]. We have performed some test calculations on a full Porphyrin–Fe^{2+} system (37 atoms, 269 basis functions for 6-31G), that yielded qualitatively similar results as for the Porphyrin model–Fe^{2+} system. For the sake of brevity, we do not report them here.

System	Number of atoms	N_d / N_s	Multiplicity (2s+1)	Basis	Number of basis functions
Pyridine – Cu^{2+}	12	34 / 1	2	6-31G	93
Pyridine – Cu^{2+}	12	34 / 1	2	cc-pVDZ	164
Pyridine – Fe^{2+}	12	31 / 4	5	6-31G	93
Pyridine – Fe^{2+}	12	31 / 4	5	cc-pVDZ	164
Pyridine – Fe^{3+}	12	30 / 5	6	6-31G	93
Pyridine – Fe^{3+}	12	30 / 5	6	cc-pVDZ	164
Porphyrin model – Fe^{2+}	29	66 / 4	5	6-31G	197
Porphyrin Fe^{2+}	37	90 / 4	5	6-31G	269

Table 2.2 – Benchmark systems used in Section 2.4.3.

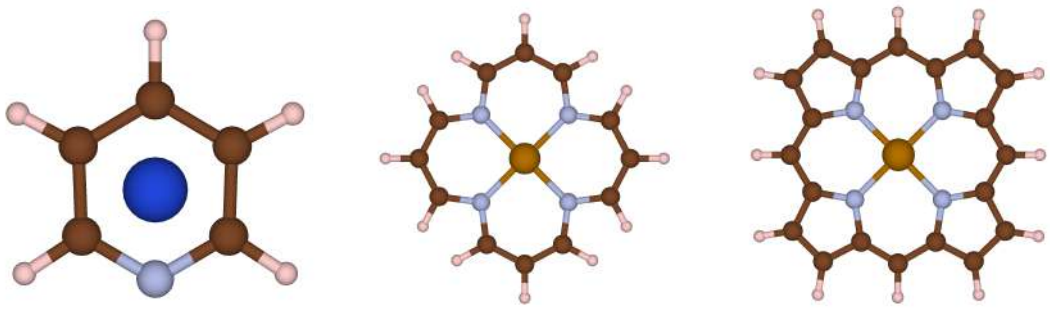


Figure 2.3 – Left: Pyridine - Cu^{2+}. Middle: Porphyrin model – Fe^{2+} taken from [LMA18]. Right: Porphyrin – Fe^{2+}. Figures have been generated with the Vesta software [MI08].

We have tested several families of basis sets representative of quantum chemistry calculations, *i.e.* the 6-31G and cc-pVDZ basis sets.

2.4.3.1 Global convergence regime

In this section, we analyze the ability of the various algorithms described in Section 2.3 to reach the vicinity of a local minimizer from the core initial guess. We consider that this is achieved if the energies of the iterates approach 0.1 Ha from the ROHF ground state energy. We compare the new algorithms proposed in this work with existing algorithms as implemented in GAMESS [Sch+93], namely the SOSCF algorithm and the DIIS schemes built from the iteration functions $g_{A,B}$ and residual function f (2.3.14). The results for the molecular systems in Table 2.2 in the 6-31G basis set are gathered in Table 2.3.

Algorithms based on $g_{A,B}$ iteration functions. We observe in the second and third columns of Table 2.3 that the results of the GAMESS implementation of DIIS are close to our DIIS implementation. For several choices of coupling coefficients A_{tt}, B_{tt}, the standard SCF+DIIS method fails to converge, and leads to oscillations. For the Pyridine$-Fe^{2+}$ and Fe^{3+} systems, and (respectively) for the Porphyrin model $-$ Fe^{2+} system, only three (resp. two) specific choices of A_{tt}, B_{tt} coefficients lead to convergence (notably Guest and Saunders and Roothaan). The results for the Pyridine$-Cu^{2+}$ system (not reported) are qualitatively the same (only Guest and Saunders, Euler, Roothaan and Canonical II choices of A_{tt}, B_{tt} coefficients lead to convergence of GAMESS DIIS or of our DIIS implementation).

Remarkably, forcing DIIS (resp. SOSCF) from the first iterations is needed in GAMESS, as the DIIS residual (resp. gradient norm) is initially much higher than the default threshold for DIIS (resp. SOSCF) activation. Let us underline that acceleration methods such as DIIS, are designed to accelerate local convergence (*i.e.* convergence when starting close enough to a local minimum). They are now well-understood mathematically in this setting [Chu+21]. In contrast, the fact that DIIS can stabilize SCF iterations starting from core initial guess in some cases (this is not always true) remains unexplained to our knowledge.

The SOSCF second-order method converges whatever the choice of A_{tt}, B_{tt} coefficients (except one, namely Canonical II, for the Pyridine$-Fe^{2+}$ system) from the core guess, although always in more than 200 iterations. Forcing DIIS (resp. SOSCF) from the first iterations is needed in GAMESS, as the DIIS residual (resp. gradient norm) is initially much higher than the default threshold for DIIS (resp. SOSCF) activation.

Algorithms based on the g_{new} iteration function. As shown in the last two columns of Table 2.3, the DIIS algorithm based on the iteration function g_{new} and the residual function f, as well as the ODA algorithm 3, provide robust schemes for all systems, except for the case of porphyrin model-Fe^{2+} with g_{new}+DIIS. Forcing a restart of the DIIS yields convergence in that case. Note that our current implementation was built as a proof-of-concept. Our method could potentially benefit from a more refined choice of preconditioning for the resolution of the subproblem (2.3.9), or from an adaptive depth DIIS approach, as introduced in [Chu+21], which we defer to future investigations.

For the other cases, the g_{new}+DIIS method is competitive with the converging standard SCF schemes in terms of iterations. The g_{new}+DIIS require more computational time than the $g_{\text{A,B}}$ standard SCFs, since each iteration involves the approximate resolution of the optimization problem (2.3.9). This is compensated by the absence of parameters in this method, and the convergence across almost all studied cases.

While the ODA method is very effective to reach the attraction basin of a local minimizer, it is very slow to converge to chemical accuracy. As the iterations approach a local minimum, the coefficient t_k of the ODA convex combination consistently equals 1, effectively reducing ODA to a simple SCF with g_{new} map and no DIIS. A good compromise is to transition from ODA to g_{new}+DIIS when sufficiently close to a local minimum (Table 2.5), mimicking the efficient EDIIS+DIIS method of [KSC02] in the RHF case. This transition can occur when the energy gradient reaches a specified tolerance, or when the ODA coefficient t_k takes the value 1 repeatedly. We chose the first option with threshold 10^{-1} in our implementation. Notably, applying ODA before g_{new}+DIIS seem to allow to target a lower local minimum, as appearing in appendix, Table 2.6.

	GAMESS [Sch+93]		This work		
	$g_{A,B}$ (2.3.3) - based methods			g_{new} (2.3.8) - based methods	
A_{tt}, B_{tt} (see Table 2.1)	SOSCF	DIIS	DIIS	DIIS	ODA
Pyridine–Fe^{2+}					
Guest and Saunders	✓(244;313)	✓(12;37)	✓(9;59)		
Roothaan	✓(212;263)	✓(28;109)	✓(13;145)		
Euler	✓(218;265)	✓(28;95)			
Mc Weeny	✓(204;254)				
Binkley	✓(262;352)			✓(8;55)	✓(4;+1000)
Faegri	✓(235;278)				
Davidson	✓(230;273)				
Canonical I	✓(262;329)				
Canonical II					
Pyridine–Fe^{3+}					
Guest and Saunders	✓(236;290)	✓(16;132)	✓(11;193)		
Roothaan	✓(221;263)	✓(19;72)	✓(17;116)		
Euler	✓(227;277)	✓(41;181)	✓(7;112)		
Mc Weeny	✓(217;273)				
Binkley	✓(216;272)			✓(8;54)	✓(8;+1000)
Faegri	✓(323;374)				
Davidson	✓(259;328)				
Canonical I	✓(246;317)				
Canonical II	✓(236;305)				
Porphyrin model–Fe^{2+}					
Guest and Saunders	✓(202;215)	✓(15;22)	✓(18;26)		
Roothaan	✓(203;219)	✓(21;34)	✓(34;49)		
Euler	✓(202;218)				
Mc Weeny	✓(202;219)				
Binkley	✓(203;213)				✓(10,+1000)
Faegri	✓(203;216)				
Davidson	✓(203;221)				
Canonical I	✓(203;212)				
Canonical II	✓(294;346)				

Table 2.3 – Convergence results starting from core initial guess (6-31G basis set). The table follows the conventions detailed in the introduction to Section 2.4. The DIIS residual function f is the one defined in (2.3.14). The DIIS maximum depth parameter $m_{\max}$ is fixed to 10 (default value in GAMESS). The notation $(n_{\text{approach}}; n_{\text{cv}})$ means that n_{approach} iterations are needed to reach 0.1 Ha accuracy, while n_{cv} iterations are necessary to reach microHartree accuracy.

	GAMESS [Sch+93]		This work		
	$g_{A,B}$ (2.3.3) - based methods		g_{new} (2.3.8) - based methods		
A_{tt}, B_{tt} (see Table 2.1)	SOSCF	DIIS	DIIS	DIIS	ODA
Pyridine–Fe^{2+}					
Guest and Saunders	✓(78)	✓(82)	✓(100)	✓(92)	✓(+1000)
Roothaan	✓(83)	✓(255)	✓(212)		
Euler	✓(40)	✓(59)	✓(68)		
McWeeny	✓(42)	✓(105)	✓(271)		
Binkley	✓(106)				
Faegri	✓(106)				
Davidson	✓(87)				
Canonical I	✓(88)				
Canonical II	✓(42)				
Pyridine–Fe^{3+}					
Guest and Saunders	✓(78)	✓(178)	✓(187)	✓(142)	✓(+1000)
Roothaan	✓(88)	✓(185)	✓(139)		
Euler	✓(50)				
McWeeny	✓(88)				
Binkley	✓(93)				
Faegri	✓(92)				
Davidson	✓(94)				
Canonical I	✓(95)				
Canonical II	✓(54)				
Porphyrin model–Fe^{2+}					
Guest and Saunders	✓(22)	✓(17)		✓(25)	✓(+1000)
Roothaan	✓(23)	✓(37)	✓(52)		
Euler	✓(29)	✓(25)	✓(72)		
Mc Weeny	✓(36)	✓(32)	✓(187)		
Binkley	✓(23)				
Faegri	✓(22)				
Davidson	✓(21)				
Canonical I	✓(24)				
Canonical II	✓(29)				

Table 2.4 – Convergence results starting an extended Hückel initial guess (6-31G basis set). The table follows the conventions detailed in the introduction to Section 2.4. The DIIS residual function f is the one defined in (2.3.14). The DIIS maximum depth parameter $m_{\max}$ is fixed to 10 (default value in GAMESS). The number of iterations in parentheses is the one needed to reach microHartree accuracy.

ODA + g_{new}-DIIS (2.3.8)			
Initial guess	Pyridine–Fe^{2+}	Pyridine–Fe^{3+}	Porphyrin model–Fe^{2+}
Core	✓(8,92)	✓(7,83)	✓(10,18)
Extended Hückel	✓(60)	✓(144)	✓(28)

Table 2.5 – Convergence results by starting with ODA iterations and switching to DIIS when the residual norm reaches a tolerance of 10^{-2}. The DIIS depth parameter $m_{\max}$ is fixed to 10 (default value in GAMESS). The number of iterations needed to reach convergence at microHartree precision is specified in parenthesis

2.4.3.2 Local convergence

We now compare the different algorithms starting from an extended Hückel initial guess, whose energy is about 1 to 2 Ha above the ground state for our test cases in the `GAMESS` implementation, and 20 to 60

Ha for PySCF. The difference between the two guesses is most notable for the Porphyrin model – Fe^{2+} system.

Algorithms based on $g_{A,B}$ iteration functions. Comparing the results in Tables 2.3 and 2.4, we observe that DIIS algorithms as implemented in GAMESS barely benefit from a better initial guess. Four different choices of A_{tt}, B_{tt} coefficients lead to convergence for Pyridine–Fe^{2+} and Porphyrin model – Fe^{2+} systems (two for Pyridine–Fe^{3+}) for DIIS.

Again, the SOSCF second-order method converges whatever the choice of A_{tt}, B_{tt} coefficients, in less than 100 iterations (thanks to the improved starting guess) except for two specific choices of coefficients (106 iterations needed with Binkley and Faegri coefficients, for Pyridine–Fe^{2+} system).

Algorithms based on the g_{new} iteration function. Both the DIIS and the ODA converge for all the four systems. As in the previous case, the ODA algorithm is very slow to converge to chemical accuracy and ODA followed by g_{new}+DIIS provides satisfactory convergence results.

2.5 Conclusion and perspectives

In this article, we have provided a geometrical derivation of the ROHF equations in the density matrix and molecular orbital formalisms. A fundamental aspect of that derivation is, for both formalisms, the characterization of the tangent space of the manifold of ROHF states at a critical point of the ROHF energy functional, as well as its orthogonal complement (for the Frobenius inner product). This analysis lead us to introduce a new, parameter-free, iteration function g_{new} (see Eq. (2.3.8)), as an alternative to Roothaan-like iteration functions $g_{A,B}$ based on the construction of a (non-physical) effective Hamiltonian $H_{A,B}$, where $A = (A_{dd}, A_{ss}, A_{dd})$ and $B = (B_{dd}, B_{ss}, B_{dd})$ collect six real empirical parameters. An important conceptual difference of the proposed new SCF algorithm with respect to previous works is that it is not based on the usual technique of diagonalization of Fock-like Hamiltonians which can lead to numerical instabilities when the *Aufbau* principle is not fulfilled. Thanks to its geometrical formulation, the present algorithm avoids the ambiguity of the orbital energies for which the Koopman theorem does not apply in the case of the ROHF framework.

The numerical results we report seem to indicate that the DIIS algorithm based on the usual g_{AB} framework with the Guest and Saunders ($A_{tt} = B_{tt} = \frac{1}{2}$) and Roothaan ($A_{tt} = \left(-\frac{1}{2}, \frac{1}{2}, \frac{3}{2}\right)$, $B_{tt} = \left(\frac{3}{2}, \frac{1}{2}, -\frac{1}{2}\right)$) iteration functions are quite robust and converge in a reasonable number of iterations, even when starting from the core initial guess. However, these observations, made on a small number of test cases (the ones reported in this paper plus a dozen of other challenging cases), do not guarantee that this algorithm will perform well for all systems and basis sets. Remarkably, the DIIS acceleration has to be enabled from the first iteration to guaranty convergence, which does not correspond to the default setting in most quantum chemistry codes, where DIIS is activated only when close enough to a local minimum.

The numerical results reported here based on our new parameter-free iteration function g_{new} are encouraging as the latter converge for all but a single systems tested in this work, which involves different open-shell transition metal ions interacting with aromatic ligands. The algorithms based on the parameter-free iteration function g_{new} may then provide a useful alternative to the $g_{A,B}$ iteration functions for very challenging systems. In particular, the ODA (involving g_{new}) seems to be extremely robust and efficient in the early iterations, to reach the attraction basin of a local minimizer. Using ODA for the first few iterations, followed by g_{new}+DIIS is a good candidate for a robust black-box SCF routine.

Acknowledgements

The authors would like to thank Michael F. Herbst, Antoine Levitt, Filippo Lipparini and Tommaso Nottoli for fruitful discussions.This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 810367).

Supplementary Material for reproductibility

The Supplementary Material contains the atomic coordinates of the benchmark systems studied in the article. The research code used to produce the numerical data is available at https://github.com/LaurentVidal95/ROHFToolkit.

Appendix A: First-order optimality conditions in MO formalism.

As seen in Section 2.2, the manifold of ROHF states in MO formalism is the quotient manifold

$$\mathcal{M}_{\mathrm{MO}} = \mathrm{St}(N_o;\mathbb{R}^{N_b})/(\mathcal{O}_{N_d}\times\mathcal{O}_{N_s}).$$

In DM formalism, $\mathcal{M}_{\mathrm{DM}}$ is embedded in $\mathcal{V}_{\mathrm{DM}}$, so that the tangent space of $\mathcal{M}_{\mathrm{DM}}$ at a point x can be directly identified with a subspace of $\mathcal{V}_{\mathrm{DM}}$ (see Fig. 2.2). Unfortunately, this is not the case for the quotient $\mathcal{M}_{\mathrm{MO}}$. Following [AMS08], a way around the problem (valid for general quotient manifolds) is to identify the tangent space $T_{[\![C_o]\!]}\mathcal{M}_{\mathrm{MO}}$ at given equivalence class $[\![C_o]\!]$ with a subspace of $T_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b})$, called the horizontal tangent space at C_o to the manifold $\mathrm{St}(N_o;\mathbb{R}^{N_b})$, and denoted $T^{\mathrm{h}}_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b})$. We therefore start by computing the expression of the tangent spaces $T_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b})$.

Tangent spaces of $\mathrm{St}(N_o;\mathbb{R}^{N_b})$. Let $C_o = (C_d, C_s) \in \mathrm{St}(N_o;\mathbb{R}^{N_b})$. The orthonormality condition ${C_o}^T C_o = I_{N_o}$ translates on C_d and C_s as $C_d^T C_d = I_{N_d}$, $C_s^T C_s = I_{N_s}$ and $C_d^T C_s = 0$. This writes at first order for a perturbation $z = (D_d|D_s) \in \mathbb{R}^{N_b\times N_o}$

$$\begin{aligned} C_d^T D_d + D_d^T C_d &= 0 \quad (1)\\ C_s^T D_s + D_s^T C_s &= 0 \quad (2)\\ C_d^T D_s + D_d^T C_s &= 0 \quad (3). \end{aligned}$$

Let C_v be the orthogonal complement of C_o such that $C = (C_d|C_s|C_v) \in \mathcal{O}_{N_b}$, and let us decompose D_d and D_s in the basis C:

$$\begin{aligned} D_d &= C_d(D_d^d)^T + C_s(D_d^s)^T + C_v(D_d^v)^T\\ D_s &= C_d(D_s^d)^T + C_s(D_s^s)^T + C_v(D_s^v)^T. \end{aligned} \tag{2.5.1}$$

Then from (1) and (2), there exists $A_d \in \mathbb{R}^{N_d\times N_d}_{\mathrm{skew}}$ and $A_s \in \mathbb{R}^{N_s\times N_s}_{\mathrm{skew}}$ such that $(D_d^d)^T = A_d$ and $(D_s^s)^T = A_s$. Now (3) writes

$$\begin{aligned} C_d^T(C_sA_s + C_d(D_s^d)^T + C_v(D_s^v)^T) + (-A_dC_d + D_d^sC_s^T + D_d^vC_v^T)C_s = D_d^s + (D_s^d)^T = 0\\ \Leftrightarrow D_d^s = -(D_s^d)^T. \end{aligned}$$

We deduce that for all $C_o = (C_d|C_s)$, the tangent space $T_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b})$ is made of all $z = (D_d|D_s) \in \mathbb{R}^{N_b\times N_o}$ such that

$$D_d = C_dA_d + C_sX^T + C_vY^T \quad \text{and} \quad D_s = -C_dX + C_sA_s + C_vZ^T \tag{2.5.2}$$

where $X \in \mathbb{R}^{N_d\times N_s}$, $Y \in \mathbb{R}^{N_d\times N_v}$ and $Z \in \mathbb{R}^{N_s\times N_v}$. This also abbreviate as

$$z = C\begin{pmatrix} A_d & -X & -Y\\ X^T & A_s & -Z\\ Y^T & Z^T & 0 \end{pmatrix}\begin{pmatrix} I_{N_o}\\ 0 \end{pmatrix}. \tag{2.5.3}$$

Horizontal tangent space. Now let $\pi : \mathrm{St}(N_o;\mathbb{R}^{N_b}) \to \mathcal{M}_{\mathrm{MO}}$ be the canonical projection on $\mathcal{M}_{\mathrm{MO}}$

$$\forall\, C_o \in \mathrm{St}(N_o;\mathbb{R}^{N_b}) \quad \pi(C_o) = [\![C_o]\!].$$

We define the vertical tangent space $T^{\mathrm{v}}_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b})$ at C_o as

$$T^{\mathrm{v}}_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b}) = T_{C_o}\pi^{-1}([\![C_o]\!]).$$

and the horizontal tangent space $T^{\mathrm{h}}_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b})$ as its orthogonal complement for the MO scalar product $\langle C_o|C_o'\rangle = \mathrm{Tr}(C_o^T C_o')$:

$$T_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b}) = T^{\mathrm{h}}_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b}) \oplus T^{\mathrm{v}}_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b}).$$

Intuitively, $T^{\mathrm{h}}_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b})$ only contains the directions of $T_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b})$ that allow escape the equivalence class $[\![C_o]\!]$, so that one has the important property [AMS08]

$$T_{[\![C_o]\!]}\mathcal{M}_{\mathrm{MO}} \simeq T^{\mathrm{h}}_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b}). \tag{2.5.4}$$

Following the same procedure as for $T_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b})$ and $T_x\mathcal{M}_{\mathrm{DM}}$, we can show that

$$T^{\mathrm{h}}_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b}) = \left\{(C_sX^T + C_vY^T| - C_dX + C_vZ^T) \text{ where } X\in\mathbb{R}^{N_d\times N_s},\, Y\in\mathbb{R}^{N_d\times N_v},\, Z\in\mathbb{R}^{N_s\times N_v}\right\} \tag{2.5.5}$$

$$T^{\mathrm{v}}_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b}) = \left\{(C_dA_d|C_sA_s) \text{ where } A_d\in\mathbb{R}^{N_d\times N_d}_{\mathrm{skew}},\, A_s\in\mathbb{R}^{N_s\times N_s}_{\mathrm{skew}}\right\} \tag{2.5.6}$$

First order optimality conditions. From (2.5.4) and (2.5.5) the first order optimality conditions write in MO formalism as

$$\nabla\mathcal{E}(C_{o*}) \in \mathcal{T}^{\mathrm{h}}_{C_{o*}}\mathrm{St}(N_o;\mathbb{R}^{N_b})^{\perp}. \tag{2.5.7}$$

A straigthforward computation shows that for all $C_o = (C_d|C_s)$, the ambiant gradient for the standard Frobenius scalar product writes

$$\nabla\mathcal{E}(C_o) = (4F_dC_d|4F_sC_s). \tag{2.5.8}$$

It now remains to find $\mathcal{T}^{\mathrm{h}}_{C_o}\mathcal{M}^{\perp}_{\mathrm{MO}}$. Once again consider $C_o = (C_d|C_s) \in \mathrm{St}(N_o;\mathbb{R}^{N_b})$ and C_v be such that $C = (C_d|C_s|C_v) \in \mathcal{O}_{N_b}$. For all $W = (W_d|W_s) \in \mathcal{V}_{\mathrm{MO}}$, decomposing W on C as in (2.5.1) yields

$$\begin{aligned}
W \in T^{\mathrm{h}}_{C_o}\mathrm{St}(N_o;\mathbb{R}^{N_b})^{\perp} &\Leftrightarrow \begin{cases} \mathrm{Tr}(X^T(W_d^s - (W_s^d)^T) + Y^TW_d^v + Z^TW_s^v) = 0, \\ \forall X\in\mathbb{R}^{N_d\times N_s},\, Y\in\mathbb{R}^{N_d\times N_v},\, Z\in\mathbb{R}^{N_s\times N_v} \end{cases} \\
&\Leftrightarrow \begin{cases} W_d^v = W_s^v = 0 \\ W_d^s = (W_s^d)^T. \end{cases} \\
&\Leftrightarrow \begin{cases} \exists M_d\in\mathbb{R}^{N_d\times N_d},\, M_s\in\mathbb{R}^{N_s\times N_s}, X\in\mathbb{R}^{N_d\times N_s} \\ \text{ such that } W = (C_dM_d^T + C_sX^T|C_dX + C_sM_s^T). \end{cases}
\end{aligned}$$

Using (2.5.7) and (2.5.8), there exists $M_d\in\mathbb{R}^{N_d\times N_d}$, $M_s\in\mathbb{R}^{N_s\times N_s}$ and $X\in\mathbb{R}^{N_d\times N_s}$ such that

$$4F_{d*}C_{d*} = C_{d*}M_d^T + C_{s*}X^T \text{ and } 4F_{s*}C_{s*} = C_{d*}X + C_{s*}M_s^T. \tag{2.5.9}$$

Multiplying both expression by C_d^T or C_s^T we obtain

$$M_d = 4C_{d*}^TF_{d*}C_{d*}, \quad M_s = 4C_{s*}^TF_{s*}C_{s*}, \quad X = 2C_{d*}^T(F_{d*}+F_{s*})C_{s^*} \tag{2.5.10}$$

so that the optimality conditions finally write

$$\begin{cases} F_{d*}C_{d*} = C_{d*}\left(C_{d*}^TF_{d*}C_{d*}\right) + \dfrac{1}{2}C_{s*}\left(C_{s*}^T(F_{d*}+F_{s*})C_{d*}\right) \\ F_{s*}C_{s*} = C_{s*}\left(C_{s*}^TF_{s*}C_{s*}\right) + \dfrac{1}{2}C_{d*}\left(C_{d*}^T(F_{d*}+F_{s*})C_{s*}\right). \end{cases} \tag{2.5.11}$$

Appendix B: List of local minima

We provide here the energies at convergence for each system, algorithm, and initial guess. Table 2.6 corresponds to the energies associated to the results of Table 2.3 while Table 2.7 corresponds to the energies associated to the results of Table 2.4. Finally, Table 2.8 corresponds to the energies reached by the ODA + g_{new}+DIIS method picture in Table 2.5

	GAMESS [Sch+93]		This work		
	$g_{A,B}$ (2.3.3) - based methods			g_{new} (2.3.8) - based methods	
A_{tt}, B_{tt} (see Table 2.1)	SOSCF	DIIS	DIIS	DIIS	ODA
Pyridine–Fe^{2+}					
Guest and Saunders	-1508.134652	-1508.134652	-1508.014203	-1508.014203	-1508.131670
Roothaan	-1508.016536	-1508.134040	-1508.131670		
Euler	-1508.016536	-1508.016536			
Mc Weeny	-1508.016536				
Binkley	-1508.134652				
Faegri	-1508.016536				
Davidson	-1508.016536				
Canonical I	-1508.134652				
Canonical II					
Pyridine–Fe^{3+}					
Guest and Saunders	-1507.414473	-1507.414091	-1507.411509	-1507.407360	-1507.411509
Roothaan	-1507.414473	-1507.343997	-1507.411889		
Euler	-1507.414473	-1507.414097	-1507.411509		
Mc Weeny	-1507.414473				
Binkley	-1507.414473				
Faegri	-1507.414473				
Davidson	-1507.414473				
Canonical I	-1507.414473				
Canonical II	-1507.414473				
Porphyrin model–Fe^{2+}					
Guest and Saunders	-1940.163309	-1940.513025	-1940.510151		-1940.510191
Roothaan	-1940.163309	-1940.335945	-1940.647646		
Euler	-1940.163309				
Mc Weeny	-1940.163309				
Binkley	-1939.977138				
Faegri	-1939.977138				
Davidson	-1939.977138				
Canonical I	-1940.075387				
Canonical II	-1940.267466				

Table 2.6 – Energies at convergence starting from a core initial guess with 6-31G basis set. The table follows the conventions detailed in the introduction to Section 2.4. The notation DIIS refers to a DIIS method using f as residual function. The DIIS depth parameter $m_{\max}$ is fixed to 10 (default value in GAMESS). All energies are expressed in Hartrees.

	GAMESS [Sch+93]		This work		
	$g_{A,B}$ (2.3.3) - based methods			g_{new} (2.3.8) - based methods	
A_{tt}, B_{tt} (see Table 2.1)	SOSCF	DIIS	DIIS	DIIS	ODA
Pyridine–Fe^{2+}					
Guest and Saunders	-1508.134652	-1508.013967	-1508.132280	-1508.131670	-1508.131670
Roothaan	-1508.016536	-1507.967145	-1508.131671		
Euler	-1508.134652	-1508.134652	-1508.002054		
Mc Weeny	-1508.134652	-1508.134652	-1508.132280		
Binkley	-1508.134652				
Faegri	-1508.134652				
Davidson	-1508.134652				
Canonical I	-1508.134652				
Canonical II	-1508.134652				
Pyridine–Fe^{3+}					
Guest and Saunders	-1507.414473	-1507.409935	-1507.411510	-1507.411509	-1507.411509
Roothaan	-1507.414473	-1507.409935	-1507.411889		
Euler	-1507.357499				
Mc Weeny	-1507.357499				
Binkley	-1507.414473				
Faegri	-1507.414473				
Davidson	-1507.414473				
Canonical I	-1507.414473				
Canonical II	-1507.357499				
Porphyrin model–Fe^{2+}					
Guest and Saunders	-1940.406548	-1940.513025		-1940.510191	-1940.510191
Roothaan	-1940.406548	-1940.335945	-1940.510151		
Euler	-1940.385615	-1940.513025	-1940.654531		
Mc Weeny	-1940.650207	-1940.513025	-1940.527432		
Binkley	-1940.513025				
Faegri	-1940.513025				
Davidson	-1939.977138				
Canonical I	-1940.513025				
Canonical II	-1940.650207				

Table 2.7 – Energies at convergence starting from an extended Hückel initial guess with 6-31G basis set. The table follows the conventions detailed in the introduction to Section 2.4. The notation DIIS refers to a DIIS method using f as residual function. The DIIS depth parameter $m_{\max}$ is fixed to 10 (default value in GAMESS). All energies are expressed in Hartrees.

ODA + g_{new}-DIIS (2.3.8)			
Initial guess	Pyridine–Fe^{2+}	Pyridine–Fe^{3+}	Porphyrin model–Fe^{2+}
Core	-1508.131670	-1507.411509	-1940.510191
Extended Hückel	-1508.131670	-1507.411509	-1940.510191

Table 2.8 – Energies at convergence obtained with a few iterations of ODA, followed by g_{new}+DIIS. The algorithm transitions from ODA to DIIS when the residual norm reaches a tolerance of 10^{-2}. The DIIS depth parameter $m_{\max}$ is fixed to 10 (default value in GAMESS). All energies are expressed in Hartrees.

Chapter 3

Optimization of atomic orbital basis sets

This chapter has been published in the proceeding [LV1]:

Eric Cancès, Geneviève Dusson, Gaspard Kemlin, and Laurent Vidal. "On basis set optimisation in quantum chemistry". In: ESAIM: Proceedings and Surveys *73 (2023), pp. 107–129*

Abstract In this chapter we propose general criteria to construct optimal atomic centered basis sets in quantum chemistry. We focus in particular on two criteria, one based on the ground-state one-body density matrix of the system and the other based on the ground-state energy. The performance of these two criteria are then numerically tested and compared on a parametrized eigenvalue problem, which corresponds to a one-dimensional toy version of the ground-state dissociation of a diatomic molecule.

Contents

3.1 Introduction

In quantum chemistry, a central problem is the computation of the electronic ground-state (GS) of a given molecular system. For many-electron systems, it is not possible to solve the N-body Schrödinger equations and most calculations are thus based on variational (e.g. Hartree–Fock) or non-variational (e.g. coupled cluster) approximations of the latter, or on Kohn–Sham density functional theory (DFT). For all these models, the continuous equations (e.g. a nonlinear elliptic eigenvalue problem in the Hartree–Fock or Kohn–Sham settings) are discretized into a finite-dimensional approximation space. Approximation spaces constructed from atomic orbitals (AO) basis sets [HJO14; Ols21] have many advantages and are therefore the most common choice in the quantum chemistry community. An AO basis set consists of a collection of functions $\chi = (\chi_\mu^z)_{z\in\mathrm{CE},\, 1\leqslant\mu\leqslant n_z}$ where CE is a set of atomic numbers (e.g. $\mathrm{CE} = \{1,\dots,92\}$ for the natural chemical elements of the periodic table), n_z a positive integer depending on the electronic shell-structure of the chemical element with atomic number z, and $\chi_\mu^z \in H^1(\mathbb{R}^3)$ a fast decaying function centered at the origin called an atomic orbital. Consider an atomic configuration ω consisting of M nuclei with nuclear charges $z_1,\dots,z_M$ (in atomic units) and positions $\mathbf{R}_1,\dots,\mathbf{R}_M$ in the three dimensional physical space. If the AO basis set χ is chosen by the user, the (spatial component of the) one-electron finite-dimensional space in which the chosen electronic structure model of a molecular system with atomic configuration ω is discretized is

$$\mathcal{X}_\omega \coloneqq \mathrm{span}(\chi_1^{z_1}(\cdot-\mathbf{R}_1),\dots,\chi_{n_{z_1}}^{z_1}(\cdot-\mathbf{R}_1),\dots,\chi_1^{z_M}(\cdot-\mathbf{R}_M),\dots,\chi_{n_{z_M}}^{z_M}(\cdot-\mathbf{R}_M)).$$

The accuracy of the approximation therefore crucially depends on the quality of the AO basis set. The main advantage of AO basis sets is that only a small number of AO per atoms (typically a dozen) are necessary to obtain a relatively accurate result on most quantities of interest. This is in sharp contrast with standard discretization methods used in the simulation of partial differential equations such as finite-element methods. To make connection with discretization methods used in mechanical and electrical engineering, AO basis set discretization methods can be considered as spectral methods [Can+07], and share common features with the modal synthesis method [ICM96b, Chapter 7], [ICM96a]. A drawback of AO basis sets is that conditioning quickly blows up when increasing the size of the basis by including polarization and diffuse basis functions, a problem known as overcompleteness [Löw70]. The numerical errors due to this large condition number can deteriorate the accuracy of the computed solutions and/or significantly increase computational times. AO basis sets can therefore not be systematically improved in a straightforward way.

In the early days, AOs were Slater functions [Sla30], with exponential decay and a cusp at the origin. It was then realized by Boys [BE50] in 1950 that it was much more efficient from a computational viewpoint to use Gaussian-type orbitals (GTO), that are linear combinations of polynomials times Gaussian functions. Indeed the multi-center overlap, kinetic and Coulomb integrals

$$\int_{\mathbb{R}^3} \chi_i^{z_a}(\mathbf{r}-\mathbf{R}_a)\,\chi_j^{z_b}(\mathbf{r}-\mathbf{R}_b)\,\mathrm{d}\mathbf{r}, \qquad \int_{\mathbb{R}^3} \nabla\chi_i^{z_a}(\mathbf{r}-\mathbf{R}_a)\cdot\nabla\chi_j^{z_b}(\mathbf{r}-\mathbf{R}_b)\,\mathrm{d}\mathbf{r},$$
$$\int_{\mathbb{R}^3} \frac{\chi_i^{z_a}(\mathbf{r}-\mathbf{R}_a)\,\chi_j^{z_b}(\mathbf{r}-\mathbf{R}_b)}{|\mathbf{r}-\mathbf{R}_k|}\,\mathrm{d}\mathbf{r}, \qquad \int_{\mathbb{R}^3\times\mathbb{R}^3} \frac{\chi_i^{z_a}(\mathbf{r}-\mathbf{R}_a)\,\chi_j^{z_b}(\mathbf{r}-\mathbf{R}_b)\,\chi_k^{z_c}(\mathbf{r}'-\mathbf{R}_c)\,\chi_\ell^{z_d}(\mathbf{r}'-\mathbf{R}_d)}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}',$$

arising in discretized electronic structure models can then be computed analytically by means of explicit calculations and recursion formulas.

However, individual Gaussian function poorly describes the cusps of the bound states electronic wave-functions at nuclear positions. *Contracted* Gaussians [McW50], that are linear combinations of Gaussians with different variances, were quickly introduced as they overcome this deficiency. Several classes of GTO basis sets have been proposed since the 50's: STO-ng basis sets [HSP69] were built as the contraction of n Gaussians that fit Clementi STO SCF AOs in an L^2 least-squares sense [Ste69]. It was quickly realized that better GTO basis sets could be obtained by minimizing atomic Hartree–Fock ground state energy. This approach led to the split-valence basis sets (e.g. 6-31G) with core and valence orbitals being approximated differently, developed by Pople et al. [BPH80]. Basis sets better suited for correlated electronic structure methods were then introduced, notably Atomic Natural Orbitals (ANO) [AT87] and Dunning basis sets [Dun89]. ANO basis sets are built by selecting occupied and virtual orbitals from Hartree–Fock and natural orbitals from correlated computations of atomic systems. Dunning bases provide a (finite) hierarchy of

bases obtained by consistently increasing the number of basis functions corresponding to different angular momenta. This optimization strategy yields the so-called correlation consistent cc-pVXZ basis sets, which are, with their *augmented* version, still commonly used nowadays.

Mathematical studies proving convergence rates or proposing systematic enrichment of GTO basis sets are so far quite limited. The approximability of solutions to electronic structure problems by Gaussian functions was studied in [Kut94], and later on in [SY17; Sha20]. An *a priori* error estimate on the approximation of Slater-type functions by Hermite and even-tempered Gaussian functions was derived in [BCS14]. A construction of Gaussian bases combined with wavelets was proposed on a one-dimensional toy model in [Pha17].

Commonly used Pople and Dunning GTO basis sets were optimized from atomic configuration energies and Hartree–Fock (and/or natural) atomic orbitals. Let us also mention [SPAS96; DCM20] where system specific optimization of AO bases has been investigated, however focusing on specific models (e.g. one electron periodic Hamiltonian) or optimization criteria. In this article, we propose a different approach, which is adaptable to any criterion one might be interested in, and involves molecular configurations. In Section 3.2, we introduce an abstract mathematical framework for the construction of optimal AO basis sets, based on the choices of

1. a set of admissible atomic configurations Ω;
2. a probability measure $\mathbb{P}$ on Ω;
3. a set of admissible AO basis sets $\mathcal{B}$;
4. a criterion $j(\chi, \omega)$ quantifying the error between the exact values of the quantities of interest when the system has atomic configuration $\omega \in \Omega$ – for the continuous model under consideration – and the ones obtained after discretization in the basis set $\chi \in \mathcal{B}$.

We also provide examples of possible choices of Ω, $\mathbb{P}$, $\mathcal{B}$, and j. As a proof of concept (Section 3.3), we apply this strategy to a simple toy model of a 1D homonuclear diatomic "molecule" with two 1D non-interacting spinless "electrons", which allows for extremely accurate reference calculations. Finally, we present numerical results in Section 3.4, where we compare the efficiency of the so-optimized AO bases compared to AO basis constructed from the occupied and unoccupied orbitals of the isolated "atom".

3.2 Optimization criteria

3.2.1 Abstract framework

We start by formulating the problem of basis set optimization in an abstract setting. The procedure can be divided into four steps.

First, we select the set Ω of all possible atomic configurations we are *a priori* interested in. For instance, depending on the foreseen applications, one can consider the set of all possible finite atomic configurations containing only hydrogen, nitrogen, carbon, and oxygen atoms, or the set of all possible periodic arrangements of chemical elements with less than 20 atoms per unit cell.

Second, we equip Ω with a probability measure $\mathbb{P}$ in order to allow for different configurations to have different weights in the optimization procedure. We will see later that our method requires the computation of very accurate reference solutions for all ω's in the support of $\mathbb{P}$. For practical reasons we therefore need to choose $\mathbb{P}$ of the form

$$\mathbb{P} = \sum_{n=1}^{N_c} \beta_n \delta_{\omega_n}, \tag{3.2.1}$$

where $\{\omega_1, \ldots, \omega_{N_c}\}$ is a finite (not too large) subset of Ω, δ_{ω_n} the Dirac mass at ω_n, and $\{\beta_1, \ldots, \beta_{N_c}\}$ are positive weights such that $\sum_{n=1}^{N_c} \beta_n = 1$. Assume that we are solely interested in reproducing accurately the dissociation curve of the HF (Hydrogen Fluoride) diatomic molecule. Then the set Ω should be identified with the interval $(0, +\infty)$, and a configuration $\omega \in \Omega$ with the H−F interatomic distance $R \in (0, +\infty)$, and $\mathbb{P}$ should be a probability measure on the interval $(0, +\infty)$. The selection of the ω_n's and β_n's can be done in various ways. An option is to

i) choose a continuous probability distribution $\mathbb{P}$ on $(0, +\infty)$ on the basis of chemical arguments, putting little weight on usually unimportant very small interatomic distances, more weight on interatomic distances close to the equilibrium distance ($d \simeq 0.92$ Å), sufficient cumulated weight on very large interatomic distance to correctly reproduce the dissociation energy, and more or less weight on intermediate interatomic distances in the range $2-8$ Å, depending on its importance for the targeted application;

ii) fix the number N_c according to the available computational means;

iii) compute the ω_n's and β_n's using e.g. quantization algorithms [MSS21] possibly based on optimal transport or clustering algorithms [Pag15].

Third, we select the set $\mathcal{B}$ of admissible AO basis sets. Restricting ourselves to the framework of GTOs, this can be done by choosing, for each chemical element arising in Ω, the number, symmetries, and contraction patterns of the Gaussian polynomials of the AO associated with this particular element. In this case, $\mathcal{B}$ has the geometry of a convex polyedron of $\mathbb{R}^d$.

Given an atomic configuration $\omega \in \Omega$ and an AO basis set $\chi \in \mathcal{B}$, we denote by χ_ω the one-electron finite-dimensional space obtained by using the AO basis set χ to describe the electronic structure of a molecular system with atomic configuration ω and an arbitrary number N of electrons.

The fourth and final step consists in choosing a criterion $j(\chi, \omega)$ quantifying the quality of the results obtained when using the basis set $\chi \in \mathcal{B}$ to compute the electronic structure of a molecular system with atomic configuration ω. The choice of the function

$$j : \mathcal{B} \times \Omega \to \mathbb{R}_+$$

depends on the quantity of interest (QoI) to the user, and on the respective weights of these quantities in the case of multicriteria analyses. For instance, if one focuses on the ground-state energy of the electrically neutral molecular system, a natural criterion is

$$j_E(\chi, \omega) \coloneqq |E_\omega - E_\omega^\chi|^2 , \tag{3.2.2}$$

where E_ω is the exact ground-state energy of the neutral system with atomic configuration ω for the chosen continuous model (e.g. Hartree–Fock, MCSCF, Kohn–Sham B3LYP...) and E_ω^χ the ground-state energy obtained with the model discretized in the AO basis set χ. Note that the absolute value of the difference is squared to make j_E differentiable. Another possible choice is to use a criterion based on the one-body reduced density matrices (1-RDM), for instance

$$j_A(\chi, \omega) \coloneqq -\mathrm{Tr}\left(\Pi_{\chi_\omega}^A \gamma_\omega \Pi_{\chi_\omega}^A A\right), \tag{3.2.3}$$

where A is a given self-adjoint, positive, definite operator on the one-particle state space $\mathcal{H}$ with form domain $Q(A)$, γ_ω the exact ground-state 1-RDM of the neutral system with atomic configuration ω for the chosen continuous model, and $\Pi_{\chi_\omega}^A : Q(A) \to \mathcal{X}_\omega \subset \mathcal{H}$ the orthogonal projector on $\mathcal{X}_\omega$ for the inner product A on $Q(A)$. If $A = I_\mathcal{H}$, then the $Q(A) = \mathcal{H}$ and $\Pi_{\chi_\omega}^A$ is the orthogonal projector on $\mathcal{X}_\omega$ for the inner product of $\mathcal{H}$. If $A = (1 - \Delta)$, then $Q(A)$ is the Sobolev space $H^1(\mathbb{R}^3)$, and $\Pi_{\chi_\omega}^A$ is the orthogonal projector on $\mathcal{X}_\omega$ for the H^1-inner product. Diagonalizing γ_ω as

$$\gamma_\omega = \sum_j n_{\omega,j} |\psi_{\omega,j}\rangle\langle\psi_{\omega,j}|, \quad 0 \leqslant n_{\omega,j} \leqslant 1, \quad \langle\psi_{\omega,j}|\psi_{\omega,j'}\rangle = \delta_{jj'},$$

where the $n_{\omega,j}$'s are the natural occupation numbers (NON) and $\psi_{\omega,j}$'s the natural orbitals (NO) for the chosen continuous model of the neutral system with atomic configuration ω, it holds

$$j_A(\chi, \omega) = -\sum_j n_{\omega,j} \|\Pi_{\chi_\omega}^A \psi_{\omega,j}\|_{Q(A)}^2.$$

Minimizing $j_A(\chi, \omega)$ thus amounts to maximizing the NON-weighted sum of the $Q(A)$-norms of $Q(A)$-orthogonal projections of the NON on the discretization space $\mathcal{X}_\omega$. Other criteria may include errors on

molecular properties, or a weighted sum of several elementary criteria, each of them targeting a specific property. The criterion should be chosen according to the intended application.

The aggregated criterion to be optimized then reads as an integral over the configuration space Ω with respect to the probability measure $\mathbb{P}$

$$J(\chi) \coloneqq \int_\Omega j(\chi, \omega) \mathrm{d}\mathbb{P}(\omega), \tag{3.2.4}$$

and the problem of basis set optimization can be formulated as

$$\boxed{\text{find } \chi_0 \in \underset{\chi \in \mathcal{B}}{\operatorname{argmin}}\, J(\chi)}$$

In the following, J_E and J_A denote the evaluation of the criterion (3.2.4) with $j = j_E$ and $j = j_A$ respectively.

Remark 3.2.1 (Reference solutions). The evaluation of criteria J_E and J_A hinges on the knowledge of exact GS energy E_ω or 1-RDM γ_ω for ω in the support of $\mathbb{P}$. In practice, these data can be approximated by very accurate off-line reference computations for a small, wisely chosen, sample of configurations ω. This is the reason why the probability measure $\mathbb{P}$ can only be a finite weighted sum of Dirac distributions, as defined in (3.2.1).

3.3 Application to 1D toy model

In this section, we focus on a linear one-dimensional toy model, mimicking a homonuclear diatomic molecule.

3.3.1 Description of the model

Let us consider a system of two 1D point-like "nuclei" and two 1D spinless non- interacting quantum "electrons". The one-particle state space is then $\mathcal{H} = L^2(\mathbb{R})$ and the configuration space $\Omega = (0, +\infty)$. In this section, a configuration of Ω will be labelled by the positive real number $a > 0$ such that the nuclei are located at $-a$ and a. The one-particle Hamiltonian at configuration a then is

$$H_a = -\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + V_a, \tag{3.3.1}$$

where V_a models the nuclei-electron interaction. We choose V_a to be a double-well potential with minima at $-a$ and $+a$, defined by

$$\forall\, x \in \mathbb{R}, \quad V_a(x) = \frac{1}{8a^2 + 4}(x - a)^2(x + a)^2. \tag{3.3.2}$$

Several considerations led us to define the potential as such. First, V_a is designed so that i) each H_a admits a non-degenerate ground-state of energy E_a, and ii) the function $a \mapsto E_a$ has the shape of the ground-state dissociation curve of a homonuclear diatomic molecule with atoms at $-a$ and $+a$. Since the two "electrons" do not interact, the ground-state energy E_a and density matrices $\gamma_a \in \mathcal{G}_2$ are given by

$$E_a = \operatorname{Tr}(H_a \gamma_a) = \min_{\gamma \in \mathcal{G}_2} \operatorname{Tr}(H_a \gamma), \tag{3.3.3}$$

where

$$\mathcal{G}_2 \coloneqq \left\{ \gamma \in \mathcal{L}(L^2(\mathbb{R})),\ \gamma^2 = \gamma = \gamma^*,\ \operatorname{Tr}(\gamma) = 2 \right\},$$

$\mathcal{L}(L^2(\mathbb{R}))$ denoting the space of bounded linear operators on $L^2(\mathbb{R})$. The existence and uniqueness of the solution to problem (3.3.3) can be shown by elementary arguments of functional analysis and spectral theory that we do not detail here. Second, $V_0(x) = \frac{1}{4}x^4$ so that (3.3.1) corresponds to the quartic oscillator, for which we have reference numerical solutions (e.g. [Bli19]). Third, V_a behaves like $x^2/2$ around $\pm a$ for large values of a and $V_a(0) \sim a^4/8 \to +\infty$ when $a \to +\infty$. Therefore, in the limit $a \to +\infty$, problem (3.3.3) is similar to two decoupled quantum harmonic oscillators centered in $-a$ and $+a$ whose bound states are all explicitly known. For the sake of illustration, we display in Figure 3.1 the potential V_a for two different values of a.

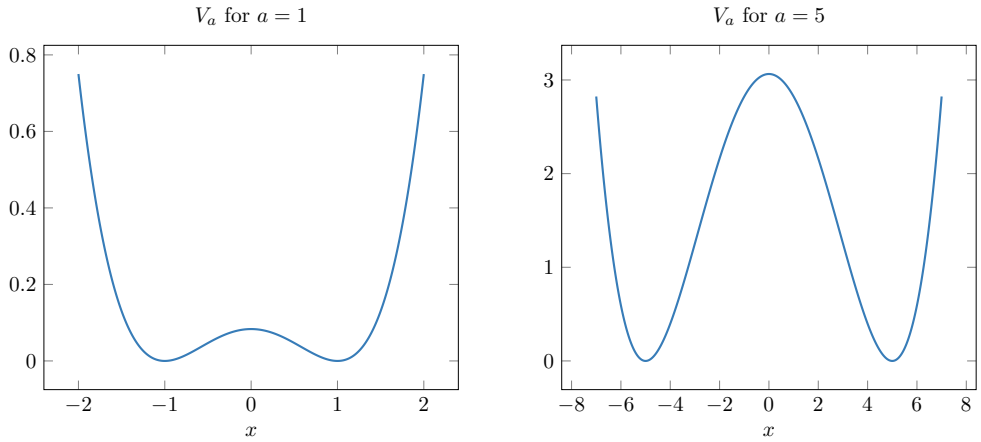


Figure 3.1 – $x \mapsto V_a(x)$ for $a = 1$ and $a = 5$.

In practice, it is convenient to compute γ_a and E_a from the lowest two eigenvalues $\lambda_{a,1} < \lambda_{a,2}$ of H_a and an associated pair $(\varphi_{a,1}, \varphi_{a,2})$ of orthonormal eigenvectors

$$\begin{cases} H_a \varphi_{a,i} = \lambda_{a,i} \varphi_{a,i}, \quad i = 1, 2 \\ \langle \varphi_{a,i} | \varphi_{a,j} \rangle = \delta_{ij}, \quad i, j = 1, 2, \end{cases} \tag{3.3.4}$$

$\langle \cdot | \cdot \rangle$ denoting the L^2 inner product. We indeed have

$$E_a = \lambda_{a,1} + \lambda_{a,2} \quad \text{and} \quad \gamma_a = |\varphi_{a,1}\rangle\langle\varphi_{a,1}| + |\varphi_{a,2}\rangle\langle\varphi_{a,2}|. \tag{3.3.5}$$

The evaluation of criteria J_A and J_E requires the computation of reference ground -state density matrices or energies, which amounts to find very accurate solutions of (3.3.4) for the configurations a_k in the support of the chosen atomic probability measure

$$\mathbb{P} = \sum_{n=1}^{N_c} \beta_n \delta_{a_n}, \quad 0 < a_1 < a_2 < \cdots < a_{N_c}, \quad \beta_n > 0, \quad \sum_{n=1}^{N_c} \beta_n = 1. \tag{3.3.6}$$

We chose to compute these reference data using a 3-point finite-difference (FD) scheme on a large enough interval $[-x_{\max}, x_{\max}]$ discretized into a uniform grid with N_g grid points:

$$x_j = -x_{\max} + j\delta x, \quad 1 \leqslant j \leqslant N_g, \quad \delta x = \frac{2x_{\max}}{N_g + 1}.$$

We then impose homogeneous Dirichlet boundary conditions at $-x_{\max}$ and $x_{\max}$. The parameter $x_{\max}$ is chosen such that $x_{\max} = a_{\max} + r_{\max}$, where $a_{\max} = \max(\mathrm{supp}(\mathbb{P}))$ and $r_{\max} > 0$ is is the radius beyond which atomic densities are zero at machine (double) precision. Note that this numerical scheme is independent of the configuration a. The FD discretization of problem (3.3.9) gives rise to the eigenvalue problem

$$\begin{cases} H_a^{\mathrm{FD}} \varphi_{a,i}^{FD} = \lambda_{a,i}^{\mathrm{FD}} \varphi_{a,i}^{\mathrm{FD}} \quad i = 1, 2 \\ \delta x \big(\varphi_{a,i}^{\mathrm{FD}}\big)^T \varphi_{a,j}^{\mathrm{FD}} = \delta_{ij}, \end{cases} \tag{3.3.7}$$

where $H_a^{\mathrm{FD}} \in \mathbb{R}_{\mathrm{sym}}^{N_g \times N_g}$ is a real symmetric matrix of size $N_g \times N_g$, and the reference data are obtained as

$$E_a^{\mathrm{FD}} = \lambda_{a,1}^{\mathrm{FD}} + \lambda_{a,2}^{\mathrm{FD}} \quad \text{and} \quad P_a^{\mathrm{FD}} = \varphi_{a,1}^{\mathrm{FD}} \big(\varphi_{a,1}^{\mathrm{FD}}\big)^T + \varphi_{a,2}^{\mathrm{FD}} \big(\varphi_{a,2}^{\mathrm{FD}}\big)^T \in \mathbb{R}_{\mathrm{sym}}^{N_g \times N_g}, \tag{3.3.8}$$

where P_a^{FD} can be interpreted as an approximation of the matrix $\gamma_a(x_j, x_{j'})$ containing the values of the (integral kernel of the) density matrix γ_a at the grid points.

3.3.2 Variational approximation in AO basis sets

For any given configuration $a \in \mathbb{R}_+$ and basis $\chi = \{\chi_\mu\}_{1\leqslant\mu\leqslant N_b} \in \mathcal{B}$, problem (3.3.4) is solved using a Galerkin method with the basis $\chi_a = \{\chi_{a,\mu}\}_{1\leqslant\mu\leqslant 2N_b}$ composed of two copies of the basis χ, the first one translated to a, and the second one to $-a$:

$$\chi_{a,1} = \chi_1(\cdot - a), \dots, \chi_{a,N_b} = \chi_{N_b}(\cdot - a),\ \chi_{a,N_b+1} = \chi_1(\cdot + a), \dots, \chi_{a,2N_b} = \chi_{N_b}(\cdot + a).$$

Defining the Hamiltonian matrix

$$H_a^\chi = \left(\left\langle \chi_{a,\mu} \middle| \left(-\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + V_a \right) \middle| \chi_{a,\nu} \right\rangle \right)_{1\leqslant\mu,\nu\leqslant 2N_b}$$

and the overlap matrix

$$S_a^\chi = \left(\langle \chi_{a,\mu} | \chi_{a,\nu} \rangle \right)_{1\leqslant\mu,\nu\leqslant 2N_b},$$

the discretization of problem (3.3.4) in the AO basis set χ then reads as the generalized eigenvalue problem: find $\left(C_{a,i}^\chi, \lambda_{a,i}^\chi\right) \in \mathbb{R}^{2N_b} \times \mathbb{R}$, $i = 1, 2$ such that

$$\begin{cases} H_a^\chi C_{a,i}^\chi = \lambda_{a,i}^\chi S_a^\chi C_{a,i}^\chi & i = 1, 2 \\ \left(C_{a,i}^\chi\right)^T S_a^\chi C_{a,j}^\chi = \delta_{ij}. \end{cases} \tag{3.3.9}$$

The approximation $\varphi_{a,i}^\chi$ of $\varphi_{a,i}$ in the AO basis set χ can then be recovered as the linear combination of atomic orbitals (LCAO)

$$\forall x \in \mathbb{R}, \quad \varphi_{a,i}^\chi(x) = \sum_{\mu=1}^{2N_b} [C_{a,i}^\chi]_\mu \chi_{a,\mu}(x). \tag{3.3.10}$$

One way to compare the LCAO ground-state 1-RDM to the reference FD solution P_a^{FD} is to simply evaluate the former at the grid points x_j, which gives rise to the matrix $P_a^\chi \in \mathbb{R}_{\mathrm{sym}}^{N_g \times N_g}$ with entries

$$[P_a^\chi]_{jj'} = \sum_{i=1}^{2} \varphi_{a,i}^\chi(x_j)\varphi_{a,i}^\chi(x_{j'}).$$

Due to numerical errors, the matrix P_a^χ is however not a rank-2 orthogonal projector. We therefore chose to follow a slightly different route (leading to very similar results). The finite difference grid gives a reference discrete setting in which any quantity of interest for any configuration and AO basis set can be expressed. For all a's, the basis χ_a is represented by a matrix $X_a \in \mathbb{R}^{N_g \times 2N_b}$. For any vectors $Y_1, Y_2 \in \mathbb{R}^{N_g}$, the discrete A inner product simply reads $\delta x Y_1^T A Y_2$ where the notation A stands for both the continuous inner product and its finite-difference discretization matrix. We denote by $\|\cdot\|_A$ the associated norm on $\mathbb{R}^{N_g}$. Solutions to (3.3.9) are then obtained by approximating respectively the Hamiltonian and overlap matrix by

$$H_a^\chi \simeq H_a^X \coloneqq \left(\delta x X_{a,\mu}^T H_a^{\mathrm{FD}} X_{a,\nu}\right)_{1\leqslant\mu,\nu\leqslant 2N_b}, \quad S_a^\chi \simeq S_a^X \coloneqq \left(\delta x X_{a,\mu}^T X_{a,\nu}\right)_{1\leqslant\mu,\nu\leqslant 2N_b},$$

and finding $(C_{a,i}^X, \lambda_{a,i}^X) \in \mathbb{R}^{2N_b} \times \mathbb{R}$, $i = 1, 2$, such that

$$\begin{cases} H_a^X C_{a,i}^X = \lambda_{a,i}^X S_a^X C_{a,i}^X, & i = 1, 2 \\ (C_{a,i}^X)^T S_a^X C_{a,j}^X = \delta_{ij}, & i, j = 1, 2, \end{cases} \tag{3.3.11}$$

from which we get the discrete approximations

$$\varphi_{a,i}^X = X_a C_{a,i}^X, \quad i = 1, 2. \tag{3.3.12}$$

Let us gather the coefficients $C_{a,i}^X$ into the $2N_b \times 2$ matrix $C_a^X = \left(C_{a,1}^X \,|\, C_{a,2}^X\right)$. The ground-state density matrix in the basis χ_a is approximated by

$$P_a^X = \varphi_{a,1}^X (\varphi_{a,1}^X)^T + \varphi_{a,2}^X (\varphi_{a,2}^X)^T = \left(X_a C_a^X\right)\left(X_a C_a^X\right)^T \in \mathbb{R}_{\mathrm{sym}}^{N_g \times N_g}. \tag{3.3.13}$$

3.3.3 Overcompleteness of Hermite Basis Sets

Before getting into basis set optimization, we introduce the following standard Hermite Basis Set (HBS), constructed from eigenfunctions of the quantum harmonic oscillator. Those functions are solutions to the eigenvalue problem $\left(-\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \frac{1}{2}x^2\right) h_n = \varepsilon_n h_n$ and are explicitly given by

$$h_n(x) = c_n p_n(x) \exp\left(-\frac{x^2}{2}\right), \quad \varepsilon_n = n + \frac{1}{2}, \quad n \in \mathbb{N}, \tag{3.3.14}$$

where p_n is the Hermite polynomial of degree n (with the same parity as n) and c_n a normalization constant such that $(h_n)_{n\in\mathbb{N}}$ forms an orthonormal basis of $L^2(\mathbb{R})$. The h_n's are the analogues of the standard atomic orbitals obtained by solving atomic electronic structure problems. Let us first consider the AO basis set made of the first N_b Hermite functions

$$\chi^{\mathrm{HBS}} = \{\chi_\mu^{\mathrm{HBS}}\}_{1\leqslant\mu\leqslant N_b} = \{h_n\}_{0\leqslant n\leqslant N_b-1}.$$

The overlap matrix for the configuration a then is of the form

$$S_a^{\chi^{\mathrm{HBS}}} = \begin{pmatrix} I_{N_b} & \Sigma_a \\ \Sigma_a^T & I_{N_b} \end{pmatrix} \quad \text{where} \quad \Sigma_a \coloneqq (\langle h_n(\cdot - a) | h_m(\cdot + a)\rangle)_{0\leqslant n,m\leqslant N_b-1}.$$

The matrix Σ_a corresponds to the overlap of functions that are localized at different atomic positions. It satisfies $\Sigma_a \simeq 0$ when a is large and $\Sigma_a \simeq I_{N_b}$ when a is close to 0, therefore causing conditioning issues on the overlap matrix $S_a^{\chi^{\mathrm{HBS}}}$, a phenomenon known as *overcompleteness*: when a is too small, the basis functions centered at $\pm a$ are almost equal, hence almost linearly dependent in the basis set. We illustrate this problem by plotting the condition number of the overlap matrix $S_a^{\chi^{\mathrm{HBS}}}$ for different values of a in Figure 3.2, which indeed blows up for small values of a. This is a well-known issue, and several methods have been proposed in the literature to cure this phenomenon, such as the standard canonical orthonormalization procedure [Löw70] or more recent work based on a Cholesky decomposition of the overlap matrix [Leh19b]. Such methods are however not directly related to the optimization procedure presented in this paper.

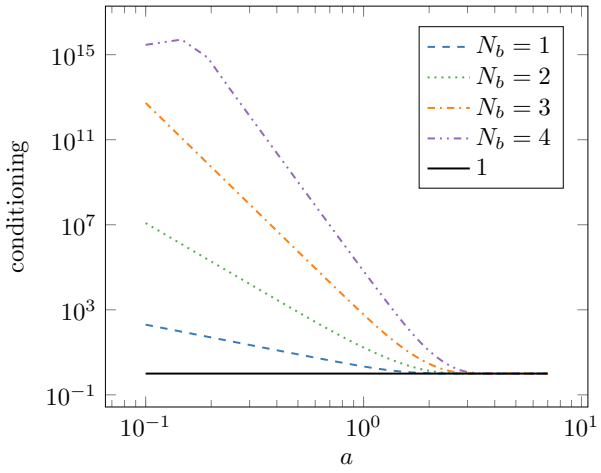


Figure 3.2 – Condition number of the HBS overlap matrix $S_a^{\chi^{\mathrm{HBS}}}$ for different values of a in log-log scale. The larger the basis set, the faster the condition number blows up for small values of a.

3.3.4 Practical computation of the criterion J_A and J_E

The rest of this section is dedicated to the rewriting and the computation of criteria J_A and J_E for our 1D model in the discrete setting.

3.3.4.1 Reference orthonormal basis

In order to avoid potential numerical stability issues, each of the N_b atomic orbital χ_μ is decomposed on a given truncated orthonormal basis of $L^2(\mathbb{R})$ of size $\mathcal{N}$ such that $N_b \ll \mathcal{N} \ll N_g$. We choose here the orthonormal basis introduced in (3.3.14). Hence, the matrix $X_a \in \mathbb{R}^{N_g \times 2N_b}$ is written as

$$X_a = B_a I_R, \tag{3.3.15}$$

with

$$B_a = \big(h_0(x_\cdot - a)|\cdots|h_{\mathcal{N}-1}(x_\cdot - a)|h_0(x_\cdot + a)|\cdots|h_{\mathcal{N}-1}(x_\cdot + a)\big) \in \mathbb{R}^{N_g \times 2\mathcal{N}},$$

and

$$I_R = \begin{pmatrix} R & 0 \\ 0 & R \end{pmatrix} \in \mathbb{R}^{(2\mathcal{N})\times(2N_b)}, \tag{3.3.16}$$

where $R \in \mathbb{R}^{\mathcal{N}\times N_b}$ gathers the coefficients of the atomic orbitals χ_μ in the truncated HBS orthonormal basis. Note that we have duplicated R in I_R as we consider the same basis at each position $\pm a$, but everything that follows can be easily adapted to the case where we would like to optimize the bases at each position separately (to deal with heteronuclear molecular systems for instance). We moreover impose that $R^T R = I_{N_b}$, so that the overlap matrix of X_a, denoted by $S(X_a)$, has the same form as in Section 3.3.3, that is

$$S(X_a) \coloneqq \delta x X_a^T X_a = \begin{pmatrix} I_{N_b} & \Sigma_a \\ \Sigma_a^T & I_{N_b} \end{pmatrix}, \tag{3.3.17}$$

where Σ_a is the overlap between functions localized at $+a$ and functions localized at $-a$. To avoid any issues arising from the conditioning of $S(X_a)$, the minimal sampled distance $a_{\min}$ should not be taken too small.

In the following, we detail the computation of each of the two criteria using the matrix R as the main variable. We will subsequently optimize the criteria J_A and J_E with respect to R to obtain optimal AO basis sets. In order to ease the reading of the following computations, every vector of $\mathbb{R}^{N_g}$ is rescaled by a factor $\sqrt{\delta x}$ so that for any given $Y_1, Y_2 \in \mathbb{R}^{N_g}$ the discrete A inner product simply reads $Y_1^T A Y_2$. The same holds for overlap matrices: with this convention, $S(X_a) = X_a^T X_a$. The output of the optimization is then scaled back to its former state by a factor $1/\sqrt{\delta x}$ to recover the original normalization.

3.3.4.2 Criterion J_A

Let $a \in \mathbb{R}_+$ be fixed and let $S^A(Y) = Y^T A Y$ denote the overlap matrix for the A-inner product of any rectangular matrix $Y \in \mathbb{R}^{N_g \times d}$. Since the columns of $X_a[S^A(X_a)]^{-\frac{1}{2}}$ are orthonormal for the A inner product, that is

$$\left(X_a[S^A(X_a)]^{-\frac{1}{2}}\right)^T A \left(X_a[S^A(X_a)]^{-\frac{1}{2}}\right) = I,$$

the projection $\Pi^A_{X_a}$ takes the simple form

$$\Pi^A_{X_a} = \left(X_a[S^A(X_a)]^{-\frac{1}{2}}\right)\left(X_a[S^A(X_a)]^{-\frac{1}{2}}\right)^T A = X_a[S^A(X_a)]^{-1}X_a^T A. \tag{3.3.18}$$

Hence, using the cyclicity of the trace and definitions (3.2.3), (3.3.8) and (3.3.18), one has

$$\begin{aligned} j_A(\chi, a) &\simeq -\mathrm{Tr}\left(P_a^{\mathrm{FD}}\,\Pi^A_{X_a} A \Pi^A_{X_a}\right) \\ &= -\mathrm{Tr}\left(P_a^{\mathrm{FD}} \times (AB_aI_R)[S^A(B_aI_R)]^{-1}(AB_aI_R)^T\right) \\ &= -\mathrm{Tr}\left(M_A^{\mathrm{offline}}(a) I_R [S^A(B_aI_R)]^{-1} I_R^T\right), \end{aligned}$$

where we have collected in the last expression all matrices independent of R into the matrix

$$M_A^{\mathrm{offline}}(a) = (AB_a)^T P_a^{\mathrm{FD}} AB_a \in \mathbb{R}^{2\mathcal{N}\times 2\mathcal{N}}. \tag{3.3.19}$$

Then, using the probability measure $\mathbb{P}$ in (3.3.6), we get

$$J_A(R) = -\int_\Omega \mathrm{Tr}\left(M_A^{\mathrm{offline}}(a) I_R [S^A(B_aI_R)]^{-1} I_R^T\right) \mathrm{d}\mathbb{P}(a) = -\sum_{n=1}^{N_c} \beta_n \mathrm{Tr}\left(M_A^{\mathrm{offline}}(a_n) I_R [S^A(B_{a_n}I_R)]^{-1} I_R^T\right)$$

and the optimization problem finally writes, with unknown $R \in \mathbb{R}^{\mathcal{N}\times N_b}$ and for a given inner product A

$$\boxed{\text{Find } R_{\text{opt}} \in \underset{R\in\mathbb{R}^{\mathcal{N}\times N_b},\ R^TR=I_{N_b}}{\operatorname{argmin}} J_A(R)} \tag{3.3.20}$$

3.3.4.3 Criterion J_E

Let again $a \in \mathbb{R}_+$ be fixed. We denote by

$$G(N_g, 2) \coloneqq \{P \in \mathbb{R}^{N_g\times N_g} \,|\, P^2 = P = P^T, \mathrm{Tr}(P) = 2\}$$

the discrete counterpart of the Grassmann manifold $\mathcal{G}_2$, and write E_a^R (resp. H_a^R) instead of E_a^χ (resp. H_a^χ), so that the dependence in the matrix R appears explicitly. Equation (3.3.3) reads in the discrete setting

$$\begin{aligned} E_a^R = \min_{P\in\, G(N_g,2)} \mathrm{Tr}\left(PH_a^R\right) &= \min_{\substack{C\in\mathbb{R}^{2N_b\times 2}\\ (C)^TS(B_aI_R)C=I_2}} \mathrm{Tr}\left(CC^T \times (B_aI_R)^T H_a^{\mathrm{FD}}(B_aI_R)\right) \\ &= \mathrm{Tr}\left(C_a^R(C_a^R)^T \times I_R^T M_E^{\text{offline}}(a) I_R\right) \end{aligned} \tag{3.3.21}$$

where, as for the previous case, all matrices independent of R have been gathered in the matrix

$$M_E^{\text{offline}}(a) = B_a^T H_a^{\mathrm{FD}} B_a, \tag{3.3.22}$$

and the matrix C_a^R is solution to the minimization problem

$$\min_{\substack{C^R\in\mathbb{R}^{2N_b\times 2}\\ (C^R)^TS(B_aI_R)C^R=I_2}} \mathrm{Tr}\left(C^R(C^R)^T \times I_R^T M_E^{\text{offline}}(a) I_R\right) \tag{3.3.23}$$

and is given in practice by $C_a^R = [S(B_aI_R)]^{-\frac{1}{2}}\,(u_{a,1}|u_{a,2})$ where $u_{a,1}$ and $u_{a,2}$ are orthonormal eigenvectors associated to the lowest two eigenvalues of

$$[S(B_aI_R)]^{-\frac{1}{2}}\, I_R M_E^{\text{offline}}(a) I_R^T\, [S(B_aI_R)]^{-\frac{1}{2}}\,.$$

From (3.3.6) and (3.3.21), one can compute

$$J_E(R) = \int_\Omega \left|E_a^{\mathrm{FD}} - E_a^R\right|^2 \mathrm{d}\mathbb{P}(a) = \sum_{n=1}^{N_c} \beta_n \left|E_{a_n}^{\mathrm{FD}} - E_{a_n}^R\right|^2$$

and the optimization problem reads

$$\boxed{\text{Find } R_{\text{opt}} \in \underset{R\in\mathbb{R}^{\mathcal{N}\times N_b},\ R^TR=I_{N_b}}{\operatorname{argmin}} J_E(R)} \tag{3.3.24}$$

3.4 Numerical results

3.4.1 Numerical setting and first results

Problems (3.3.20) and (3.3.24) are solved by direct minimization algorithms over the Stiefel manifold [AMS08]

$$\mathrm{St}(\mathcal{N}, N_b) = \{R \in \mathbb{R}^{\mathcal{N}\times N_b} \,|\, R^TR = I_{N_b}\}.$$

The explicit computation of the gradients of J_A and J_E with respect to R is detailed in the Appendix. We used a L–BFGS algorithm (with tolerance 10^{-7} on the norm of the projected gradient), as implemented in the *Optim.jl* package [MR18] in the *Julia* language [Bez+17]. As initial guess, we picked the first N_b Hermite functions introduced in Section 3.3.3.

In this subsection, we choose a probability distribution $\mathbb{P}$ supported in the interval $\mathcal{I} = [1.5, 5]$ so as to retain the physics of interest that takes place around the equilibrium configuration $a_0 \simeq 1.925$ and all

the way to dissociation. In particular $a_{\min} = 1.5$ is taken sufficiently large to avoid the conditioning issues on the overlap matrices described in Section 3.3.3. More precisely, all the results in this subsection are obtained with the probability

$$\mathbb{P} = \frac{1}{10}\sum_{n=1}^{10} \delta_{a_n} \quad \text{with } a_n = 1.5 + (n-1)\frac{3.5}{9}. \tag{3.4.1}$$

The quantities $M_A^{\text{offline}}(a_n)$ and $M_E^{\text{offline}}(a_n)$ are computed offline beforehand. We will discuss this choice and consider other probability measures $\mathbb{P}$ in Sections 3.4.2.2 and 3.4.2.3.

The finite-difference grid is a uniform grid on the interval $[-20, 20]$ discretized into $N_g = 1999$ points ($\delta x = 0.02$). Finally, we decompose the basis functions to be optimized in the HBS $\{h_n\}_{0\leqslant n\leqslant \mathcal{N}-1}$ of $L^2(\mathbb{R})$ of size $\mathcal{N} = 10$. Regarding the choice of the inner product for the first criterion J_A, we used the standard $L^2(\mathbb{R})$ and the $H^1(\mathbb{R})$ inner products, and denoted by J_{L^2} and J_{H^1} the corresponding. This translates at the discrete level by choosing $A = I_{N_g}$ for J_{L^2} and $A = I_{N_g} - \Delta$ for J_{H^1} where Δ is the 3-point finite-difference discretization matrix of the 1D Laplace operator. Once obtained, the optimal bases are used to solve the variational problem (3.3.11) on a much finer sampling of $\mathcal{I}$ and their accuracy is compared to the HBS. The code performing the simulations and plotting the results is available online[1]. Also, for the sake of clarity in the plots, $\widetilde{E}_a$ (resp. $\widetilde{\rho}_a$) denotes the GS energy (resp. the density) in the configuration a with a given basis (specified by the context) and E_a (resp. ρ_a) stands for the reference energy (resp. density) on the finite difference grid. Note that we write HBS for the (nonoptimized) Hermite basis set, and L^2-OBS, H^1-OBS or E-OBS for optimized basis sets with respect to the criterion J_{L^2}, J_{H^1}, or J_E.

Figure 3.3 displays the dissociation curve and the energy difference on the interval $\mathcal{I}$ for different values of N_b, the size of the AO basis set. For $N_b = 1$, i.e. only one basis function at $\pm a$, criterion J_E shows better performance than the criterion J_A, regardless of the choice of norm to perform the projections. It also very closely matches the accuracy of the standard HBS. When N_b becomes larger however, the different criteria behave in a similar fashion and we observe that they approach the dissociation curve better than the Hermite basis. Comparing the values of criterion J_E for all bases, which directly measures the distance to the dissociation curve, we see in Table 3.1 that all optimized bases give an increased accuracy of roughly four orders of magnitude over the interval $\mathcal{I}$ for $N_b = 4$.

In Figure 3.4, we plot the density for a given value of a and the error on the density for different norms, with varying values of N_b. The error is plotted with respect to three different distances: the L^1-norm, which corresponds to the L^2-norm on eigenvectors, the H^1-norm of the error on the density, as it is common to compute the forces $\int_{\mathbb{R}} \rho \nabla_a V_a$ with good estimates on the H^{-1}-norm of $\nabla_a V_a$ (see e.g. [Can+21b]), and the distance

$$\|\nabla\sqrt{\rho_1} - \nabla\sqrt{\rho_2}\|_{L^2}$$

(recall that the von-Weizsäcker kinetic energy reads $\frac{1}{2}\int_{\mathbb{R}} |\nabla\sqrt{\rho}|^2$). We observe similar behaviors between these different distances. For $N_b = 1$, both bases obtained with the first criterion behave slightly better than the standard Hermite basis and the basis computed with the second criterion. For $N_b = 3$, we observe again that all optimal bases yield better accuracy than the Hermite basis. Table 3.1 gives the confirmation that each basis for a given criterion indeed performs better than the other bases for that particular criterion. As for dissociation curves, we read from the values of J_{L^2} and J_{H^1} that the optimized bases yield similar results for large N_b, all of them giving lower values than the HBS. Note that the optimal bases for criterion J_{L^2} and J_{H^1} give similar results for any number of basis functions N_b, so that the L^2 and H^1 norm optimizations seem equivalent.

In terms of computational time, first note that criterion J_{H^1} is always more expensive to compute than J_{L^2} as it requires additional matrix-vector products with the matrix A, this having noticeable impact on the computational time. Second, criterion J_E requires less off-line data as it only needs to be given the reference eigenvalues while criterion J_A requires the reference GS eigenvectors (or density matrices). In addition, the use of orthonormality constraints as detailed in appendix allows one to compute the gradient of J_E at very low cost. In turn, criterion J_E is more than twice faster to minimize than criterion J_{L^2} in our implementation.

Finally, for the sake of completeness, we plot in Figure 4.1 the different basis functions built with each criterion for different values of N_b, confirming again the previous observations that the optimal basis functions are quite close to the standard Hermite basis functions.

[1] https://github.com/gkemlin/1D_basis_optimization

The main conclusion of these observations is that, for N_b large enough, there is no real difference between the proposed criteria. Still, if the bases we built do not seem to be very different from the standard Hermite basis (Figure 4.1), building optimal bases allows to increase accuracy on the quantities of interest we focused on by one order of magnitude in average.

Value of J_{L^2} for the different basis sets

Basis	$N_b = 1$	$N_b = 2$	$N_b = 3$	$N_b = 4$
HBS	-7.40829	-7.70051	-7.74312	-7.77138
L^2-OBS	-7.43954	-7.76479	-7.77725	-7.77773
H^1-OBS	-7.43928	-7.76466	-7.77724	-7.77772
E-OBS	-7.39410	-7.76425	-7.77720	-7.77772

Value of J_{H^1} for the different basis sets

Basis	$N_b = 1$	$N_b = 2$	$N_b = 3$	$N_b = 4$
HBS	-10.5613	-11.0566	-11.1451	-11.2402
L^2-OBS	-10.6256	-11.2338	-11.2630	-11.2650
H^1-OBS	-10.6265	-11.2342	-11.2630	-11.2651
E-OBS	-10.5334	-11.2313	-11.2626	-11.2650

Value of J_E for the different basis sets

Basis	$N_b = 1$	$N_b = 2$	$N_b = 3$	$N_b = 4$
HBS	3.77956×10^{-2}	3.98301×10^{-3}	1.86537×10^{-3}	1.35309×10^{-4}
L^2-OBS	6.52016×10^{-2}	2.18282×10^{-4}	1.01365×10^{-6}	3.22260×10^{-8}
H^1-OBS	6.83537×10^{-2}	2.40548×10^{-4}	1.27251×10^{-6}	3.91885×10^{-8}
E-OBS	3.69610×10^{-2}	1.92087×10^{-4}	6.93394×10^{-7}	2.54014×10^{-8}

L–BFGS iterations

Basis	$N_b = 1$	$N_b = 2$	$N_b = 3$	$N_b = 4$
L^2-OBS	4	13	48	219
H^1-OBS	7	17	235	not converged after 500 it
E-OBS	6	19	52	134

Table 3.1 – (Top & Middle) Values of the different criteria for the HBS and optimal bases, for increasing values of N_b. (Bottom) Number of iterations of L–BFGS required for each criterion to achieve convergence up to requested tolerance (10^{-7} on the ℓ^2-norm of the gradient).

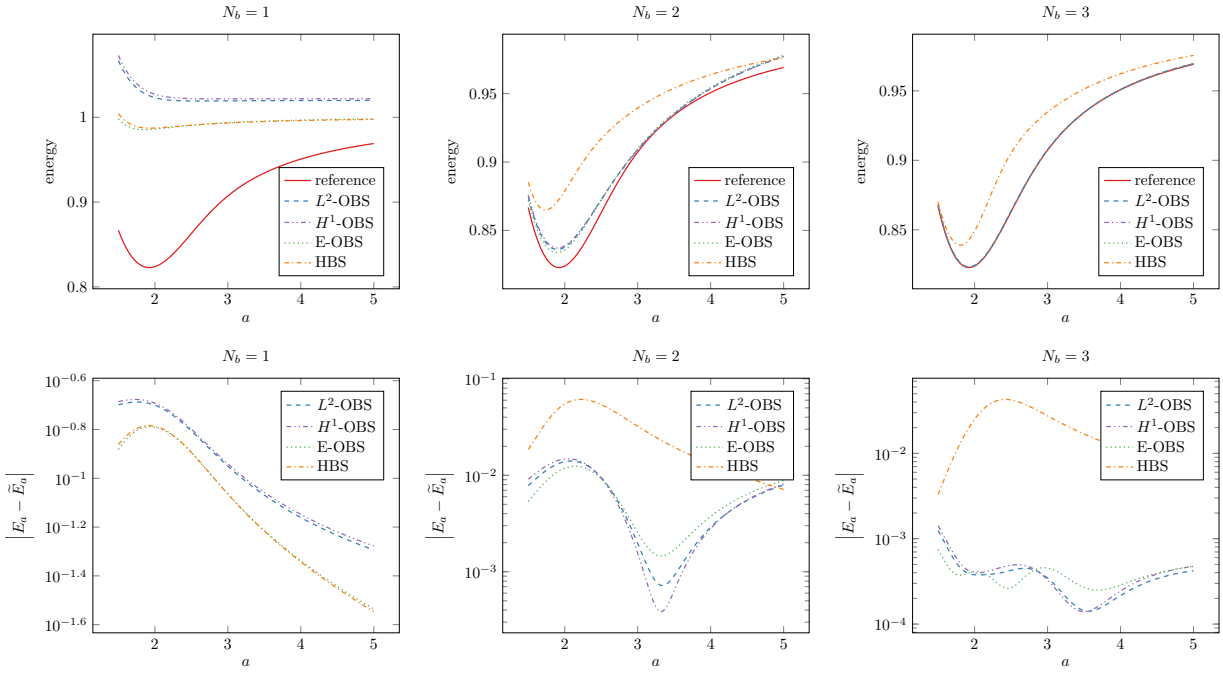


Figure 3.3 – Energies for the optimal bases obtained with the different criteria. (Top) Dissociation curve. (Bottom) Errors on the energy on the range of configuration $\mathcal{I} = [1.5, 5]$.

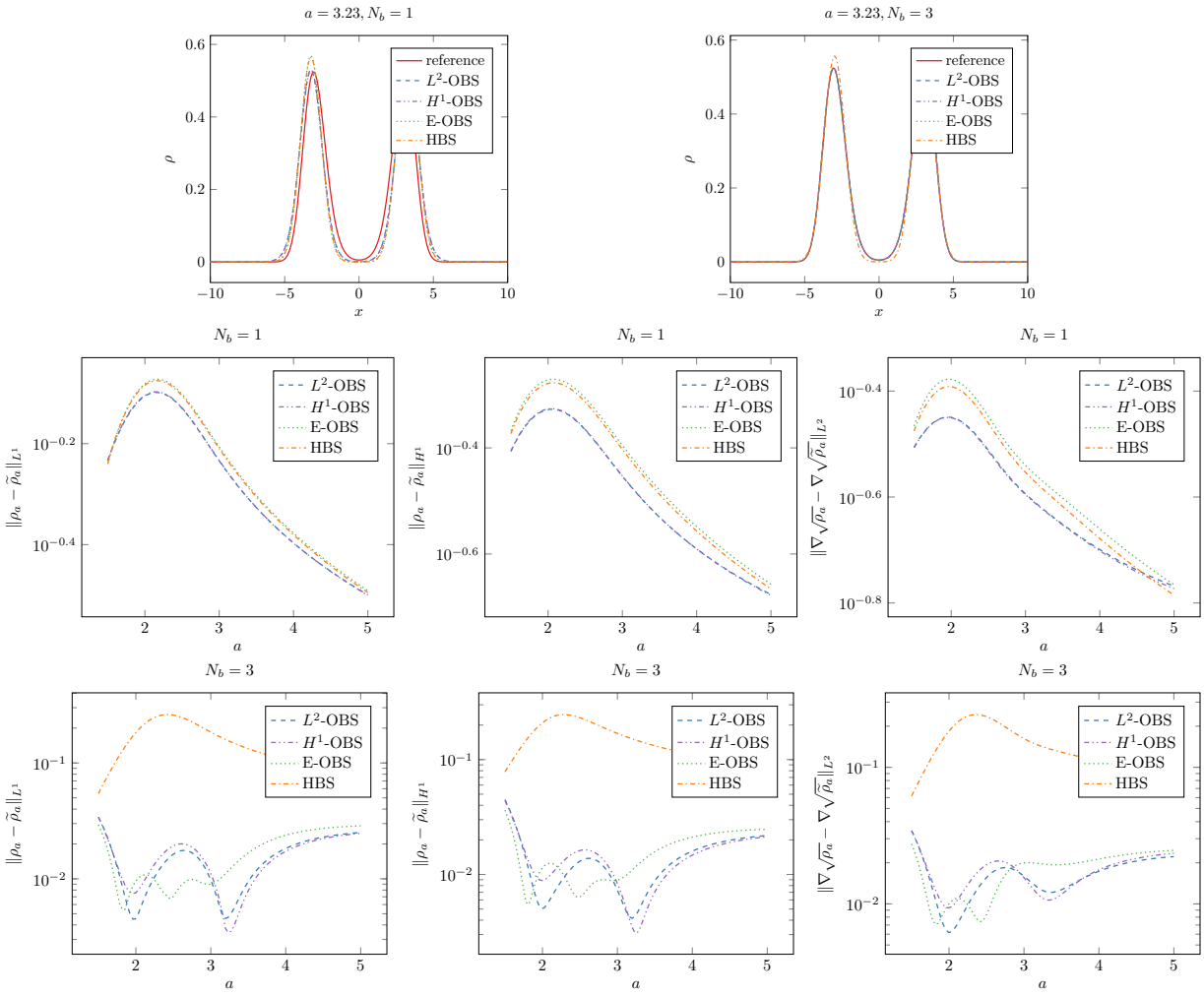


Figure 3.4 – (Top) Densities for the optimal bases obtained with the different criteria. (Middle) Errors on the density for different norms with $N_b = 1$. (Bottom) Error on the density for different norms with $N_b = 3$.

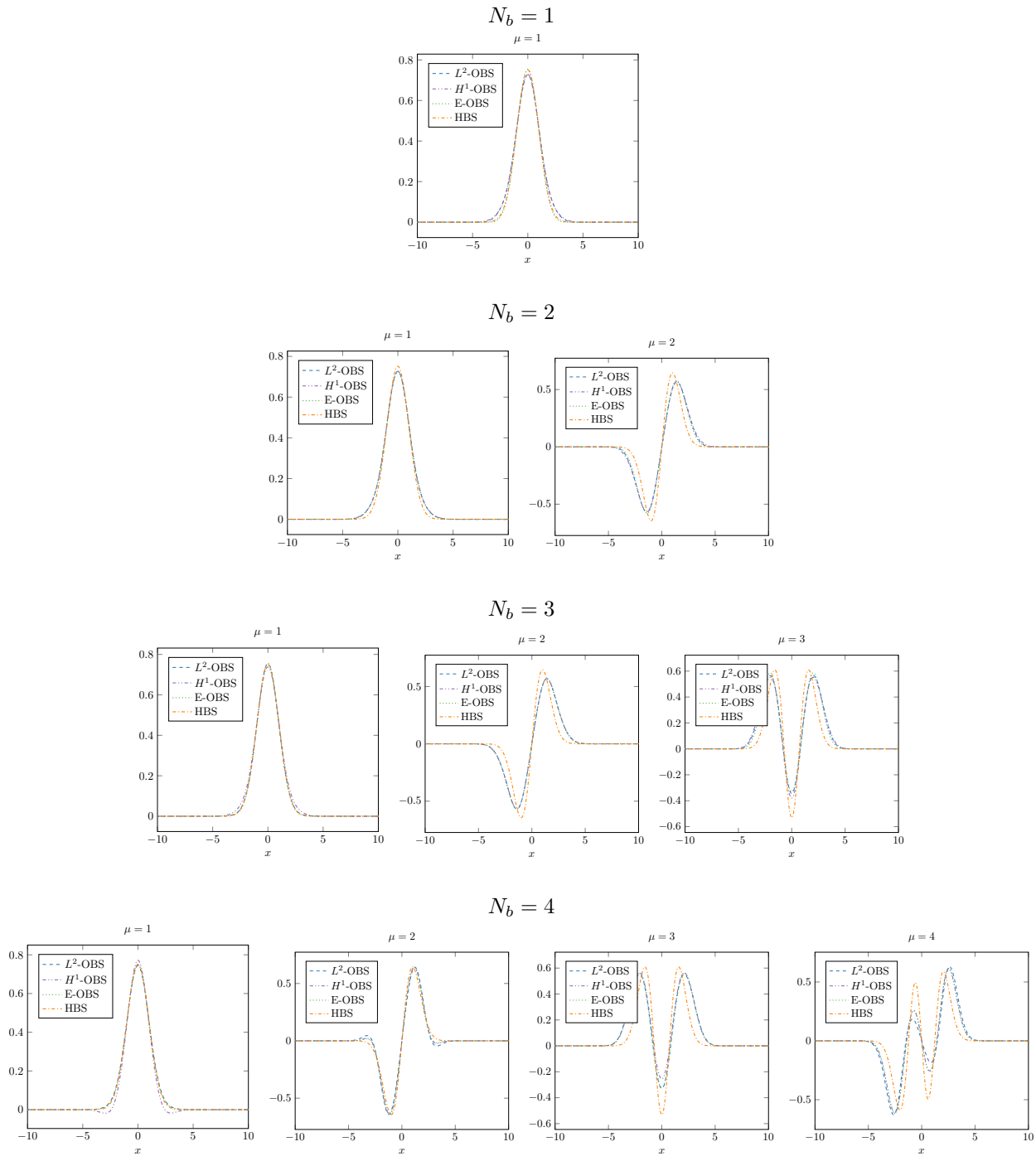


Figure 3.5 – Optimal basis functions for different criteria, each of them being optimized for different values of N_b.

3.4.2 Influence of numerical parameters

3.4.2.1 Random starting points

In Section 3.4.1, we used the first N_b Hermite functions as a starting point for the optimization procedures. We obtain the same solutions if we start from a random matrix R on the Stiefel manifold, in the sense that the optimal values reached for each criterion are the same, as well as the error plots. However, the L–BFGS algorithm requires more iterations to converge. The basis functions obtained from the optimization algorithms are different from those observed in Figure 4.1, but still span the same space as the variational solutions are equal.

3.4.2.2 Extrapolating the parameter space $\mathcal{I}$

In Section 3.4.1, we chose a probability measure $\mathbb{P}$ supported in the interval $[1.5, 5]$ in order to avoid conditioning issues. Indeed, taking smaller values of a results in the L–BFGS algorithm having convergence problems when N_b increases. This phenomenon was observed already for $N_b = 3$ or $N_b = 4$ when including $a = 1$ in the support of $\mathbb{P}$. In practice, this problem can be solved by using preconditioning or getting rid of overcompleteness by pre-processing the basis χ_a (e.g. selecting a smaller basis by filtering out the very small singular values of the original overlap matrix), but for brevity we will not elaborate further in this direction.

However, once we have computed optimal bases for a reasonable interval $\mathcal{I}$, it is possible to use these bases to solve the variational problem (3.3.9) and extrapolate the energy and the density to smaller values of a that are not in the set $\mathcal{I}$. The results are plotted in Figure 3.6. We notice that the quantities of interest are better approximated on $\mathcal{I} = [1.5, 5]$, but for smaller a's, there is no more gain in accuracy with respect to the standard HBS.

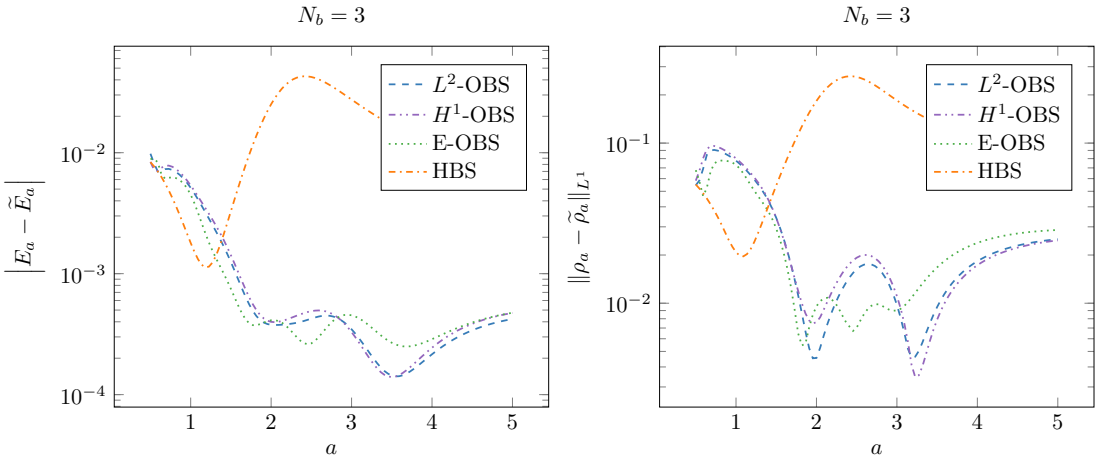


Figure 3.6 – Energy and densities error with extrapolation up to $a = 0.5$, with basis functions optimized on $\mathcal{I} = [1.5, 5]$.

3.4.2.3 Choice of the probability $\mathbb{P}$

The major drawback of our AO basis optimization lies in the necessity to compute very accurate reference solutions for all configurations in the support of $\mathbb{P}$. This is not an issue for our 1D toy model but it can be very time consuming for real systems if the support of $\mathbb{P}$ is too large. It is therefore crucial to reduce as much as possible the support of $\mathbb{P}$.

In this section, we study the influence of the probability measure $\mathbb{P}$ on the quality of the optimized bases. For simplicity, we restrict ourselves to uniform samplings of the interval $\mathcal{I} = [1.5, 5]$. Numerical tests show that increasing the sample size above the reference sampling with $N_c = 10$ points used in Section 3.4.1 (see Eq. (3.4.1)) brings no significant accuracy improvement. Therefore we chose to investigate in the following the performance of the optimal AO basis sets obtained with very sparse sampling. Figure 3.7 pictures the error of approximation of the dissociation curve and densities for three samplings: first, the two extreme

points of the interval $\mathcal{I} = [1.5, 5]$; second, two points around the equilibrium distance $a_0 \simeq 1.925$; third, a single point near the equilibrium distance. All curves are plotted for a fixed number of basis functions $N_b = 3$.

It appears that the latter sampling already provides satisfactory accuracy. The criteria J_{L^2} and J_{H^1} are equal to -5×10^{-6} for optimized basis to be compared with -1.8×10^{-3} for standard HBS. Hence they provide a gain of accuracy in energy of three orders of magnitude over the whole dissociation curve.

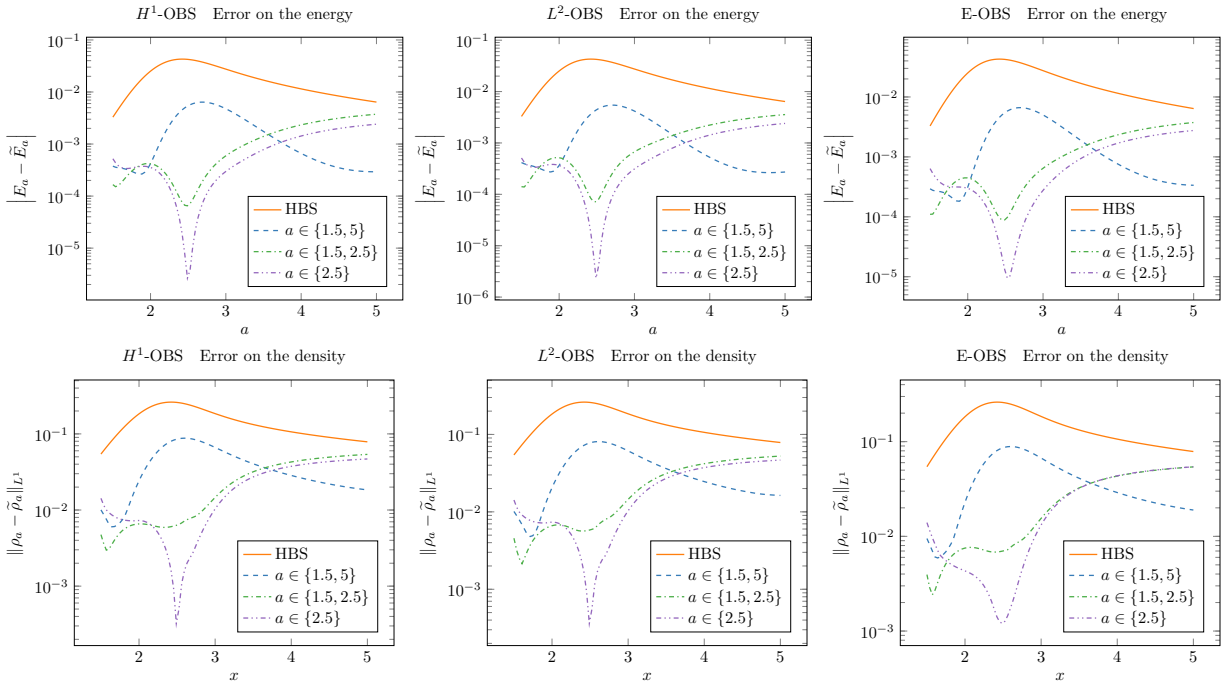


Figure 3.7 – Error plots for probability measures $\mathbb{P}$ corresponding to very sparse samplings of the interval $\mathcal{I} = [1.5, 5]$: i) the two endpoints of $\mathcal{I}$ ii) two points near the equilibrium distance and iii) one point near the equilibrium distance. (Top line) Error on energy. (Bottom line) Error on density in L^1 norm. (Left) OBS for J_{H^1}. (Middle) OBS for J_{L^2}. (Right) OBS for J_E. The "a" in legends are the sampled configurations a.

3.4.2.4 Number of Hilbert basis functions

We now take the same setting as in Section 3.4.1, except that we set $\mathcal{N} = 5$ instead of $\mathcal{N} = 10$. This provides similar results as those collected in Table 3.1, see Table 3.2. However, the values of the criteria J_A and J_E are higher than for $\mathcal{N} = 10$, in particular for $N_b = 4$, where criterion J_A cannot be optimized further than -10^{-5}, which makes sense as the space over which the optimization algorithms are performed is smaller. Calculations with $\mathcal{N} = 15$ were also performed: for $N_b = 1, 2, 3$, the criteria are slightly improved but for $N_b = 4$, convergence issues were noticed, due to ill conditioning of the overlap matrices for $a = 1.5$ as the number $\mathcal{N}$ of functions used to describe the optimal bases is larger.

Acknowledgements

The authors thank Susi Lehtola and Etienne Polack for fruitful discussions. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement EMC2 No 810367). This work was supported by the French Investissements dAvenir program, project Agence Nationale de la Recherche (ISITE-BFC) (contract ANR-15-IDEX-0003). The work of the second author was also supported by the Ecole des Ponts-ParisTech.

Value of J_{L^2} for the different basis sets

Basis	$N_b = 1$	$N_b = 2$	$N_b = 3$	$N_b = 4$
HBS	-7.40829	-7.70051	-7.74312	-7.77138
L^2-OBS	-7.43933	-7.76304	-7.77554	-7.77618
H^1-OBS	-7.43923	-7.76258	-7.77525	-7.77612
E-OBS	-7.39401	-7.76259	-7.77545	-7.77615

Value of J_{H^1} for the different basis sets

Basis	$N_b = 1$	$N_b = 2$	$N_b = 3$	$N_b = 4$
HBS	-10.5613	-11.0566	-11.1451	-11.2402
L^2-OBS	-10.6237	-11.2225	-11.2541	-11.2577
H^1-OBS	-10.6240	-11.2244	-11.2555	-11.2581
E-OBS	-10.5328	-11.2234	-11.2547	-11.2580

Value of J_E for the different basis sets

Basis	$N_b = 1$	$N_b = 2$	$N_b = 3$	$N_b = 4$
HBS	3.77956×10^{-2}	3.98301×10^{-3}	1.86537×10^{-3}	1.35309×10^{-4}
L^2-OBS	6.43832×10^{-2}	2.46466×10^{-4}	1.58667×10^{-5}	1.01128×10^{-5}
H^1-OBS	6.13025×10^{-2}	2.45930×10^{-4}	1.62235×10^{-5}	1.00611×10^{-5}
E-OBS	3.69681×10^{-2}	1.30365×10^{-4}	1.41935×10^{-5}	9.74560×10^{-6}

Table 3.2 – Value of the different criteria for the different local (optimized and Hermite) bases, with $\mathcal{N} = 5$ and increasing values of N_b.

Appendix

In this appendix, we will use extensively the two symmetries of the trace: for any matrices M and N such that MN and NM are defined,

$$\mathrm{Tr}(MN) = \mathrm{Tr}(NM) \quad \text{and} \quad \mathrm{Tr}(M^T) = \mathrm{Tr}(M).$$

Computation of the gradient of J_A

Let $R, H \in \mathbb{R}^{\mathcal{N}\times N_b}$ and define $I_H = \begin{pmatrix} H & 0 \\ 0 & H \end{pmatrix}$. One has

$$J_A(R+H) - J_A(R) = -\int_\Omega \mathrm{Tr}\left(M_A^{\text{offline}}(a)\left(2I_R[S^A(B_aI_R)]^{-1}I_H^T + I_R\left[\mathrm{d}[S^A]^{-1}(B_aI_R)\cdot(B_aI_H)\right]I_R^T\right)\right)\mathrm{d}\mathbb{P}(a) + O(\|H\|^2) \tag{3.4.2}$$

Considering that

$$(M+H)^{-1} - M^{-1} = -M^{-1}HM^{-1} + O(\|H\|^2) \text{ and } S^A(BI_{R+H}) - S^A(B_aI_R) = I_H^T S^A(B)I_R + I_R^T S^A(B)I_H + O(\|H\|^2),$$

it follows from the chain rule that

$$\mathrm{d}[S^A]^{-1}(B_aI_R)\cdot(B_aI_H) = -[S^A(B_aI_R)]^{-1}\left(I_H^T S^A(B_a)I_R + I_R^T S^A(B_a)I_H\right)[S^A(B_aI_R)]^{-1}.$$

From this computation, we obtain that the integrand in expression (3.4.2) writes for all a

$$\begin{aligned} &2\mathrm{Tr}\left(M_A^{\text{offline}}(a)\left[I_R[S^A(B_aI_R)]^{-1}I_H^T - I_R[S^A(B_aI_R)]^{-1}I_H^T S^A(B_a)I_R[S^A(B_aI_R)]^{-1}I_R^T\right]\right) \\ &= 2\mathrm{Tr}\left(M_A^{\text{offline}}(a)I_R[S^A(B_aI_R)]^{-1}I_H^T - I_H^T S^A(B_a)I_R[S^A(B_aI_R)]^{-1}I_R^T M_A^{\text{offline}}(a)I_R[S^A(B_aI_R)]^{-1}\right). \end{aligned} \tag{3.4.3}$$

The idea is now to write the expression (3.4.3) as the inner product of H with a given matrix of $\mathbb{R}^{\mathcal{N}\times N_b}$, which we will identify as the integrand of the gradient of J_A. Changing from I_H to H imposes to decompose each matrix by block and to write the trace in (3.4.3) as the sum of traces over the diagonal blocks. To this end we introduce the superscripts "++", "+−", "−+" and "−−" associated with one of the four identically shaped blocks of a generic matrix

$$M = \begin{pmatrix} M^{++} & M^{+-} \\ M^{-+} & M^{--} \end{pmatrix}. \tag{3.4.4}$$

Expression (3.4.3) therefore immediately reads

$$\begin{aligned} &2\mathrm{Tr}\Bigg(I_H^T \underbrace{\left[M_A^{\text{offline}}(a)I_R[S^A(B_aI_R)]^{-1} - S^A(B_a)I_R[S^A(B_aI_R)]^{-1}I_R^T M_A^{\text{offline}}(a)I_R[S^A(B_aI_R)]^{-1}\right]}_{M_A(a,R)}\Bigg) \\ &= 2\mathrm{Tr}\left(H^T\left(M_A(a,R)^{++} + M_A(a,R)^{--}\right)\right). \end{aligned} \tag{3.4.5}$$

One can verify that $M_A(a,R)^{++} + M_A(a,R)^{--}$ is in $\mathbb{R}^{\mathcal{N}\times N_b}$ and we conclude by identification that

$$\boxed{\nabla J_A(R) = -2\int_\Omega \left(M_A(a,R)^{++} + M_A(a,R)^{--}\right)\mathrm{d}\mathbb{P}(a).} \tag{3.4.6}$$

Computation of the gradient of J_E

Let $R, H \in \mathbb{R}^{\mathcal{N}\times N_b}$ and define $I_H = \begin{pmatrix} H & 0 \\ 0 & H \end{pmatrix}$. We immediately have that

$$\nabla J_E(R) = -2\int_\Omega \nabla E_a(R)\left(E_a - E_a(R)\right)\mathrm{d}\mathbb{P}(a), \tag{3.4.7}$$

where

$$E_a(R) = \mathrm{Tr}\left(C_a(R)(C_a(R))^T \times \mathcal{H}_a(R)\right), \tag{3.4.8}$$

with $C_a(R)$ defined in Section 3.3.2 and $\mathcal{H}_a(R) \coloneqq I_R^T M_E^{\text{offline}}(a) I_R$. Therefore, if we define $\mathcal{E}_a(R, C) = \mathrm{Tr}(CC^T \mathcal{H}_a(R))$, then $E_a(R) = \mathcal{E}_a(R, C_a(R))$ and we have, by the chain rule,

$$\nabla E_a(R) \cdot H = \nabla_R \mathcal{E}_a(R, C_a(R)) \cdot H + \nabla_C \mathcal{E}_a(R, C_a(R)) \cdot (\mathrm{d}C_a(R) \cdot H).$$

We now detail the computations of the two gradients of $\mathcal{E}_a$, namely $\nabla_R \mathcal{E}_a$ and $\nabla_C \mathcal{E}_a$.

Computation of the first gradient $\nabla_R \mathcal{E}_a$ Using notation (3.4.4), we introduce

$$M_a \coloneqq M_E^{\text{offline}}(a) \text{ and } \Sigma(H) \coloneqq I_H^T M_a I_R = \begin{pmatrix} H^T M_a^{++} R & H^T M_a^{+-} R \\ H^T M_a^{-+} R & H^T M_a^{--} R \end{pmatrix} \in \mathbb{R}^{(2N_b)\times(2N_b)},$$

so that, with $P = CC^T$,

$$\begin{aligned} \mathrm{Tr}\left(P[\mathrm{d}\mathcal{H}_a(R) \cdot H]\right) &= \mathrm{Tr}\left(P[\Sigma(H) + \Sigma(H)^T]\right) = 2\mathrm{Tr}(P\Sigma(H)) \\ &= 2\mathrm{Tr}\left(H^T \left(M_a^{++} R P^{++} + M_a^{-+} R P^{+-} + M_a^{+-} R P^{-+} + M_a^{--} R P^{--}\right)\right). \end{aligned}$$

In the end,

$$\nabla_R \mathcal{E}_a(R, C) = 2\left(M_a^{++} R (CC^T)^{++} + M_a^{+-} R (CC^T)^{-+} + M_a^{-+} R (CC^T)^{+-} + M_a^{--} R (CC^T)^{--}\right) \in \mathbb{R}^{\mathcal{N}\times N_b}.$$

Computation of the second gradient $\nabla_C \mathcal{E}_a$ The Euler–Lagrange equation of the minimization problem (3.3.21) yields that there exist a symmetric matrix $\Lambda_a(R) \in \mathbb{R}^{2\times 2}$ such that

$$\nabla_C \mathcal{E}_a(R, C_a(R)) = 2\mathcal{H}_a(R) = 2S(B_a I_R) C_a(R) \Lambda_a(R),$$

where $\Lambda_a(R)$ is actually a diagonal matrix whose diagonal is composed of the two lowest eigenvalues of $\mathcal{H}_a(R)$. Moreover, if we differentiate the constraint $C_a(R)^T S(B_a I_R) C_a(R) = \mathrm{Id}_2$, we get

$$C_a(R)^T S(B_a I_R)(\mathrm{d}C_a(R) \cdot H) + (\mathrm{d}C_a(R) \cdot H)^T S(B_a I_R) C_a(R) = -C_a(R)^T (\mathrm{d}S(B_a I_R) \cdot H) C_a(R),$$

so that

$$\begin{aligned} \nabla_C \mathcal{E}_a(R, C_a(R)) \cdot (\mathrm{d}C_a(R) \cdot H) &= 2\mathrm{Tr}\left((S(B_a I_R) C_a(R) \Lambda_a(R))^T (\mathrm{d}C_a(R) \cdot H)\right) \\ &= -\mathrm{Tr}\left((\mathrm{d}S(B_a I_R) \cdot H)\, C_a(R) \Lambda_a(R) C_a(R)^T\right). \end{aligned}$$

Now, let us recall that

$$\mathrm{d}S(B_a I_R) \cdot H = I_H^T S(B_a) I_R + I_R^T S(B_a) I_H.$$

Thus, by denoting $Q_a(R) = C_a(R) \Lambda_a(R) C_a(R)^T$, we get that

$$\begin{aligned} \nabla_C \mathcal{E}_a(R, C_a(R)) \cdot (\mathrm{d}C_a(R) \cdot H) = -2\mathrm{Tr}\big(H^T \big(&S(B_a)^{++} R Q_a(R)^{++} + S(B_a)^{+-} R Q_a(R)^{-+} \\ &+ S(B_a)^{-+} R Q_a(R)^{+-} + S(B_a)^{--} R Q_a(R)^{--}\big)\big) \end{aligned}$$

which ends the computations of the second gradient.

Final gradient Compiling the computations of the two previous paragraphs, we obtain

$$\boxed{\begin{aligned} \nabla_R E_a(R) = &\, 2\left(M_a^{++} R P_a(R)^{++} + M_a^{+-} R P_a(R)^{-+} + M_a^{-+} R P_a(R)^{+-} + M_a^{--} R P_a(R)^{--}\right) \\ &- 2\left(S_a^{++} R Q_a(R)^{++} + S_a^{+-} R Q_a(R)^{-+} + S_a^{-+} R Q_a(R)^{+-} + S_a^{--} R Q_a(R)^{--}\right) \end{aligned}} \tag{3.4.9}$$

where $P_a(R) = C_a(R) C_a(R)^T$, $M_a = M_A^{\text{offline}}(a)$, $S_a = S(B_a)$ and $Q_a(R) = C_a(R) \Lambda_a(R) C_a(R)^T$, and the gradient of J_E is computed with (3.4.7).

Part II

Solid State Physics

Chapter 4

Modified-operator method for the calculation of band diagrams of crystalline materials

This chapter has been published in the article [LV2]:

Eric Cancès, Muhammad Hassan, and Laurent Vidal. "Modified-operator method for the calculation of band diagrams of crystalline materials". In: Mathematics of Computation *(2023)*

The preliminary results of the last section are not from [LV2].

Abstract In solid state physics, electronic properties of crystalline materials are often inferred from the spectrum of periodic Schrödinger operators. As a consequence of Bloch's theorem, the numerical computation of electronic quantities of interest involves computing derivatives or integrals over the Brillouin zone of so-called energy bands, which are piecewise smooth, Lipschitz continuous periodic functions obtained by solving a parametrized elliptic eigenvalue problem on a Hilbert space of periodic functions. Classical discretization strategies for resolving these eigenvalue problems produce approximate energy bands that are either non-periodic or discontinuous, both of which cause difficulty when computing numerical derivatives or employing numerical quadrature. In this article, we study an alternative discretization strategy based on an ad hoc operator modification approach. While specific instances of this approach have been proposed in the physics literature, we introduce here a systematic formulation of this operator modification approach. We derive a priori error estimates for the resulting energy bands and we show that these bands are periodic and can be made arbitrarily smooth (away from band crossings) by adjusting suitable parameters in the operator modification approach. Numerical experiments involving a toy model in 1D, graphene in 2D, and silicon in 3D validate our theoretical results and showcase the efficiency of the operator modification approach.

Contents

4.1 Introduction

In solid state physics, macroscopic properties such as the electrical and thermal conductivities, heat capacity, magnetic susceptibility, and optical absorption of crystalline materials are often explained through the use of an independent electron model (see, e.g., [Har80, Part II, Chapter5], [Mar20, Chapter 12], [KMM96, Chapter7], and [GP13, Chapter 5]). This model consists of treating the crystalline material as an infinite, perfect crystal and modeling the electrons as independent of each other (quasiparticle approach) and evolving under the influence of an *effective* periodic potential. The behavior of each electron is thus determined by the spectrum of an unbounded, self-adjoint, periodic Schrödinger operator acting on $L^2(\mathbb{R}^3)$ (see, e.g., [RS78, ChapterXIII]). Although the independent electron assumption might seem naive, this model has achieved great success in explaining basic phenomena such as the difference between conductors, semi-conductors and insulators, as well as describing the electronic properties of many ubiquitous non-strongly correlated materials (see, e.g., [Har80, PartIII], [Mar20, Part V], [GP13, Chapters 10-12], and [KMM96, Chapters 6]). In addition, Kohn-Sham Density Functional Theory (DFT) provides a method to parameterize this independent-electron model and obtain *quantitatively* accurate results for a very large class of materials of practical interest (see, for instance, [Kax03; DG12]).

In the independent electron model, the practical computation of electronic quantities of interest is based on the use of the Bloch-Floquet transform (see, e.g., [RS78, Chapter XIII]). The Bloch-Floquet transform essentially yields an explicit block-diagonal decomposition of the underlying Schrödinger operator into so-called Bloch *fibers*, which are self-adjoint operators, bounded from below, acting on a space of periodic square-integrable functions. Thus, the problem of computing the spectrum of the periodic Schrödinger operator is reduced to one of calculating the low-lying eigenvalues of the Bloch fibers. These Bloch fibers are typically indexed by a parameter $\mathbf{k}$ that belongs to a d-dimensional torus (the Brillouin zone), and therefore each resulting eigenvalue (often referred to as an energy) can be viewed as a periodic function on the d-dimensional Brillouin Zone. It is thus common in the solid-state physics literature to speak of *energy bands*.

Energy bands provide both qualitative and quantitative information about the electronic properties of the crystalline material being studied (see, e.g., the references quoted above). Insulators and conductors for instance, are characterized by the presence or absence, respectively, of an energy *band gap*. Other electronic quantities of interest can be expressed in terms of integrals (over the Brillouin zone) or derivatives involving the energy bands (see, e.g., [Can+20]). In order to estimate important quantities such as the *integrated density of states* or the *integrated density of energy* (see Section 4.2 for precise definitions of these quantities), it is therefore necessary to

- sample the energy bands at different $\mathbf{k}$-points which corresponds to solving approximately the $\mathbf{k}$-fiber eigenvalue problems posed on a periodic domain;
- use suitable numerical quadrature to approximate integrals involving these energy bands.

Concerning the first step, the famous Monkhorst-Pack numerical scheme [MP76] is widely used to select the specific $\mathbf{k}$-points at which the eigenvalue problem is to be solved. For the second step, a number of numerical quadrature methods for integration in the Brillouin zone have been proposed including the well-known linear tetrahedron method (see, e.g., [LT72]) and the improvement due to Blöchl et al. [BJA94], and smearing methods (see, e.g., [Mor+18; PP99; Hen01; MP89]).

From a mathematical and computational point of view, two natural questions now arise. First, which discretization method should be employed in the actual numerical resolution of the $\mathbf{k}$-fiber eigenvalue problems, and second, what can be said about the convergence rate of the various numerical quadrature methods that are in use? For technical reasons, these questions become particularly relevant for *metallic* systems (see, e.g., [GL16] for an analysis involving insulators and semi-conductors), and in this case, the latter question has recently been addressed by the first author and coworkers in [Can+20]. The analysis carried out in [Can+20] revealed that the *periodicity* (with respect to the Brillouin zone) and *regularity* properties of the energy bands play a crucial role in the quadrature convergence rates, which of course is consistent with the experience of classical integration schemes in numerical analysis. Given that different eigenvalue discretization methods can conceivably produce (and in fact *do* produce, as we show in Section 4.3) energy bands that possess different regularity properties or may be altogether aperiodic, the choice of discretization scheme becomes vitally important. This article is concerned precisely with the study of approximation strategies for energy bands in the Brillouin zone.

The remainder of this article is organized as follows: We begin in Section 4.2 by introducing our notation and stating precisely the problem setting and governing equations. We then present in Section 4.3 two classical Galerkin discretization strategies for approximating the $\mathbf{k}$-fiber eigenvalue problems, and we show the problems associated with the energy bands produced by these classical approaches. Next, in Section 4.4, we present an alternative discretization scheme, systematizing ideas first introduced in the physics literature (see Remark 4.4.1 below), which is based on modifying in a controlled manner the underlying $\mathbf{k}$-fiber operator. In Section 4.5, we present our two main results on the analysis of this alternative approach: we derive a priori error estimates with respect to a discretization cutoff for the modified energy bands, and we show that these bands are periodic and can be made arbitrarily smooth (away from band crossings) by adjusting suitable parameters in the operator modification approach. Numerical experiments in Section 4.6 involving a 1D toy model, and two real materials (graphene and face-centered cubic silicon), validate our theoretical results and showcase the efficiency of the operator modification approach. Finally, in Section 4.7, we present the proofs of our main results.

4.2 Problem Formulation and Setting

Perfect crystals are structures composed of a periodic arrangement of atoms. Such structures can therefore be described very conveniently through the use of a suitable lattice: assuming a d-dimensional lattice with $d \in \mathbb{N}^* = \{1, 2, 3, \dots\}$, we denote by $\{\mathbf{a}_i\}_{i=1}^d$ a collection of d linearly independent primitive vectors in $\mathbb{R}^d$, and we denote by $\{\mathbf{b}_i\}_{i=1}^d \subset \mathbb{R}^d$ the corresponding reciprocal vectors, i.e., vectors in $\mathbb{R}^d$ that satisfy $\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\delta_{ij}$ $\forall i, j \in \{1, \dots, d\}$. The primitive lattice $\mathbb{L} \subset \mathbb{R}^d$ and reciprocal lattice $\mathbb{L}^* \subset \mathbb{R}^d$ are then defined as

$$\mathbb{L} := \{\mathbb{Z}\mathbf{a}_1 + \ldots + \mathbb{Z}\mathbf{a}_d\} \qquad \text{and} \qquad \mathbb{L}^* := \{\mathbb{Z}\mathbf{b}_1 + \ldots + \mathbb{Z}\mathbf{b}_d\}.$$

We denote by $\Omega \subset \mathbb{R}^d$ and $\Omega^* \subset \mathbb{R}^d$ the first Wigner-Seitz unit cell of the primitive and reciprocal lattice respectively. Recall that the first Wigner-Seitz unit cell of a lattice in $\mathbb{R}^d$ is the locus of points in $\mathbb{R}^d$ that are closer to the origin of the lattice than to any other lattice point. The first Wigner-Seitz cell Ω^* of the reciprocal lattice is called the (first) *Brillouin zone*.

Finally for clarity of the subsequent exposition, let us introduce the so-called translation operator and the related notion of lattice periodicity: Given any $\mathbf{y} \in \mathbb{R}^d$ and denoting $\mathcal{D}(\mathbb{R}^d) := \mathscr{C}_c^\infty(\mathbb{R}^d)$ the space of complex-valued smooth compactly-supported functions on $\mathbb{R}^d$, we define the translation operator $\tau_{\mathbf{y}} \colon \mathcal{D}(\mathbb{R}^d) \to \mathcal{D}(\mathbb{R}^d)$ as the mapping with the property that

$$\forall \Phi \in \mathcal{D}(\mathbb{R}^d) \colon \qquad \big(\tau_{\mathbf{y}}\Phi\big)(\mathbf{x}) := \Phi(\mathbf{x} - \mathbf{y}) \qquad \text{for a.e. } \mathbf{x} \in \mathbb{R}^d.$$

It follows that for any $\mathbf{y} \in \mathbb{R}^d$, the translation operator extends by duality as a mapping $\tau_{\mathbf{y}} \colon \mathcal{D}'(\mathbb{R}^d) \to \mathcal{D}'(\mathbb{R}^d)$.

Given now some $\Phi \in \mathcal{D}'(\mathbb{R}^d)$, we will say that Φ is $\mathbb{L}^*$-periodic (resp. $\mathbb{L}$-periodic) if $\tau_{\mathbf{G}}\Phi = \Phi$ for all $\mathbf{G} \in \mathbb{L}^*$ (resp. $\tau_{\mathbf{R}}\Phi = \Phi$ for all $\mathbf{R} \in \mathbb{L}$).

4.2.1 Function spaces and norms

We define the function space $L^2_{\mathrm{per}}(\Omega)$ as the set of (equivalence classes of) functions given by

$$L^2_{\mathrm{per}}(\Omega) := \big\{ f \in L^2_{\mathrm{loc}}(\mathbb{R}^d) \text{ such that } f \text{ is } \mathbb{L}\text{-periodic} \big\},$$

equipped with the inner-product

$$\forall f, g \in L^2_{\mathrm{per}}(\Omega) \colon \quad (f, g)_{L^2_{\mathrm{per}}(\Omega)} := \int_\Omega \overline{f(\mathbf{x})} g(\mathbf{x})\, d\mathbf{x},$$

where $L^2_{\mathrm{loc}}(\mathbb{R}^d)$ denotes the space of complex-valued, locally square-integrable functions on $\mathbb{R}^d$, and $\overline{f(\cdot)}$ indicates the complex conjugate of $f(\cdot)$. The spaces $L^p_{\mathrm{per}}(\Omega)$, $p \in [1, 2) \cup (2, \infty]$ are defined analogously.

We denote by $\mathcal{B}$, the orthonormal Fourier basis of $L^2_{\mathrm{per}}(\Omega)$, i.e.,

$$\mathcal{B} := \left\{ e_{\mathbf{G}}(\mathbf{x}) := \frac{1}{|\Omega|^{\frac{1}{2}}} e^{\imath \mathbf{G} \cdot \mathbf{x}} \colon \quad \mathbf{G} \in \mathbb{L}^* \right\}.$$

It follows from the definition of the reciprocal lattice $\mathbb{L}^*$ that the basis set $\mathcal{B}$ consists precisely of $\mathbb{L}$-periodic plane-waves. Thus, given any $f \in L^2_{\mathrm{per}}(\Omega)$, we will frequently express f in the form

$$f = \sum_{\mathbf{G}\in\mathbb{L}^*} \widehat{f}_{\mathbf{G}} e_{\mathbf{G}}, \qquad \text{where } \widehat{f}_{\mathbf{G}} := \int_\Omega f(\mathbf{x})\overline{e_{\mathbf{G}}(\mathbf{x})}\,d\mathbf{x}, \qquad \text{and}$$
$$\sum_{\mathbf{G}\in\mathbb{L}^*} \left|\widehat{f}_{\mathbf{G}}\right|^2 = \int_\Omega |f(\mathbf{x})|^2\,d\mathbf{x} < \infty.$$

Periodic Sobolev spaces of positive orders are constructed analogously. Indeed, we define for each $s > 0$ the set

$$H^s_{\mathrm{per}}(\Omega) := \left\{ f \in L^2_{\mathrm{per}}(\Omega) \colon \sum_{\mathbf{G}\in\mathbb{L}^*} \left(1 + |\mathbf{G}|^2\right)^s \left|\widehat{f}_{\mathbf{G}}\right|^2 < \infty \right\},$$

equipped with the inner-product

$$\forall f, g \in H^s_{\mathrm{per}}(\Omega) \colon \quad (f,g)_{H^s_{\mathrm{per}}(\Omega)} := \sum_{\mathbf{G}\in\mathbb{L}^*} \left(1 + |\mathbf{G}|^2\right)^s \overline{\widehat{f}_{\mathbf{G}}}\widehat{g}_{\mathbf{G}}.$$

Naturally, we have $H^0_{\mathrm{per}}(\Omega) := L^2_{\mathrm{per}}(\Omega)$, and we define periodic Sobolev spaces of negative orders through duality, i.e., for each $s > 0$ we define $H^{-s}_{\mathrm{per}}(\Omega) := \left(H^s_{\mathrm{per}}(\Omega)\right)'$, and we equip $H^{-s}_{\mathrm{per}}(\Omega)$ with the canonical dual norm.

Finally, given a Banach space X, we will write $\mathcal{L}(X)$ to denote the Banach space of bounded linear operators from X to X, equipped with the usual operator norm.

4.2.2 Governing operators and quantities of interest

In this section, we assume that the electronic properties of the crystal that we study are encoded in an effective one-body Schrödinger operator

$$\mathrm{H} := -\frac{1}{2}\Delta + V \qquad \text{acting on } L^2(\mathbb{R}^d) \quad \text{with domain } H^2(\mathbb{R}^d), \tag{4.2.1}$$

where $V \in L^\infty_{\mathrm{per}}(\Omega)$ is an $\mathbb{L}$-periodic *effective potential*. Many electronic properties of the crystal we study can be computed from the spectral decomposition of this one-body Hamiltonian operator H, and we are therefore interested in its analysis and computation. The classical approach to this problem relies on the use of the Bloch-Floquet transform (see, e.g., [RS78, Chapter XIII]), which we will now briefly present. The following exposition is based on the article [Can+21a].

We begin by introducing for each $\mathbf{G} \in \mathbb{L}^*$, the unitary multiplication operator $T_{\mathbf{G}} \colon L^2_{\mathrm{per}}(\Omega) \to L^2_{\mathrm{per}}(\Omega)$ defined as

$$\forall v \in L^2_{\mathrm{per}}(\Omega) \colon \qquad \left(T_{\mathbf{G}} v\right)(\mathbf{x}) = e^{-\imath \mathbf{G}\cdot\mathbf{x}} v(\mathbf{x}) \qquad \text{for a.e. } \mathbf{x} \in \mathbb{R}^d.$$

Next, we introduce the Hilbert space of $\mathbb{L}^*$-quasi-periodic, $L^2_{\mathrm{per}}(\Omega)$-valued functions on $\mathbb{R}^d$ as the vector space

$$L^2_{\mathrm{qp}}(\mathbb{R}^d; L^2_{\mathrm{per}}(\Omega)) := \Big\{ \mathbb{R}^d \ni \mathbf{k} \mapsto u_{\mathbf{k}} \in L^2_{\mathrm{per}}(\Omega) \colon \int_{\Omega^*} \|u_{\mathbf{k}}\|^2_{L^2_{\mathrm{per}}(\Omega)} \mathrm{d}\mathbf{k} < \infty \quad \text{and}$$
$$u_{\mathbf{k}+\mathbf{G}} = T_{\mathbf{G}} u_{\mathbf{k}} \ \forall\ \mathbf{G} \in \mathbb{L}^* \ \text{ and a.e. } \mathbf{k} \in \mathbb{R}^d \Big\},$$

equipped with the inner product

$$\forall u, v \in L^2_{\mathrm{qp}}(\mathbb{R}^d; L^2_{\mathrm{per}}(\Omega)) \colon \qquad (u,v)_{L^2_{\mathrm{qp}}(\mathbb{R}^d; L^2_{\mathrm{per}}(\Omega))} = \fint_{\Omega^*} (u_{\mathbf{k}}, v_{\mathbf{k}})_{L^2_{\mathrm{per}}(\Omega)}\,d\mathbf{k},$$

where we have denoted $\fint_{\Omega^*} := \frac{1}{|\Omega^*|}\int_{\Omega^*}$ and we have used the subscript 'qp' to highlight *quasi-periodicity*.

The Bloch-Floquet transform is now the unitary mapping from $L^2(\mathbb{R}^d)$ to $L^2_{\mathrm{qp}}(\mathbb{R}^d; L^2_{\mathrm{per}}(\Omega))$ with the property that any $u \in \mathscr{D}(\mathbb{R}^d)$ is mapped to the element of $L^2_{\mathrm{qp}}(\mathbb{R}^d; L^2_{\mathrm{per}}(\Omega))$ defined as

$$\mathbb{R}^d \ni \mathbf{k} \mapsto u_{\mathbf{k}} := \sum_{\mathbf{R}\in\mathbb{L}} u\big(\,\bullet + \mathbf{R}\big) e^{-\imath \mathbf{k}\cdot(\bullet+\mathbf{R})} \in L^2_{\mathrm{per}}(\Omega).$$

Any bounded linear operator $\mathrm{A}\colon L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ that is $\mathbb{L}$-periodic, i.e., one that commutes with the translation operator $\tau_{\mathbf{R}}$ for all $\mathbf{R} \in \mathbb{L}$, is decomposed by the Bloch-Floquet transform in the following sense: denoting by $L^\infty_{\mathrm{qp}}\left(\mathbb{R}^d; \mathscr{L}\big(L^2_{\mathrm{per}}(\Omega)\big)\right)$, the vector space defined as

$$L^\infty_{\mathrm{qp}}\left(\mathbb{R}^d; \mathscr{L}\big(L^2_{\mathrm{per}}(\Omega)\big)\right) := \Big\{ \mathbb{R}^d \ni \mathbf{k} \mapsto \mathrm{A}_{\mathbf{k}} \in \mathscr{L}\left(L^2_{\mathrm{per}}(\Omega)\right) : \sup_{\mathbf{k}\in\Omega^*} \|\mathrm{A}_{\mathbf{k}}\|_{\mathscr{L}\left(L^2_{\mathrm{per}}(\Omega)\right)} < \infty \\ \text{and} \quad \mathrm{A}_{\mathbf{k}+\mathbf{G}} = T_{\mathbf{G}} \mathrm{A}_{\mathbf{k}} T^*_{\mathbf{G}} \ \forall\, \mathbf{G} \in \mathbb{L}^* \ \text{ and a.e. } \mathbf{k} \in \mathbb{R}^d \Big\},$$

there exists a function $\mathbf{k} \mapsto \mathrm{A}_{\mathbf{k}}$ in $L^\infty_{\mathrm{qp}}\left(\mathbb{R}^d; \mathscr{L}\big(L^2_{\mathrm{per}}(\Omega)\big)\right)$ such that for any $u \in L^2(\mathbb{R}^d)$, all $\mathbf{G} \in \mathbb{L}^*$ and a.e. $\mathbf{k} \in \mathbb{R}^d$ it holds that

$$(\mathrm{A}u)_{\mathbf{k}} = \mathrm{A}_{\mathbf{k}} u_{\mathbf{k}}. \tag{4.2.2}$$

Here, the operators $(\mathrm{A}_{\mathbf{k}})_{\mathbf{k}\in\mathbb{R}^d} \in \mathscr{L}\left(L^2_{\mathrm{per}}(\Omega)\right)$ are called the Bloch fibers of A.

The Bloch decomposition (4.2.2) can also be extended to *unbounded*, $\mathbb{L}$-periodic self-adjoint operators such as the one-body electronic Hamiltonian defined through Equation (4.2.1). In this case, the fibers $\mathrm{H}_{\mathbf{k}}$, $\mathbf{k} \in \mathbb{R}^d$ of the electronic Hamiltonian H are *unbounded* operators on $L^2_{\mathrm{per}}(\Omega)$ given by

$$\mathrm{H}_{\mathbf{k}} := \frac{1}{2}\left(-\imath\nabla + \mathbf{k}\right)^2 + V, \qquad \text{with domain } H^2_{\mathrm{per}}(\Omega). \tag{4.2.3}$$

A detailed proof of this technical result can be found in [RS78, Chapter XIII].

Thanks to the Bloch-Floquet decomposition (4.2.2), the spectral properties of the Hamiltonian H can be deduced using properties of the fibers $\mathrm{H}_{\mathbf{k}}$, $\mathbf{k} \in \mathbb{R}^d$. Indeed, it is a classical result (see, e.g., [RS78, Chapter XIII]) that

- each $\mathrm{H}_{\mathbf{k}}$ is a self-adjoint operator on $L^2_{\mathrm{per}}(\Omega)$ with domain $H^2_{\mathrm{per}}(\Omega)$ and form domain $H^1_{\mathrm{per}}(\Omega)$ (see, e.g., [RS72, Chapter VIII.6] for a definition of the form domain). Additionally, each $\mathrm{H}_{\mathbf{k}}$ is bounded below and has compact resolvent so that each $\mathrm{H}_{\mathbf{k}}$ has a purely discrete spectrum with eigenvalues accumulating at $+\infty$ and eigenfunctions that form an orthonormal basis for $L^2_{\mathrm{per}}(\Omega)$;
- H is a self-adjoint operator, bounded from below, on $L^2(\mathbb{R}^d)$ with domain $H^2(\mathbb{R}^d)$ and form domain $H^1(\mathbb{R}^d)$. Additionally, H has a purely absolutely continuous spectrum, and it holds that $\sigma(\mathrm{H}) = \sigma_{\mathrm{ac}}(\mathrm{H}) = \bigcup\limits_{\mathbf{k}\in\overline{\Omega^*}} \sigma(\mathrm{H}_{\mathbf{k}})$.

From the point of view of applications, the Bloch-Floquet decomposition (4.2.2) also allows for the calculation of electronic properties of interest of the underlying crystal using only spectral information from the fibers $\mathrm{H}_{\mathbf{k}}$, $\mathbf{k} \in \mathbb{R}^d$ (see below and also, e.g., [RS78; Can+20] for details). As a consequence, it suffices to focus our attention purely on the resolution of the eigenvalue problem for the operators $\mathrm{H}_{\mathbf{k}}$, $\mathbf{k} \in \mathbb{R}^d$ defined through Equation (4.2.3).

Given a fiber $\mathrm{H}_{\mathbf{k}}$, $\mathbf{k} \in \mathbb{R}^d$ defined through Equation (4.2.3), we seek an $L^2_{\mathrm{per}}(\Omega)$-orthonormal basis $\{(\varepsilon_{n,\mathbf{k}}, u_{n,\mathbf{k}})\}_{n\in\mathbb{N}^*} \subset \left(\mathbb{R} \times L^2_{\mathrm{per}}(\Omega)\right)^{\mathbb{N}^*}$ of eigenmodes of $\mathrm{H}_{\mathbf{k}}$:

$$\begin{aligned} \mathrm{H}_{\mathbf{k}} u_{n,\mathbf{k}} &= \varepsilon_{n,\mathbf{k}} u_{n,\mathbf{k}} \qquad && \text{and} \\ (u_{n,\mathbf{k}}, u_{m,\mathbf{k}})_{L^2_{\mathrm{per}}(\Omega)} &= \delta_{nm} && \forall n, m \in \mathbb{N}^*. \end{aligned} \tag{4.2.4}$$

Equipped with such a basis, we can introduce several important electronic properties of interest. To this end, we first require a convention and some notation.

Convention 1. Consider the setting of the eigenvalue problem (4.2.4). By convention, for every $\mathbf{k} \in \mathbb{R}^d$ we order the eigenvalues $\varepsilon_{n,\mathbf{k}},\ n \in \mathbb{N}^*$ (counting multiplicities) such that

$$\varepsilon_{1,\mathbf{k}} \leqslant \varepsilon_{2,\mathbf{k}} \leqslant \varepsilon_{3,\mathbf{k}} \leqslant \varepsilon_{4,\mathbf{k}} \cdots.$$

Moreover, for every $n \in \mathbb{N}^*$ we will write $\varepsilon_n \colon \mathbb{R}^d \to \mathbb{R}$ for the mapping $\mathbf{k} \mapsto \varepsilon_{n,\mathbf{k}}$, and we will call ε_n the n^{th} energy band. Since the functions $\varepsilon_n,\ n \in \mathbb{N}^*$ are continuous (as a straightforward consequence of the Courant-Fisher min-max theorem), we have

$$\sigma(\mathrm{H}) = \bigcup_{n \in \mathbb{N}^*} \mathrm{Ran}\,(\varepsilon_n),$$

and for each $n \in \mathbb{N}^*$ it holds that $\mathrm{Ran}\,(\varepsilon_n) = [\min \varepsilon_n, \max \varepsilon_n]$ is an interval.

We will use a similar convention for any subsequent eigenvalue problem that we introduce in the sequel.

The energy bands play a key role in the definition of various electronic properties of a perfect crystal. Indeed, given $\mathbf{k} \in \mathbb{R}^d$ and the Bloch fiber $\mathrm{H}_{\mathbf{k}}$ defined through Equation (4.2.3), the $\mathbf{k}^{\text{th}}$ fiber of the **one-body ground-state density matrix** at chemical potential $\mu \in \mathbb{R}$ is defined as the bounded self-adjoint operator $\gamma_{\mathbf{k}}(\mu) \colon L^2_{\text{per}}(\Omega) \to L^2_{\text{per}}(\Omega)$ given by

$$\gamma_{\mathbf{k}}(\mu) := \mathbb{1}(\mathrm{H}_{\mathbf{k}} \leqslant \mu) = \sum_{n \in \mathbb{N}^*} \mathbb{1}\big(\varepsilon_n(\mathbf{k}) \leqslant \mu\big)\, |u_{n,\mathbf{k}}\rangle \langle u_{n,\mathbf{k}}|\,.$$

The **integrated density of states** is defined as the function $\mathcal{n} \colon \mathbb{R} \to \mathbb{R}_+$ given by

$$\forall \mu \in \mathbb{R} \colon \qquad \mathcal{n}(\mu) := \sum_{n \in \mathbb{N}^*} \fint_{\Omega^*} \mathbb{1}\big(\varepsilon_n(\mathbf{k}) \leqslant \mu\big) d\mathbf{k}.$$

Lastly, the **integrated density of energy** is defined as the function $\mathcal{E} \colon \mathbb{R} \to \mathbb{R}$ given by

$$\forall \mu \in \mathbb{R} \colon \qquad \mathcal{E}(\mu) := \sum_{n \in \mathbb{N}^*} \fint_{\Omega^*} \varepsilon_n(\mathbf{k}) \mathbb{1}\big(\varepsilon_n(\mathbf{k}) \leqslant \mu\big) d\mathbf{k}.$$

Often the above quantities are computed for $\mu = \mu_{\mathrm{F}} \in \mathbb{R}$ where μ_{F} is known as the *Fermi level* and is defined through the relation $\mathcal{n}(\mu_{\mathrm{F}}) = N$ with N being the number of electrons (or electrons pairs if spin is taken into account) per unit cell. Naturally, computing any of these physical observables requires the approximation, through numerical quadrature, of integrals over the Brillouin zone Ω^* that involve the energy bands $\{\varepsilon_n\}_{n \in \mathbb{N}^*}$. This is a highly non-trivial problem in the case of *metallic* systems for which the Fermi level μ_{F} is an interior point of $\sigma(\mathrm{H})$, and several numerical methods have been proposed for Brillouin zone integration (see, for instance, the previously cited articles [LT72; BJA94; Mor+18; PP99; Hen01]). From the point of view of numerical analysis, it is natural to ask for error bounds for the various numerical methods in the literature, and such an error analysis has recently been carried out in [Can+20] under the assumption that the values of the functions $\varepsilon_n,\ n \in \mathbb{N}^*$ can be computed exactly at any $\mathbf{k} \in \Omega^*$.

As is typically the case in the analysis of quadrature methods, the error analysis in [Can+20] makes use of functional properties of the energy bands $\{\varepsilon_n\}_{n \in \mathbb{N}^*}$. This analysis shows that there are two properties of these energy bands in particular that are necessary in order to deduce higher order convergence rates for numerical quadrature in the Brillouin zone.

Property one (Periodicity of the eigenvalues)**.**

Consider the setting of the eigenvalue problem (4.2.4) and let Convention 1 hold. Then for each $n \in \mathbb{N}^*$, the function ε_n is $\mathbb{L}^*$-periodic.

The proof follows in view of Convention 1 by recognizing that for any $\mathbf{k} \in \mathbb{R}^d$ and any $\mathbf{G} \in \mathbb{L}^*$, the operators $\mathrm{H}_{\mathbf{k}}$ and $\mathrm{H}_{\mathbf{k}+\mathbf{G}}$ are unitarily equivalent through the unitary multiplication operator $T_{\mathbf{G}} \colon L^2_{\text{per}}(\Omega) \to L^2_{\text{per}}(\Omega)$ defined in Section 4.2.2.

Property two (Continuity of the eigenvalues)**.**

Consider the setting of the eigenvalue problem (4.2.4), and let the maps $\{\varepsilon_n\}_{n \in \mathbb{N}^*}$ be defined according to Convention 1. Then, each function ε_n is Lipschitz continuous on $\mathbb{R}^d$. Additionally, if $\mathbf{k}_n \in \mathbb{R}^d$ is such that

$$\varepsilon_{n,\mathbf{k}_n} \neq \varepsilon_{m,\mathbf{k}_n} \quad \forall\, \mathbb{N}^* \ni m \neq n, \qquad \big(\text{No energy band crossings at } (\mathbf{k}_n, \varepsilon_{n,\mathbf{k}_n})\big),$$

then ε_n is locally real-analytic at $\mathbf{k}_n$, i.e., $\exists \delta_n > 0$ such that ε_n is real-analytic on the open ball $\mathbb{B}_{\delta_n}(\mathbf{k}_n)$.

A proof of this statement can, for instance, be found in [Can+20, Lemma 3.2].

The $\mathbb{L}^*$-periodicity and real-analyticity away from crossings of the energy bands $\{\varepsilon_n\}_{n\in\mathbb{N}^*}$ has significant consequences for evaluating Brillouin zone integrals involving these functions. Indeed as can readily be deduced from [KR09], for d-dimensional periodic integrands of class $\mathscr{C}^r$, the uniform grid quadrature rule converges as $\mathcal{O}\Big((\Delta x)^{\frac{r}{d}}\Big)$ when integrating over an entire period. For real-analytic periodic integrands, a uniform grid quadrature rule even recovers exponential convergence (see, e.g., [TW14]). This fact is essential in understanding the approximability of the integrals appearing in the definitions of the various electronic properties of interest defined above, and is a key element of the higher order convergence rates for numerical quadrature obtained in [Can+20].

Of course in practice, we typically do not have access to the *exact* energy bands $\{\varepsilon_n\}_{n\in\mathbb{N}^*}$, these being solutions to infinite-dimensional eigenvalue problems. Instead, the eigenvalue problem (4.2.4) is typically discretized in some M-dimensional basis for specific values of $\mathbf{k} \in \Omega^* \subset \mathbb{R}^d$ corresponding to the grid points of our chosen quadrature method. This naturally raises the question of how the resulting *approximate* energy bands $\{\varepsilon_n^{\text{approx}}\}_{i=1}^M$ compare to the exact bands $\{\varepsilon_n\}_{n\in\mathbb{N}^*}$, and in particular whether **Properties one** and **two** also hold for the approximate bands $\{\varepsilon_n^{\text{approx}}\}_{i=1}^M$, these properties being essential to the quadrature error analysis. This is the topic of the next section.

4.3 Classical Discretization Strategies

We will now describe two well-known discretization strategies for resolving the eigenvalue problem (4.2.4). Throughout this section, we assume the setting of Section 4.2, and we recall in particular that we will use Convention 1 in ordering and labelling all eigenvalue problems that appear in this section.

Definition 4.3.1 (Uniform basis set)**.**
Let $\mathrm{E_c} > 0$ denote a scalar cutoff. We define the basis set $\mathcal{B}_0^{\mathrm{E_c}} \subset H^1_{\text{per}}(\Omega)$ as

$$\mathcal{B}_0^{\mathrm{E_c}} := \left\{ e_{\mathbf{G}} : \quad \mathbf{G} \in \mathbb{L}^* \quad \text{with} \quad \frac{1}{2}|\mathbf{G}|^2 < \mathrm{E_c} \right\},$$

and we define the subspace spanned by this basis set as $\mathrm{X}_0^{\mathrm{E_c}} := \operatorname{span} \mathcal{B}_0^{\mathrm{E_c}}$.

Notation 4.3.1 (Projections involving the uniform basis set).
Consider the setting of Definition 4.3.1. We denote by $\Pi_{\mathrm{E_c}} \colon L^2_{\text{per}}(\Omega) \to L^2_{\text{per}}(\Omega)$ the L^2_{per}-orthogonal projection operator onto $\mathrm{X}_0^{\mathrm{E_c}}$, and we denote by $\Pi_{\mathrm{E_c}}^{\perp}$ its complement, i.e., $\Pi_{\mathrm{E_c}}^{\perp} := \mathbb{I} - \Pi_{\mathrm{E_c}}$.

Additionally, for each $\mathbf{k} \in \mathbb{R}^d$ we denote by $\breve{\mathrm{H}}_{\mathbf{k}}^{\mathrm{E_c}}$ the two-sided projection of the Hamiltonian fiber $\mathrm{H}_{\mathbf{k}}$ in $\mathrm{X}_0^{\mathrm{E_c}}$, i.e., $\breve{\mathrm{H}}_{\mathbf{k}}^{\mathrm{E_c}} := \Pi_{\mathrm{E_c}} \mathrm{H}_{\mathbf{k}} \Pi_{\mathrm{E_c}}$.

Equipped with the uniform basis sets defined through Definition 4.3.1, we can propose the following elementary Galerkin discretization of the eigenvalue problem (4.2.4).

Uniform Galerkin discretization of the eigenvalue problem (4.2.4)**.**
Given a fiber $\mathrm{H}_{\mathbf{k}}$, $\mathbf{k} \in \mathbb{R}^d$ defined through Equation (4.2.3) and a scalar cutoff $\mathrm{E_c} > 0$, we seek an orthonormal basis $\left\{ \left(\breve{\varepsilon}_{n,\mathbf{k}}^{\mathrm{E_c}}, \breve{u}_{n,\mathbf{k}}^{\mathrm{E_c}} \right) \right\} \subset \mathbb{R} \times \mathrm{X}_0^{\mathrm{E_c}}$ of eigenvectors of $\breve{\mathrm{H}}_{\mathbf{k}}^{\mathrm{E_c}}$:

$$\begin{aligned} \breve{\mathrm{H}}_{\mathbf{k}}^{\mathrm{E_c}} \breve{u}_{n,\mathbf{k}}^{\mathrm{E_c}} &= \breve{\varepsilon}_{n,\mathbf{k}}^{\mathrm{E_c}} \breve{u}_{n,\mathbf{k}}^{\mathrm{E_c}} \qquad && \text{and} \\ \left(\breve{u}_{n,\mathbf{k}}^{\mathrm{E_c}}, \breve{u}_{m,\mathbf{k}}^{\mathrm{E_c}} \right)_{L^2_{\text{per}}(\Omega)} &= \delta_{nm} && \forall n, m \in \{1, \ldots, \dim \breve{\mathrm{H}}_{\mathbf{k}}^{\mathrm{E_c}}\}. \end{aligned} \tag{4.3.1}$$

An alternative to the uniform Galerkin discretization (4.3.1) is provided by the use of so-called **k**-dependent basis sets.

Definition 4.3.2 (**k**-dependent basis set)**.**

Let $\mathrm{E_c} > 0$ denote a scalar cutoff, and let $\mathbf{k} \in \mathbb{R}^d$. We define the basis set $\mathcal{B}_{\mathbf{k}}^{\mathrm{E_c}} \subset H^1_{\mathrm{per}}(\Omega)$ as

$$\mathcal{B}_{\mathbf{k}}^{\mathrm{E_c}} := \left\{ e_{\mathbf{G}} : \quad \mathbf{G} \in \mathbb{L}^* \quad \text{with} \quad \frac{1}{2}|\mathbf{k}+\mathbf{G}|^2 < \mathrm{E_c} \right\},$$

and we define the subspace spanned by this basis set as $\mathrm{X}_{\mathbf{k}}^{\mathrm{E_c}} := \operatorname{span} \mathcal{B}_{\mathbf{k}}^{\mathrm{E_c}}$. Additionally, we write $M_{\mathrm{E_c}}(\mathbf{k})$ to denote the cardinality of $\mathcal{B}_{\mathbf{k}}^{\mathrm{E_c}}$, and we refer to $\mathcal{B}_{\mathbf{k}}^{E_c}$ as a $\mathbf{k}$-dependent basis set.

Remark 4.3.1. Consider the setting of Definition 4.3.2. It can readily be seen that for a fixed $\mathrm{E_c}$, the cardinality $M_{\mathrm{E_c}}(\mathbf{k})$ of the basis $\mathcal{B}_{\mathbf{k}}^{\mathrm{E_c}}$ is not fixed and depends indeed on $\mathbf{k} \in \mathbb{R}^d$. In the sequel, we will therefore regard $M_{\mathrm{E_c}}(\cdot)$ as a piecewise constant mapping from $\mathbb{R}^d$ to $\mathbb{N}^*$, which is moreover uniformly bounded below and above by optimal constants $M^-_{\mathrm{E_c}}$ and $M^+_{\mathrm{E_c}}$ respectively that depend on $\mathrm{E_c}$. A visual example of the uniform and $\mathbf{k}$-dependent basis sets is given in Figure 4.1.

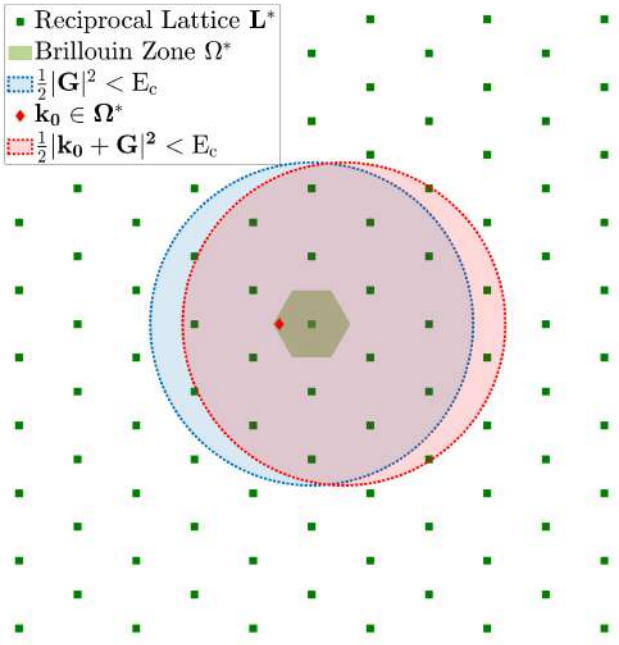


Figure 4.1 – An example of the uniform and k-dependent basis sets. The reciprocal lattice $\mathbb{L}^*$ is triangular and indicated in dark green with the corresponding Brillouin zone Ω^* shaded light green. The blue disk contains all $\mathbf{G} \in \mathbb{L}^*$ which belong to the basis set $\mathcal{B}_0^{\mathrm{E_c}}$ while the red disk contains all $\mathbf{G} \in \mathbb{L}^*$ which belong to the basis set $\mathcal{B}_{\mathbf{k}_0}^{\mathrm{E_c}}$ for a given $\mathbf{k}_0$ in the Brillouin zone Ω^*. Notice that the $\mathbf{k}$-dependent basis set $\mathcal{B}_{\mathbf{k}_0}^{\mathrm{E_c}}$ contains an additional four $\mathbf{G} \in \mathbb{L}^*$ that are missing from the uniform basis set $\mathcal{B}_0^{\mathrm{E_c}}$. Similarly, $\mathcal{B}_0^{\mathrm{E_c}}$ contains two $\mathbf{G} \in \mathbb{L}^*$ that are missing from the basis set $\mathcal{B}_{\mathbf{k}_0}^{\mathrm{E_c}}$.

Notation 4.3.2 (Projections involving the $\mathbf{k}$-dependent basis set).

Consider the setting of Definition 4.3.2. For each $\mathbf{k} \in \mathbb{R}^d$, we denote by $\Pi_{\mathbf{k},\mathrm{E_c}} \colon L^2_{\mathrm{per}}(\Omega) \to L^2_{\mathrm{per}}(\Omega)$ the L^2_{per}-orthogonal projection operator onto $\mathrm{X}_{\mathbf{k}}^{\mathrm{E_c}}$, and we denote by $\Pi^{\perp}_{\mathbf{k},\mathrm{E_c}}$ its complement, i.e., $\Pi^{\perp}_{\mathbf{k},\mathrm{E_c}} := \mathbb{I} - \Pi_{\mathbf{k},\mathrm{E_c}}$.

Additionally, for each $\mathbf{k} \in \mathbb{R}^d$ we denote by $\mathrm{H}_{\mathbf{k}}^{\mathrm{E_c}}$ the two-sided projection of the Hamiltonian fiber $\mathrm{H}_{\mathbf{k}}$ in $\mathrm{X}_{\mathbf{k}}^{\mathrm{E_c}}$, i.e., $\mathrm{H}_{\mathbf{k}}^{\mathrm{E_c}} := \Pi_{\mathbf{k},\mathrm{E_c}} \mathrm{H}_{\mathbf{k}} \Pi_{\mathbf{k},\mathrm{E_c}}$.

$\mathbf{k}$-dependent Galerkin discretization of the eigenvalue problem (4.2.4).

Given a fiber $\mathrm{H}_{\mathbf{k}}$, $\mathbf{k} \in \mathbb{R}^d$ defined through Equation (4.2.3) and a scalar cutoff $\mathrm{E_c} > 0$, we seek an orthonormal basis $\left\{\left(\varepsilon^{\mathrm{E_c}}_{n,\mathbf{k}}, u^{\mathrm{E_c}}_{n,\mathbf{k}}\right)\right\} \subset \mathbb{R} \times \mathrm{X}^{\mathrm{E_c}}_{\mathbf{k}}$ of eigenvectors of $\mathrm{H}^{\mathrm{E_c}}_{\mathbf{k}}$:

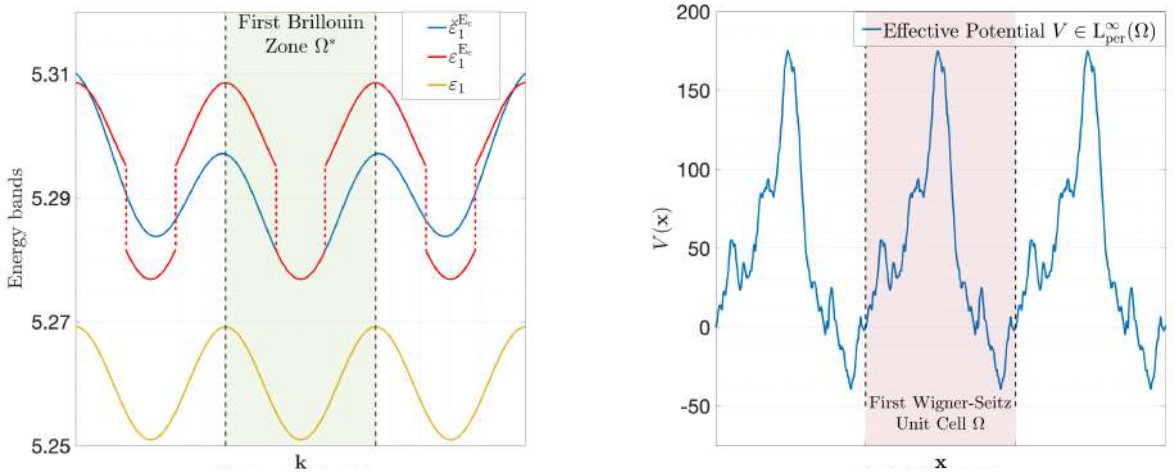
$$\begin{aligned} \mathrm{H}^{\mathrm{E_c}}_{\mathbf{k}} u^{\mathrm{E_c}}_{n,\mathbf{k}} &= \varepsilon^{\mathrm{E_c}}_{n,\mathbf{k}} u^{\mathrm{E_c}}_{n,\mathbf{k}} \qquad && \text{and} \\ \left(u^{\mathrm{E_c}}_{n,\mathbf{k}}, u^{\mathrm{E_c}}_{m,\mathbf{k}}\right)_{L^2_{\mathrm{per}}(\Omega)} &= \delta_{nm} \qquad && \forall n, m \in \{1, \ldots, M_{\mathrm{E_c}}(\mathbf{k})\}. \end{aligned} \tag{4.3.2}$$

The Galerkin discretizations (4.3.1) and (4.3.2) are both well-posed, and a straightforward analysis reveals the following error bound: for any fixed $\mathbf{k} \in \mathbb{R}^d$, any $n \in \mathbb{N}^*$, there exists $\mathrm{E}^* > 0$ such that for scalar cutoffs $\mathrm{E_c} \geqslant \mathrm{E}^*$, we have eigenvalue bounds of the form:

$$|\breve{\varepsilon}^{\mathrm{E_c}}_{n,\mathbf{k}} - \varepsilon_{n\mathbf{k}}| \lesssim \left(\mathrm{E_c}\right)^{-s} \qquad \text{and} \qquad |\varepsilon^{\mathrm{E_c}}_{n,\mathbf{k}} - \varepsilon_{n\mathbf{k}}| \lesssim \left(\mathrm{E_c}\right)^{-s}, \tag{4.3.3}$$

where the convergence rate $s \geqslant 0$ depends on the regularity of the effective potential $V \in L^\infty_{\mathrm{per}}(\Omega)$.

Unfortunately, in spite of the availability of the error estimate (4.3.3), a closer study of the Galerkin discretizations (4.3.1) and (4.3.2) reveals a serious deficiency that may not have been immediately apparent: the approximate energy bands $\left\{\breve{\varepsilon}^{\mathrm{E_c}}_n\right\}_{n \in M_{\mathrm{E_c}}(0)}$ and $\left\{\varepsilon^{\mathrm{E_c}}_n\right\}_{n \in M_{\mathrm{E_c}}(k)}$ do not preserve **Properties one** and **two** of the exact energy bands $\{\varepsilon_n\}_{n \in \mathbb{N}^*}$ respectively. An example of this phenomenon is displayed in Figure 4.2A where we plot the exact and approximate ground state energy bands for a simple one-dimensional example.



(A) Approximate and exact energy bands

(B) The $\mathbb{L}$-periodic potential $V \in L^\infty_{\mathrm{per}}(\Omega)$

Figure 4.2 – Lowest energy bands for a simple 1-D example with effective potential $V \in L^\infty_{\mathrm{per}}(\Omega)$ as shown. The effective potential satisfies the regularity property $V \in H^{1-\epsilon}_{\mathrm{per}}(\Omega)$ for every $\epsilon > 0$.

The core problem is that while the exact fibers $\mathrm{H}_{\mathbf{k}}$ and $\mathrm{H}_{\mathbf{k}+\mathbf{G}}$ are unitarily equivalent for all $\mathbf{k} \in \mathbb{R}^d$ and $\mathbf{G} \in \mathbb{L}^*$, the same is not always true for the uniform-basis projected fibers. Indeed, $\breve{\mathrm{H}}^{\mathrm{E_c}}_{\mathbf{k}} = \Pi_{\mathrm{E_c}} \mathrm{H}_{\mathbf{k}} \Pi_{\mathrm{E_c}}$ is not, in general, unitarily equivalent to $\breve{\mathrm{H}}^{\mathrm{E_c}}_{\mathbf{k}+\mathbf{G}}$ as can readily be verified by a direct calculation in the case $V \equiv 0$. Unitary equivalence is, conversely, preserved for the $\mathbf{k}$-basis projected fibers $\mathrm{H}^{\mathrm{E_c}}_{\mathbf{k}} = \Pi_{\mathbf{k},\mathrm{E_c}} \mathrm{H}_{\mathbf{k}} \Pi_{\mathbf{k},\mathrm{E_c}}$ but in this case, the rank of $\mathrm{H}^{\mathrm{E_c}}_{\mathbf{k}}$ changes as a function of $\mathbf{k}$ (recall Remark 4.3.1). This causes the continuity argument used to prove **Property two** to break down.

In other words, the choice of basis set (uniform or $\mathbf{k}$-dependent) represents a trade-off between the $\mathbb{L}^*$-periodicity and regularity of the resulting approximate energy bands. Both choices are obviously sub-optimal from the point of view of numerical quadrature in the Brillouin zone, and it is therefore of great interest to develop an alternative discretization scheme that might result in approximate energy bands that are both $\mathbb{L}^*$-periodic and of class $\mathscr{C}^r$ for some $r \geqslant 0$. One such methodology which has been proposed in the physics literature (see [Ber+95] for the original proposal and [Jan+16] for a recent presentation) and also implemented in several quantum chemistry simulation softwares (see [Abi; Qbo]) relies on the

idea of modifying the 'diagonal' part of the fiber $\mathrm{H}_{\mathbf{k}}$ in a controlled manner. A systematic description of this discretization method is the subject of the next section.

4.4 Operator Modification Approach

We begin this section by defining a one-dimensional "blow-up" function that will be central to the construction of a modified Hamiltonian operator. Throughout this section, we will assume the settings of Sections 4.2 and 4.3.

Definition 4.4.1 (Blow-up function)**.**
Let $m \in \mathbb{N}$, let f_2 denote the quadratic monomial, i.e., $f_2(x) = x^2$ for all $x \in \mathbb{R}$, and denote by $\hbar \colon \left[\frac{1}{2}, 1\right) \to \mathbb{R}$ a function with the following four properties:

1. It holds that $\hbar \in \mathscr{C}^m\left(\frac{1}{2}, 1\right)$.
2. It holds that $\lim_{x \to 1^-} \left((1-x)^m \, \hbar(x)\right) = +\infty$.
3. It holds that $\hbar(x) \geq f_2(x)$ for all $x \in \left(\frac{1}{2}, 1\right)$.
4. For all $j \in \{0, \ldots, m\}$ it holds that $\hbar^{(j)}\left(\frac{1}{2}\right) = f_2^{(j)}\left(\frac{1}{2}\right)$, where $\hbar^{(j)}(\cdot)$ and $f_2^{(j)}(\cdot)$ denote the j^{th} derivative of $\hbar$ and f_2 respectively.

Then we define the blow-up function $\mathscr{G} \colon \mathbb{R} \to \mathbb{R}$ as the mapping given by

$$\mathscr{G}(x) = \begin{cases} f_2(x) & \text{for } |x| \in [0, \frac{1}{2}] \ \cup \ [1, \infty), \\ \hbar(|x|) & \text{for } |x| \in \left(\frac{1}{2}, 1\right), \end{cases} \tag{4.4.1}$$

where we have suppressed the dependency of $\mathscr{G}$ on m and $\hbar$ by assuming once and for all that m and $\hbar$ are fixed for the remainder of our analysis.

Consider Definition 4.4.1 of the blow-up function $\mathscr{G}$. We emphasize three properties of $\mathscr{G}$ that will be useful in the sequel: first, that it is of class $\mathscr{C}^m$ on the interval $[0, 1) \subset \mathbb{R}$; second, that it is *point-wise* bounded below by the quadratic map $x \mapsto x^2$ on all of $\mathbb{R}$, and third that it blows up as $x \to 1^-$ at a rate greater than $\frac{1}{(1-x)^m}$.

Remark 4.4.1 (Blow-up functions in the physics literature). It is pertinent at this point to contrast our rather general definition of the blow-up function $\mathscr{G}$ with those that have been proposed in the literature (see [Ber+95; Jan+16]) and implemented in electronic structure calculation codes (see [Abi; Qbo]). In fact, the functions $\widetilde{\mathscr{G}}$ proposed in [Ber+95; Jan+16] are not 'blow-up' functions at all, in the sense that $\lim_{x \to 1^-} \widetilde{\mathscr{G}}(x) \neq +\infty$. Instead, both papers propose the use of the error function to construct $\widetilde{\mathscr{G}}$ such that $\lim_{x \to 1} \widetilde{\mathscr{G}}(x) = c \gg 1$ but with $c < \infty$. The implementation in the quantum chemistry code QBOX [Qbo] is based on similar ideas. In contrast, the software suite ABINIT [Abi] employs a true 'blow-up' function that satisfies the conditions of Definition 4.4.1 for $m = 1$.

We will now propose a modified Galerkin discretization for the eigenvalue problem (4.2.4). To this end, we first require a definition, and we recall in particular Definition 4.3.2 of the $\mathbf{k}$-dependent basis set on $L^2_{\text{per}}(\Omega)$ and Notation 4.3.2.

Definition 4.4.2 (Modified Hamiltonian operator)**.**
Let $\mathrm{E_c} > 0$ and let the blow-up function $\mathscr{G}$ be defined according to Equation (4.4.1). For each $\mathbf{k} \in \mathbb{R}^d$, we define the operator $\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G}, \mathrm{E_c}} \colon \mathrm{X}_{\mathbf{k}}^{\mathrm{E_c}} \to \mathrm{X}_{\mathbf{k}}^{\mathrm{E_c}}$ as the mapping with the property that

$$\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G}, \mathrm{E_c}} := \Pi_{\mathbf{k}, \mathrm{E_c}} \left(\mathrm{E_c} \, \mathscr{G}\left(\frac{|-\imath\nabla + \mathbf{k}|}{\sqrt{2\mathrm{E_c}}} \right) + V \right) \Pi_{\mathbf{k}, \mathrm{E_c}}. \tag{4.4.2}$$

Remark 4.4.2. Consider the setting of Definition 4.4.2. In Equation (4.4.2), the term $\mathscr{G}\left(\frac{|-\imath\nabla+\mathbf{k}|}{\sqrt{2\mathrm{E_c}}}\right)$ should be understood in the sense of functional calculus. In particular, given some $\Phi = \sum_{\substack{\mathbf{G}\in\mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}|^2<\mathrm{E_c}}} \widehat{\Phi}_{\mathbf{G}} e_{\mathbf{G}} \in \mathrm{X}_{\mathbf{k}}^{\mathrm{E_c}} \subset H^2_{\mathrm{per}}(\Omega)$, we have

$$\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}\Phi = \Pi_{\mathbf{k},\mathrm{E_c}}\left(\sum_{\substack{\mathbf{G}\in\mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}|^2<\mathrm{E_c}}} \widehat{\Phi}_{\mathbf{G}}\left(\mathrm{E_c}\mathscr{G}\left(\frac{|\mathbf{G}+\mathbf{k}|}{\sqrt{2\mathrm{E_c}}}\right)+V\right) e_{\mathbf{G}}\right).$$

Additionally, recalling the definition of the Bloch-Floquet fibers $\mathrm{H}_{\mathbf{k}}$, $\mathbf{k}\in\mathbb{R}^d$ given by Equation (4.2.3), we notice that for each $\mathbf{k}\in\mathbb{R}^d$, thanks to the definition of the blow-up function $\mathscr{G}$, we have that $\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}} \geqslant \mathrm{H}_{\mathbf{k}}^{\mathrm{E_c}} := \Pi_{\mathbf{k},\mathrm{E_c}}\mathrm{H}_{\mathbf{k}}\Pi_{\mathbf{k},\mathrm{E_c}}$, i.e., for all $\Phi\in\mathrm{X}_{\mathbf{k}}^{\mathrm{E_c}}\subset H^1_{\mathrm{per}}(\Omega)$ it holds that

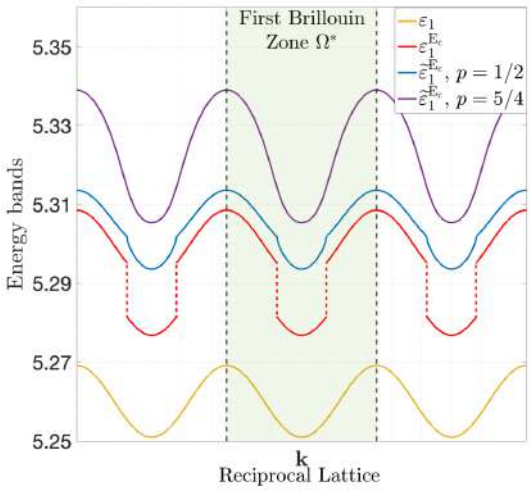
$$\left(\Phi, \widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}\Phi\right)_{L^2_{\mathrm{per}}(\Omega)} \geqslant \left(\Phi, \mathrm{H}_{\mathbf{k}}^{\mathrm{E_c}}\Phi\right)_{L^2_{\mathrm{per}}(\Omega)}.$$

k-dependent modified Galerkin discretization of eigenvalue problem (4.2.4)

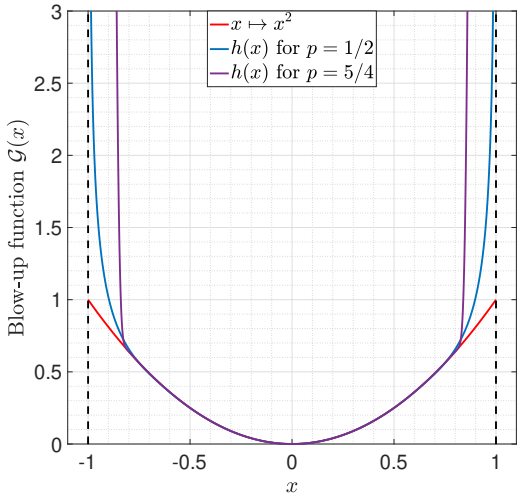
Let $\mathrm{E_c}>0$, let the blow-up function $\mathscr{G}$ be defined according to Equation (4.4.1), and let the modified Hamiltonian operator $\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}$, $\mathbf{k}\in\mathbb{R}^d$ be defined through Definition 4.4.2. We seek an $L^2_{\mathrm{per}}(\Omega)$-orthonormal basis $\left\{\left(\widetilde{\varepsilon}_{n,\mathbf{k}}^{\mathrm{E_c}}, \widetilde{u}_{n,\mathbf{k}}^{\mathrm{E_c}}\right)\right\}\subset\mathbb{R}\times\mathrm{X}_{\mathbf{k}}^{\mathrm{E_c}}$ of eigenmodes of $\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}$:

$$\begin{aligned} \widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}\widetilde{u}_{n,\mathbf{k}}^{\mathrm{E_c}} &= \widetilde{\varepsilon}_{n,\mathbf{k}}^{\mathrm{E_c}}\widetilde{u}_{n,\mathbf{k}}^{\mathrm{E_c}} \qquad \text{and} \\ \left(\widetilde{u}_{n,\mathbf{k}}^{\mathrm{E_c}}, \widetilde{u}_{m,\mathbf{k}}^{\mathrm{E_c}}\right)_{L^2_{\mathrm{per}}(\Omega)} &= \delta_{nm} \qquad \forall n,m\in\{1,\dots,M_{\mathrm{E_c}}(\mathbf{k})\}. \end{aligned} \tag{4.4.3}$$

The eigenvalue problem (4.4.3) can now be solved for different choices of the parameter $\mathrm{E_c}>0$ and blow-up function $\mathscr{G}$. Figure 4.3 displays the approximations of the lowest energy band $\mathbf{k}\mapsto\varepsilon_{1,\mathbf{k}}$ for two different choices of $\mathscr{G}$ and the same $\mathrm{E_c}>0$ and effective potential $V\in L^\infty_{\mathrm{per}}(\Omega)$ as chosen to produce Figure 4.2. The most interesting feature of Figure 4.3 is the fact that– in contrast to the approximate energy band $\varepsilon_1^{\mathrm{E_c}}$– the approximate energy band $\widetilde{\varepsilon}_1^{\mathrm{E_c}}$ remains $\mathbb{L}^*$-periodic, while also appearing to no longer be discontinuous.



(A) Approximate and exact energy bands



(B) Examples of blow-up functions $\hbar$.

Figure 4.3 – Lowest energy bands for the same 1D effective potential V used to produce Figure 4.2. The blow-up functions were of the form $\hbar(x)=\mathrm{C}(1-x)^{-p}$ in the vicinity of 1^-, with $\mathrm{C}>0$.

In the next section, we will present our two main results on the analysis of the modified discretization (4.4.3). Our first result is on the error analysis of this modified discretization where we prove that the

approximate modified energy bands $\{\widetilde{\varepsilon}_n^{\mathrm{E_c}}\}_{n\in\mathbb{N}^*}$ converge with the expected asymptotic rate to the exact energy bands $\{\varepsilon_n\}_{n\in\mathbb{N}^*}$ in a point-wise sense when the cutoff energy $\mathrm{E_c}$ goes to infinity. Our second result is a precise characterization of regularity properties of each energy band $\widetilde{\varepsilon}_n^{\mathrm{E_c}}$ with respect to blow-up functions of different singularity orders.

4.5 Main Results on the Analysis of the Operator Modification Approach

Throughout this section, we will assume the setting of Section 4.4. We recall in particular Definition 4.4.1 of the blow-up function $\mathscr{G}$ as well as the modified Hamiltonian matrices $\{\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}\}_{\mathrm{E_c}>0,\mathbf{k}\in\mathbb{R}^d}$ defined through Equation (4.4.2).

Our first main result concerns the error analysis of the $\mathbf{k}$-dependent modified Galerkin discretization (4.4.3) of the exact eigenvalue problem (4.2.4).

Theorem 4.5.1 (Error estimate for approximate, modified energy bands)**.**

Consider the settings of the exact eigenvalue problem (4.2.4) and the modified discretization (4.4.3) with a blow-up function $\mathscr{G}$ satisfying the conditions in Definition 4.4.1. Assume that the effective potential $V\in L^\infty_{\mathrm{per}}(\Omega)\cap H^r_{\mathrm{per}}(\Omega)$ for some $r>\frac{d}{4}-1$. Let $n\in\mathbb{N}^*$, and for each $\mathbf{k}\in\mathbb{R}^d$, let the subspace $\mathrm{Y}_n^{\mathbf{k}}\subset H^2_{\mathrm{per}}(\Omega)$ be defined as the span of the first n eigenfunctions of the exact fiber $\mathrm{H}_{\mathbf{k}}$, i.e.,

$$\mathrm{Y}_n^{\mathbf{k}} := \mathrm{span}\{u_{j,\mathbf{k}}\colon j\in\{1,\dots,n\}\}.$$

Then there exists $\mathrm{E_c}^*>0$ and a constant $\mathrm{C}>0$ such that for every $\mathrm{E_c}\geqslant \mathrm{E_c}^*$ and all $\mathbf{k}\in\mathbb{R}^d$ it holds that

$$0\leqslant \widetilde{\varepsilon}_{n,\mathbf{k}}^{\mathrm{E_c}}-\varepsilon_{n,\mathbf{k}}\leqslant\left(\frac{\mathrm{C}}{\mathrm{E_c}}\right)^{r+1-\frac{d}{4}}\max_{\substack{\Phi\in\mathrm{Y}_n^{\mathbf{k}}\\ \|\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}}\|\Phi\|^2_{H^{r+2}_{\mathrm{per}}(\Omega)}. \tag{4.5.1}$$

An immediate consequence of Theorem 4.5.1 is that the modified energy bands $\{\widetilde{\varepsilon}_n^{\mathrm{E_c}}\}_{n\in\mathbb{N}^*}$ converge at the same asymptotic rate as the unmodified energy bands $\{\varepsilon_n^{\mathrm{E_c}}\}_{n\in\mathbb{N}^*}$, with respect to $\mathrm{E_c}$, to the exact bands $\{\varepsilon_n\}_{n\in\mathbb{N}^*}$. Additionally, Theorem 4.5.1 informs us that the approximate energy bands $\{\widetilde{\varepsilon}_n^{\mathrm{E_c}}\}_{n\in M_{\mathrm{E_c}}^-}$ are bounded functions of $\mathbb{R}^d$ (recall from Remark 4.3.1 that $M_{\mathrm{E_c}}^-$ denotes the minimal dimension of the $\mathbf{k}$-dependent approximation space $\mathrm{X}_{\mathbf{k}}^{\mathrm{E_c}}$). This latter fact will be of use in the proof of our second main result (see Section 4.7).

Next, we present our second main result, which concerns the regularity properties of these energy bands.

Theorem 4.5.2 (Regularity of approximate, modified energy bands)**.**

Consider the setting of the $\mathbf{k}$-dependent modified discretization (4.4.3) with a blow-up function $\mathscr{G}$ satisfying the conditions in Definition 4.4.1. Let $\mathrm{E_c}>0$ be such that $M_{\mathrm{E_c}}^->0$, and let $n\in\{1,\dots,M_{\mathrm{E_c}}^-\}$. If $\mathbf{k}_0\in\mathbb{R}^d$ is such that

$$\widetilde{\varepsilon}_{n,\mathbf{k}_0}^{\mathrm{E_c}}\neq\widetilde{\varepsilon}_{\widetilde{n},\mathbf{k}_0}^{\mathrm{E_c}}\quad\forall\widetilde{n}\in\{1,\dots,M_{\mathrm{E_c}}^-\}\text{ with }\widetilde{n}\neq n\qquad\big(\text{no band crossings at }(\mathbf{k}_0,\widetilde{\varepsilon}_{n,\mathbf{k}_0}^{\mathrm{E_c}})\big),$$

then the approximate energy band $\widetilde{\varepsilon}_n^{\mathrm{E_c}}$ is of class $\mathscr{C}^m$ in a neighborhood of $\mathbf{k}_0$.

If on the other hand, $\mathbf{k}_0\in\mathbb{R}^d$ is such that

$$\exists\widetilde{n}\in\{1,\dots,M_{\mathrm{E_c}}^-\}\text{ with }\widetilde{n}\neq n\colon\quad\widetilde{\varepsilon}_{n,\mathbf{k}_0}^{\mathrm{E_c}}=\widetilde{\varepsilon}_{\widetilde{n},\mathbf{k}_0}^{\mathrm{E_c}}\qquad\big(\text{band crossing at }(\mathbf{k}_0,\widetilde{\varepsilon}_{n,\mathbf{k}_0}^{\mathrm{E_c}})\big),$$

then the approximate energy band $\widetilde{\varepsilon}_n^{\mathrm{E_c}}$ is

$$\begin{cases}\text{Lipschitz continuous in a neighborhood of }\mathbf{k}_0 & \text{if } m\geqslant 1,\\ \text{continuous in a neighborhood of }\mathbf{k}_0 & \text{otherwise.}\end{cases} \tag{4.5.2}$$

Theorem 4.5.2 indicates that by designing a blow-up function $\mathscr{G}$ that satisfies Properties (1)-(4) from Definition 4.4.1, and in particular has a blow-up singularity of the correct order, we can obtain modified energy bands $\{\widetilde{\varepsilon}_n^{\mathrm{E_c}}\}_{n\in\mathbb{N}^*}$ of arbitrarily high regularity away from band crossings. Moreover, thanks to Theorem 4.5.1, we also have a precise *a priori* characterization of the error in a given band $\widetilde{\varepsilon}_n^{\mathrm{E_c}}$ with respect to varying cutoff energies $\mathrm{E_c}$.

In order to prove the continuity result in Theorem 4.5.2, we will make use of the following general lemma, which is also valid for non-Hermitian matrices and could be used to extend the present analysis to more advanced electronic structure models such as the so-called GW model (see [CGS16] for a mathematical analysis of the latter model). Note however that the continuity result in Theorem 4.5.2 can also be obtained by using arguments specific to Hermitian matrices based on spectral inequalities and residual estimates.

Lemma 1. Let $M \in \mathbb{N}^*$, let $p \in \mathbb{N}^*$ be such that $1 \leq p < M$, and let $(\mathrm{H}_n)_{n\in\mathbb{N}} \in \left(\mathbb{C}^{M\times M}\right)^{\mathbb{N}}$ be a sequence of matrices that admit the block decomposition

$$\mathrm{H}_n = \left[\begin{array}{c|c} \mathrm{A}_n & \mathrm{B}_n \\ \hline \widetilde{\mathrm{B}_n} & \mathrm{C}_n \end{array}\right],$$

where $\mathrm{A}_n \in \mathbb{C}^{p\times p}$, $\mathrm{B}_n \in \mathbb{C}^{p\times(M-p)}$, $\widetilde{\mathrm{B}_n} \in \mathbb{C}^{(M-p)\times p}$, and $\mathrm{C}_n \in \mathbb{C}^{(M-p)\times(M-p)}$ are sub-matrices such that

$$\exists \mathrm{A} \in \mathbb{R}^{p\times p} \quad \text{such that} \quad \lim_{n\to\infty} \|\mathrm{A}_n - \mathrm{A}\|_2 = 0,$$

$$\sup_{n\in\mathbb{N}} \|\mathrm{B}_n\|_2 < \infty, \quad \sup_{n\in\mathbb{N}} \|\widetilde{\mathrm{B}}_n\|_2 < \infty,$$

$$\mathrm{C}_n \text{ is invertible for each } n \text{ and } \lim_{n\to\infty} \|\mathrm{C}_n^{-1}\|_2 = 0,$$

with $\|\cdot\|_2$ denoting the usual matrix 2-norm. Then

1. for every $\rho > 0$ sufficiently small, there exists $N(\rho) \in \mathbb{N}$ such that for any eigenvalue λ^{A} of the matrix A with algebraic multiplicity $Q \in \mathbb{N}^*$ and all $n \geqslant N(\rho)$, the open disc $\mathbb{B}_\rho(\lambda^{\mathrm{A}}) \subset \mathbb{C}$ contains exactly Q eigenvalues of the matrix H_n counting algebraic multiplicities;
2. for every $\Upsilon > 0$ sufficiently large, there exists $\widetilde{N}(\Upsilon) \in \mathbb{N}$ such that for all $n \geqslant \widetilde{N}(\Upsilon)$, there are exactly $M - p$ eigenvalues of H_n with magnitude larger than or equal to Υ.

Before stating the proofs of Theorems 4.5.1 and 4.5.2 and Lemma 1, we will present some numerical results on the use of the operator modification approach that we have described. The aim of these numerical studies is to provide numerical support for the conclusions of our main results Theorems 4.5.1 and 4.5.2. These numerical studies are the subject of the next section.

4.6 Numerical Results

Throughout this section, we assume the setting described in Sections 4.2-4.5. Our goal is now two-fold. First, we wish to present numerical results supporting the conclusions of Theorem 4.5.1 and Theorem 4.5.2. Second, we would like to demonstrate the effectiveness of the operator modification methodology described in Section 4.4 for computing the energy bands of realistic materials such as graphene and face-centered cubic (FCC) silicon crystals.

4.6.1 Validation of theoretical results in one spatial dimension

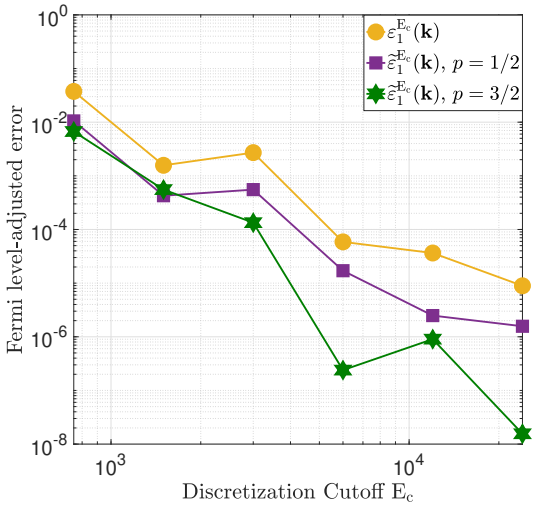
We begin by considering a simple one-dimensional geometric setting. We set $\mathbb{L} = \mathbb{Z}$ which results in $\mathbb{L}^* = 2\pi\mathbb{Z}$, $\Omega = [-\frac{1}{2}, \frac{1}{2})$ and $\Omega^* = [-\pi, \pi)$. The effective potential V is chosen such that $V \in L^\infty_{\mathrm{per}}(\Omega) \cap H^{1-\epsilon}_{\mathrm{per}}(\Omega)$ for all $\epsilon > 0$. Figure 4.2B in Section 4.3 displays a plot of the chosen potential and

demonstrates that the magnitude of V remains between -50 and 200. For all subsequent simulations, the eigenvalue solver tolerance was set to machine (double) precision and reference eigenvalues $\{\varepsilon_{n,k}\}_{n\in\mathbb{N}^*,\ \mathbf{k}\in\mathbb{R}}$ were computed using the uniform Galerkin discretization (4.3.1) with $\mathrm{E_c} = 72,000$. Unless stated otherwise, the approximate bands were computed using $\mathrm{E}_c = 750$ and $\mathbf{k}$-point mesh-width equal to $\Delta = 10^{-3}$. For comparison, the average kinetic energy of the reference lowest energy band is about 12. All blow-up functions $\mathscr{G}$ have regularity $\mathscr{C}^6$ on the interval $(0,1)$ and are of the form $\mathrm{C}(1-x)^{-p}$ in the vicinity of 1^-, with $\mathrm{C} > 0$.

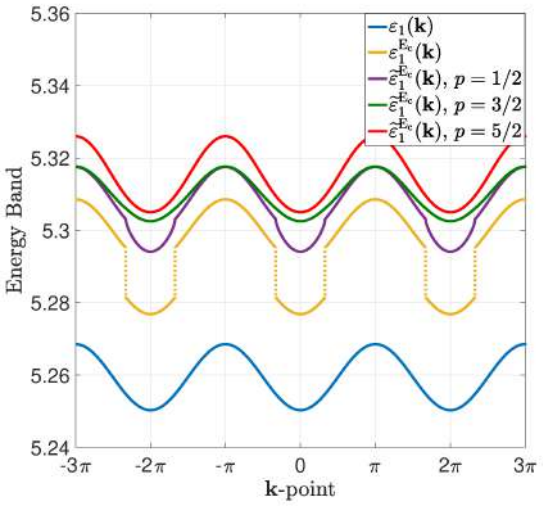
Error convergence with respect to $\mathrm{E_c}$

Our first set of numerical experiments is designed to demonstrate the dependence of the eigenvalue errors in the operator modification approach as a function of the discretization parameter $\mathrm{E_c}$. We compute the lowest energy bands $\widetilde{\varepsilon}_1^{\mathrm{E_c}}$ and $\varepsilon_1^{\mathrm{E_c}}$ for different values of the cutoff energy $\mathrm{E_c}$. Additionally, assuming a single electron per unit cell, we also compute the Fermi levels $\widetilde{\mu}_{\mathrm{F}}^{\mathrm{E_c}}, \mu_{\mathrm{F}}^{\mathrm{E_c}}$ corresponding to the energy bands $\widetilde{\varepsilon}_1^{\mathrm{E_c}}$ and $\varepsilon_1^{\mathrm{E_c}}$ respectively. Note that the band diagrams of periodic physical systems such as this are well-defined only up to an arbitrary additive shift since the potential V is itself defined up to an additive constant in solid-state physics. In band structure calculations, this reference is usually taken as the Fermi level of the system. Therefore, in order to evaluate the accuracy of the modified operator methods we consider the *Fermi-level shifted* errors

$$\int_{\Omega^*} \left|(\varepsilon_{n,\mathbf{k}} - \mu_{\mathrm{F}}) - \left(\widetilde{\varepsilon}_{n,\mathbf{k}}^{\mathrm{E_c}} - \widetilde{\mu}_{\mathrm{F}}^{\mathrm{E_c}}\right)\right| d\mathbf{k} \quad \text{and} \quad \int_{\Omega^*} \left|(\varepsilon_{n,\mathbf{k}} - \mu_{\mathrm{F}}) - \left(\varepsilon_{n,\mathbf{k}}^{\mathrm{E_c}} - \mu_{F}^{\mathrm{E_c}}\right)\right| d\mathbf{k}.$$



(A) Fermi level-adjusted error of the $\mathbf{k}$-dependent and modified energy bands as a function of $\mathrm{E_c}$.

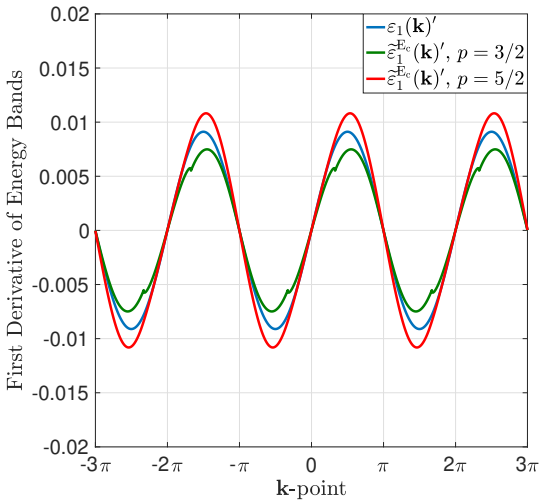


(B) Lowest modified energy bands for the $\mathbf{k}$-dependent discretization and modified operator discretization.

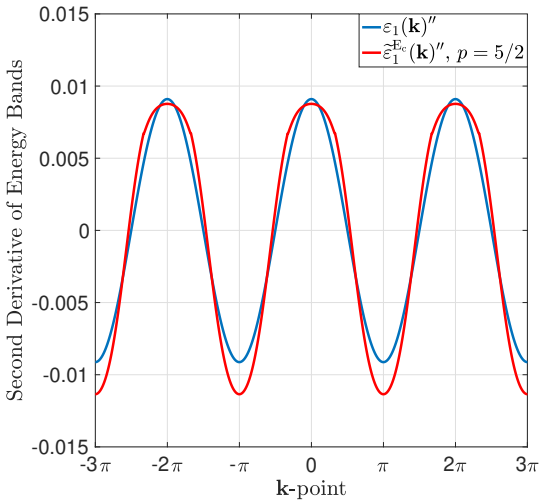
Figure 4.4 – Fermi level-adjusted error in lowest energy bands as a function of $\mathrm{E_c}$ (left) and the lowest energy bands for different blow-up functions of the form $\hbar(x) = \mathrm{C}(1-x)^{-p}$ in the vicinity of 1^-. Note that $\varepsilon_1^{\mathrm{E_c}}$ has a jump discontinuity, while the modified energy bands $\widetilde{\varepsilon}_1^{\mathrm{E_c}}$ are at least continuous.

Figure 4.4A displays our results for two different choice of blow-up function $\mathscr{G}\colon \mathbb{R} \to \mathbb{R}$, one of which has a singularity blow-up of order $|\cdot|^{-\frac{1}{2}}$ and thus satisfies Properties (1)-(4) in Definition 4.4.1 for $m = 0$, and the other with a singularity blow-up of order $|\cdot|^{-\frac{3}{2}}$ which thus satisfies Properties (1)-(4) in Definition 4.4.1 for $m = 1$. We observe from Figure 4.4A that the asymptotic convergence rate with respect to $\mathrm{E_c}$ of both the $\mathbf{k}$-dependent Galerkin discretization scheme (4.3.2) and the modified discretization scheme (4.4.3) are identical, and thus the use of the operator modification approach does not result in any asymptotic degradation of the discretization error. Additionally, we see that for a given cutoff energy $\mathrm{E_c}$, the error of the $\mathbf{k}$-dependent Galerkin discretization (4.3.2) is strictly larger than that of the modified discretization (4.4.3).

Regularity of energy bands as a function of blow-up function singularity



(A) First derivative of the lowest modified energy bands for different choices of the blow-up function $\mathscr{G}$.



(B) Second derivative of the lowest modified energy bands for a higher order blow-up function $\mathscr{G}$.

Figure 4.5 – First and second derivatives of the lowest energy bands for different blow-up functions of the form $\hbar(x) = \mathrm{C}(1-x)^{-p}$ in the vicinity of 1^-, with $\mathrm{C} > 0$

. (Left) The energy band $\widetilde{\varepsilon}_1^{\mathrm{E_c}}$ produced using $p = \frac{3}{2}$ is of class $\mathscr{C}^1(\mathbb{R})$ since it has a kink in the first derivative. (Right) The energy band $\widetilde{\varepsilon}_1^{\mathrm{E_c}}$ produced using $p = \frac{5}{2}$ is of class $\mathscr{C}^2(\mathbb{R})$ since it has a kink in the second derivative.

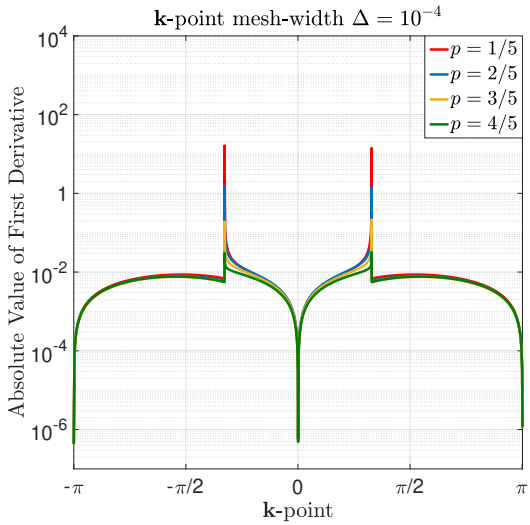
Our next set of numerical simulations is designed to support the conclusion of Theorem 4.5.2 concerning the regularity of the energy bands produced by the modified Galerkin discretization (4.4.3). We consider the regularity of the lowest energy band $\widetilde{\varepsilon}_1^{\mathrm{E_c}}$ for cutoff energy $\mathrm{E_c} = 750$ and three different choices of blow-up functions $\mathscr{G}\colon \mathbb{R} \to \mathbb{R}$. More precisely, we consider blow-up functions $\mathscr{G}$ that satisfy Properties (1)-(4) from Definition 4.4.1 for $m = 0, 1$, and 2 respectively. Our theoretical results indicate that the resulting energy bands should be of class $\mathscr{C}^0(\mathbb{R})$, $\mathscr{C}^1(\mathbb{R})$ and $\mathscr{C}^2(\mathbb{R})$ respectively since there are no energy band crossings for non-trivial effective potentials in one dimension (see, e.g., [AM76, Chapters 8-9]). Figures 4.4B, 4.5A, and 4.5B display our results and show perfect agreement with the conclusions of Theorem 4.5.2.

Considering the energy bands displayed in Figure 4.4B, it is natural to ask if the use of a blow-up function that satisfies Properties (1)-(4) from Definition 4.4.1 only for $m = 0$ results in energy bands that are *Lipschitz* continuous rather than simply continuous, and a similar question can be asked for the derivatives of the energy bands when using blow-up functions with stronger singularities.

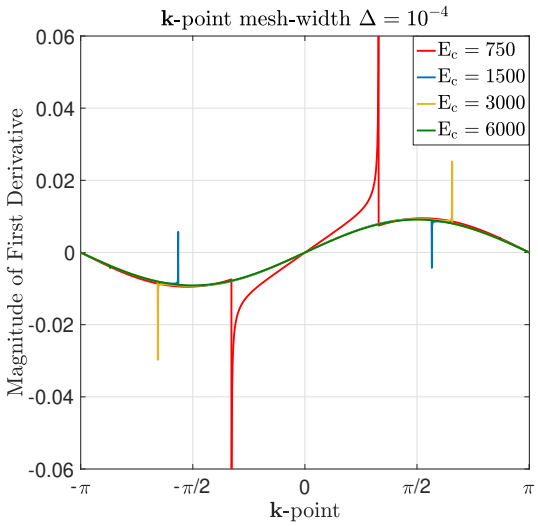
In order to answer this question, we compute the lowest energy band $\widetilde{\varepsilon}_1^{\mathrm{E_c}}$ resulting from the modified discretization (4.4.3) for cutoff energy $\mathrm{E_c} = 750$ and four different choices of blow-up functions $\mathscr{G}$. The blow up functions $\mathscr{G}$ are constructed such that they satisfy Properties (1), (3) and (4) from Definition 4.4.1 for $m = 6$, and such that the singularity of $\mathscr{G}(x)$ at $x = 1$ is of order $|\cdot|^{-p}$ for $p = \frac{1}{5}, \frac{2}{5}, \frac{3}{5}$, and $\frac{4}{5}$ respectively. Figure 4.6A displays the absolute values of the first derivatives of the resulting lowest energy band $\widetilde{\varepsilon}_1^{\mathrm{E_c}}$ for a $\mathbf{k}$-point mesh-width $\Delta = 10^{-4}$. The figure indicates that the derivative of $\widetilde{\varepsilon}_1^{\mathrm{E_c}}$ exhibits peaks at the two points of discontinuity, although the magnitude of the peaks at the points of discontinuity seems to decrease with increasing $\mathrm{E_c}$. In fact, although we do not display the plot here, the magnitudes of these peaks increase as the mesh width Δ is decreased, which indicates that the first derivative is truly unbounded at these points.

Regularity of energy bands as a function of the cutoff energy

The goal of the final set of numerical experiments in this subsection is to explore, for a fixed choice of blow-up function $\mathscr{G}\colon \mathbb{R} \to \mathbb{R}$, how the regularity of the modified energy bands $\{\widetilde{\varepsilon}_n^{\mathrm{E_c}}\}_{n\in\mathbb{N}^*}$ varies as a function of the cutoff energy $\mathrm{E_c}$. To this end, we consider once again the regularity of the lowest energy band $\widetilde{\varepsilon}_1^{\mathrm{E_c}}$ resulting from the modified Galerkin discretization (4.4.3). For these experiments, we set the $\mathbf{k}$-point mesh-width $\Delta = 10^{-4}$, and use a blow-up function $\mathscr{G}$ with blow-up of order $|\cdot|^{-\frac{1}{2}}$.



(A) First derivative of the lowest energy bands for different choices of blow-up function $\mathscr{G}$.



(B) First derivative of the lowest energy band for different choices of cutoff energies E_c.

Figure 4.6 – Non-Lipschitz energy bands ($m = 0$): (Left) First derivative of the lowest energy band for different blow-up functions of the form $\hbar(x) = C(1-x)^{-p}$ for $p < 1$ in the vicinity of 1^-, with $C > 0$. The cutoff energy was taken as $E_c = 750$. (Right) First derivative of the lowest energy band for different choices of cutoff energies E_c and with a blow-up function of the form $\hbar(x) = C(1-x)^{-\frac{1}{2}}$ in the vicinity of 1^-. Although not shown here, in both cases the magnitudes of the peaks increase when **k**-point mesh-width is decreased indicating truly unbounded derivatives.

Figure 4.6B displays the first derivative of the energy band $\widetilde{\varepsilon}_1^{E_c}$ for different values of E_c. The blow-up function chosen for these simulations satisfies Properties (1)-(4) from Definition 4.4.1 only for $m = 0$, so Theorem 4.5.2 indicates that the energy bands should be continuous and not differentiable. It is readily seen that this is indeed the case, although the first derivative of $\widetilde{\varepsilon}_1^{E_c}$ is noticeably regularized by increasing the value of E_c. In fact, the finite magnitude of the peak is a numerical artifact since (although not shown here) the magnitude of the peaks increases as the **k**-point mesh-width is decreased, which indicates that the derivative is truly unbounded for finite E_c at these points. Note that the points of discontinuity occur at different **k**-points depending on the chosen energy cutoff E_c.

4.6.2 Numerical experiments on real materials

We now investigate the effectiveness of the operator modification approach introduced in Section 4.4 on two realistic crystalline systems: a single layer of graphene and face-centered cubic (FCC) crystalline silicon. All subsequent computations are performed using the plane-wave density functional theory (DFT) package *DFTK.jl* [HLC21] in the *Julia* language [Bez+17]. The numerical results of this section can be reproduced by downloading the repository [Mod] and following the instructions therein. The exact definition of the blow-up function used to produce these numerical results is provided in the *utils/blowup.jl* file. A modified kinetic approach guaranteeing $\mathscr{C}^2$ regularity has been directly integrated into the *DFTK.jl* package available at [Dft].

Remark 4.6.1. Two comments are in order:

- *DFTK.jl* makes use of norm-conserving pseudopotentials consisting of a local component (a periodic multiplication operator) and a non-local component (a periodized finite-rank operator). The Kohn-Sham Hamiltonians obtained with *DFTK.jl* are therefore not Schrödinger operators of the form (4.2.1) with V a periodic function. Our theoretical results can easily be extended to handle general pseudopotentials at the price of slightly more cumbersome proofs. For the sake of simplicity, we chose to restrict our analysis to the case of purely local potentials;

- graphene is a real material living in the three-dimensional physical space, but its Bravais lattice is the two-dimensional hexagonal lattice, hence the name 2D materials used to refer to graphene, hexagonal boron nitride, transition metal dichalcogenides, phosphorene, and other atomic-thin layered materials. The Bloch fibers of a periodic 2D material are labelled by a 2D quasi-momentum $\mathbf{k} \in \mathbb{R}^2$ and can be expressed as

$$\mathrm{H}_{\mathbf{k}} = \frac{1}{2}(-i\nabla_{\mathbf{x}_\parallel} + \mathbf{k})^2 - \frac{1}{2}\partial_{x_3}^2 + V$$

  where $\mathbf{x}_\parallel = (x_1, x_2)$ denotes the in-plane position, and x_3 the out-of-plane coordinate. The operators $\mathrm{H}_{\mathbf{k}}$ act on $L^2_{\mathrm{per}}(\Omega)$ where $\Omega = \omega \times \mathbb{R}$ with $\omega \subset \mathbb{R}^2$ being the Wigner-Seitz cell of the 2D Bravais lattice. They do not have compact resolvents and do not admit spectral decompositions of the form (4.2.4). However, the Bloch fibers of the Kohn-Sham Hamiltonian of a real 2D material have discrete eigenvalues below the bottom of their essential spectrum forming the so-called valence bands and low-energy conduction bands. Our theoretical results can thus easily be extended to 2D materials.

Numerical setting

We begin by computing a reference ground-state effective potential using a Kohn-Sham DFT self-consistent field (SCF) procedure with cutoff energy $\mathrm{E_c^{ref}} = 30$ Ha and a fine Monkhorst-Pack $\mathbf{k}$–point grid with 12 points in each sampled dimension. The SCF tolerance is set to machine (double) precision. We choose to work with a Perdew-Burke-Ernzerhof (PBE) functional [PBE96] that is standard in solid state electronic structure computations and Hartwigsen-Goedecker-Teter-Hutter separable dual-space Gaussian pseudopotentials [HGH98]. In a second step, the obtained effective potential is used to construct the reference Hamiltonian $\mathrm{H}_{\mathbf{k}}^{\mathrm{E_c^{ref}}}$ whose eigenvalues will serve as reference data.

The same potential is used to construct the Hamiltonian operator $\mathrm{H}_{\mathbf{k}}^{\mathrm{E_c}}$ defined through (4.3.2), and the modified Hamiltonian operator $\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathrm{E_c}}$ defined through Equation (4.4.3) for a custom set of $\mathbf{k}$–points and for a fixed $\mathrm{E_c} \ll \mathrm{E_c^{ref}}$. The chosen $\mathbf{k}$-points are located on the band-structure paths automatically generated by the *Brillouin.jl* package using the crystallography based method introduced in [Hin+17]. For reference, the $\mathbf{k}$-paths in the Brillouin zone of graphene and FCC silicon and the corresponding band structures are displayed in Figures 4.7 and 4.8 respectively.

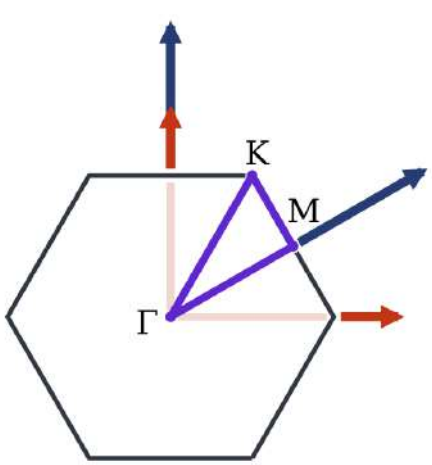
Regularity of energy bands as a function of blow-up function singularity

We choose a very low $\mathrm{E_c} = 5$ Ha in order to clearly highlight the expected irregularities of the energy bands of the standard Hamiltonian operator $\mathrm{H}_{\mathbf{k}}^{\mathrm{E_c}}$. Figure 4.9 displays the abrupt changes in the size of the $\mathbf{k}$-dependent plane-wave basis $\mathcal{B}_{\mathbf{k}}^{\mathrm{E_c}}$ along the band-structure paths of graphene and FCC silicon for this choice of $\mathrm{E_c}$. As in the 1D case, the energy bands produced by the $\mathbf{k}$-dependent Galerkin discretization (4.3.2) are highly irregular, as we read from Figures 4.10A and 4.11A. On the other hand, the modified energy bands produced by the Galerkin discretization (4.4.3) (also displayed in Figures 4.10 and 4.11) appear to be smooth in accordance with the choice of blow-up function $\mathscr{G}$ and in agreement with the theoretical result of Theorem 4.5.2.

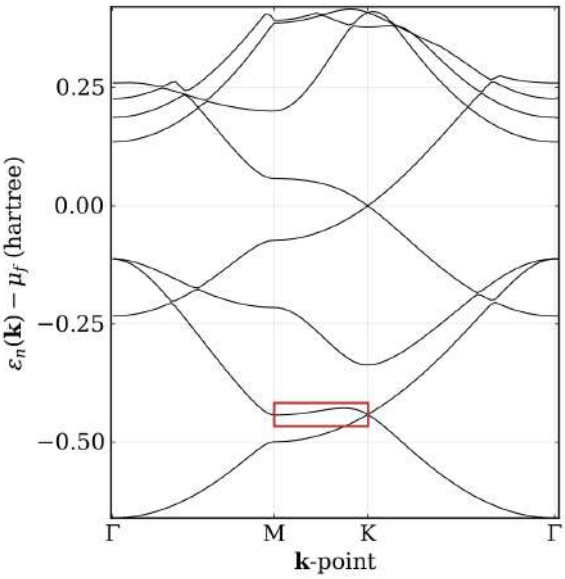
Let us remark that for small values of $\mathrm{E_c}$, the low-energy eigenfunctions $u_{n,\mathbf{k}}$ have non-negligible projections on Fourier modes $e_{\mathbf{G}} \in \mathcal{B}_{\mathbf{k}}^{\mathrm{E_c}}$ with 'modified' kinetic energy that is much higher than the standard kinetic energy. This artificially higher kinetic energy is due to the fact that the blow-up function $\mathscr{G}$ diverges from the $x \mapsto x^2$ curve when x goes to 1. As a consequence, the modified-Hamiltonian energy bands appear significantly over-estimated in comparison to the approximate energy bands produced by the standard Hamiltonian matrix $\mathrm{H}_{\mathbf{k}}^{\mathrm{E_c}}$. As stated previously however, the band diagrams of the considered periodic physical systems are defined up to an arbitrary additive constant, so that a mere shift in energies is unimportant. On the other hand, for large values of $\mathrm{E_c}$, the Fourier modes $e_{\mathbf{G}}$ in $\mathcal{B}_{\mathbf{k}}^{\mathrm{E_c}}$ with over-estimated kinetic energies do not contribute much to the Fourier expansions of the low-energy eigenfunctions $u_{n,\mathbf{k}}$. This suggests that the proper range of application of the modified-operator approach is typically in the regime where $\mathrm{E_c}$ is not too small so that the modified discretization scheme matches the accuracy of the standard Galerkin discretization while ensuring the targeted regularity.

Regularity of energy bands as a function of the volume of the unit cell

Another possible application of the modified-operator approach concerns the computations of energy bands as a function of the volume of the unit cell, that can be used to estimate the macroscopic volumetric mass density of a crystalline material or its bulk modulus [Kax03, Chapter 5.6]. The size of the $\mathbf{k}$-dependent



The graphene **k**-path in the Brillouin zone.



The graphene reference band structure.

Figure 4.7 – Graphene **k**-path (left) and the corresponding band structure (right). The paths are automatically generated by the *Brillouin.jl* package based on crystallographic considerations. The red (resp. blue) arrows display the Cartesian coordinate axes (resp. the reciprocal basis vectors). All bands are shifted so that the Fermi level appears at zero Hartree on the graph. The red box shows the part of the band diagram on which Figure 4.10 will focus.

discretization basis at a given **k**-point depends on the unit cell volume through its associated reciprocal lattice, so that within the standard Galerkin approximation, the energy per unit volume is a rough function of the crystalline parameters. For both graphene and FCC silicon, the unit cell is parameterized by a single lattice parameter a. Figure 4.12 displays the energy of graphene and FCC silicon per unit volume as a function of a around the experimental value a_0 of the equilibrium lattice parameter. In both cases, we observe high oscillations of the energy per unit volume. We see that the modified-operator approach produces much smoother energy curves.

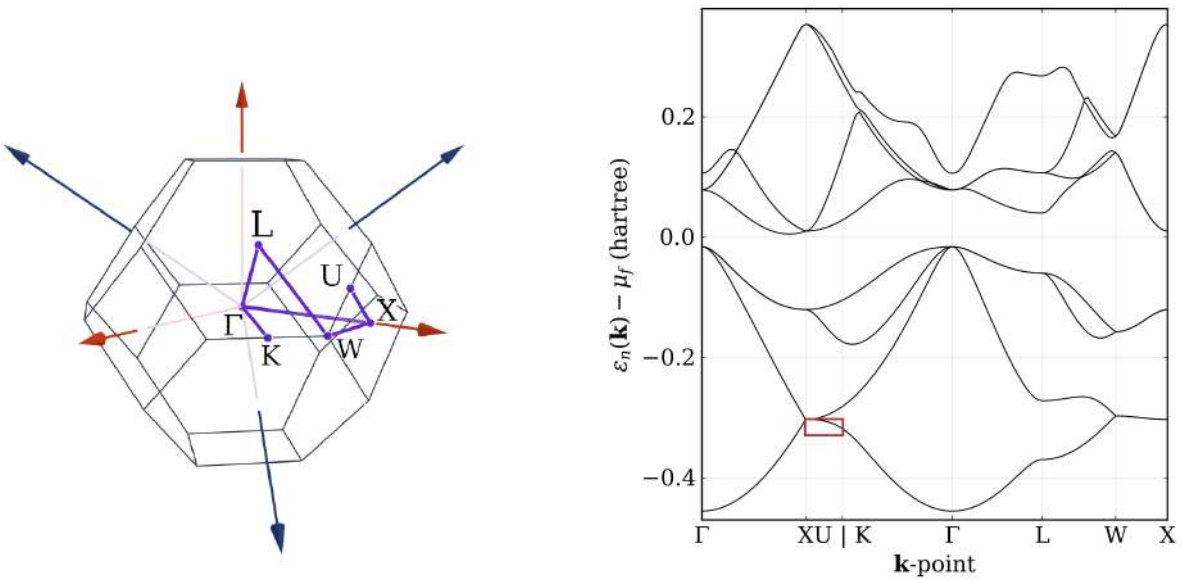
Effect of the operator modification approach on the computational cost

For the final set of numerical experiments, we evaluate the impact of the operator modification strategy on the total computational cost of solving the underlying Bloch fiber eigenvalue problems. To do so, we consider again Graphene and FCC crystalline silicon, and we compute the total number of linear solver iterations required to compute, for a given $\mathrm{E_c}$, the first 8 eigenvalues in the 'modified' discrete eigenvalue problem (4.4.3) for all **k**-points on a Monkhorst-Pack grid in the Brillouin zone. Note that the eigensolver is considered converged when the Frobenius distance between consecutive one-body density matrices falls bellow 10^{-8}. Our results are shown in Figures 4.13A and 4.13B and indicate that while the number of iterations does increase with the blow-up rate, the increase is not catastrophic and largely ranges between 5% and 25% for the blow-up rates considered in this paper. Further testing indicates however that a combination of very high blow-up rates and very high values of $\mathrm{E_c}$ can considerably worsen the convergence of the eigensolver, which suggests again that the proper range of application of this operator modification approach is likely in the moderate blow-up and medium $\mathrm{E_c}$ regime.

4.7 Proofs of the Main Results

We will begin with the proof of the *a priori* error estimate Theorem 4.5.1.

Proof of Theorem 4.5.1.



(A) The silicon **k**-path in the Brillouin zone.

(B) The silicon reference band structure along this path.

Figure 4.8 – FCC crystalline silicon **k**-path (left) and the corresponding band structure (right). The paths are automatically generated by the *Brillouin.jl* package based on crystallographic considerations. Red and blue arrows display the Cartesian coordinate axes and the reciprocal basis vectors respectively. All bands are again shifted so that the Fermi level appears at zero Hartree on the graph, and the red box shows the part of the band diagram on which Figure 4.11 focuses.

The lower bound in Inequality (4.5.1) follows directly from Remark 4.4.2 so we need only prove the upper bound. Also, since both $\widetilde{\varepsilon}_n^{\mathrm{E_c}}$ and ε_n are $\mathbb{L}^*$-periodic functions on $\mathbb{R}^d$, it suffices to establish the error estimate (4.5.1) only for $\mathbf{k} \in \Omega^*$.

From Definition 4.3.2 of the **k**-dependent basis set $\mathcal{B}_{\mathbf{k}}^{\mathrm{E_c}}$ (see also Remark 4.3.1), we deduce the existence of an $\mathrm{E_c'} > 0$ such that for all $\mathbf{k} \in \mathbb{R}^d$, we have that $n = \dim \mathrm{Y}_n^{\mathbf{k}} \leqslant M_{\mathrm{E_c'}}^-$ where we recall that $M_{\mathrm{E_c'}}^-$ denotes the minimal size of the basis $\mathcal{B}_{\mathbf{k}}^{\mathrm{E_c'}}$ over all $\mathbf{k} \in \mathbb{R}^d$. Thus, for each $\mathbf{k} \in \Omega^*$ there also exists $\mathrm{E_c^{\mathbf{k}}} \geqslant \mathrm{E_c'}$ such that

$$\mathrm{Y}_n^{\mathbf{k},\mathrm{E_c}} := \{\Psi := \Pi_{\mathbf{k},\mathrm{E_c^{\mathbf{k}}}}\Phi \colon \Phi \in \mathrm{Y}_n^{\mathbf{k}}\},$$

is an n-dimensional subspace of $H^1_{\mathrm{per}}(\Omega)$. We set $\widetilde{\mathrm{E_c}} := \sup_{\mathbf{k}\in\Omega^*} \mathrm{E_c^{\mathbf{k}}}$. Using now the min-max theorem, we deduce that for all $\mathrm{E_c} \geqslant \widetilde{\mathrm{E_c}}$ and any $\mathbf{k} \in \Omega^*$ it holds that

$$\widetilde{\varepsilon}_{n,\mathbf{k}}^{4\mathrm{E_c}} \leqslant \max_{\substack{\Psi\in\mathrm{Y}_n^{\mathbf{k},\mathrm{E_c}} \\ \|\Psi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} \left(\Psi, \widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},4\mathrm{E_c}}\Psi\right)_{L^2_{\mathrm{per}}(\Omega)}. \tag{4.7.1}$$

Notice here that we consider the modified energy band $\widetilde{\varepsilon}_{n,\mathbf{k}}^{4\mathrm{E_c}}$ which corresponds to the modified Hamiltonian matrix $\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},4\mathrm{E_c}}$. The appearance of the factor 4 is linked to Definition 4.4.1 of the blow-up function $\mathscr{G}\colon \mathbb{R} \to \mathbb{R}$ in which we impose that $\mathscr{G}(x) = x^2$ for $|x| \in [0, \frac{1}{2}] \cup [1, \infty)$. Indeed, as we shall demonstrate below, considering the modified Hamiltonian matrix $\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},4\mathrm{E_c}}$ rather than $\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}$ allows us to obtain the required error estimates through relatively simple arguments. Let us nevertheless remark that any pre-factor greater than or equal to 4 is equally valid for the subsequent arguments.

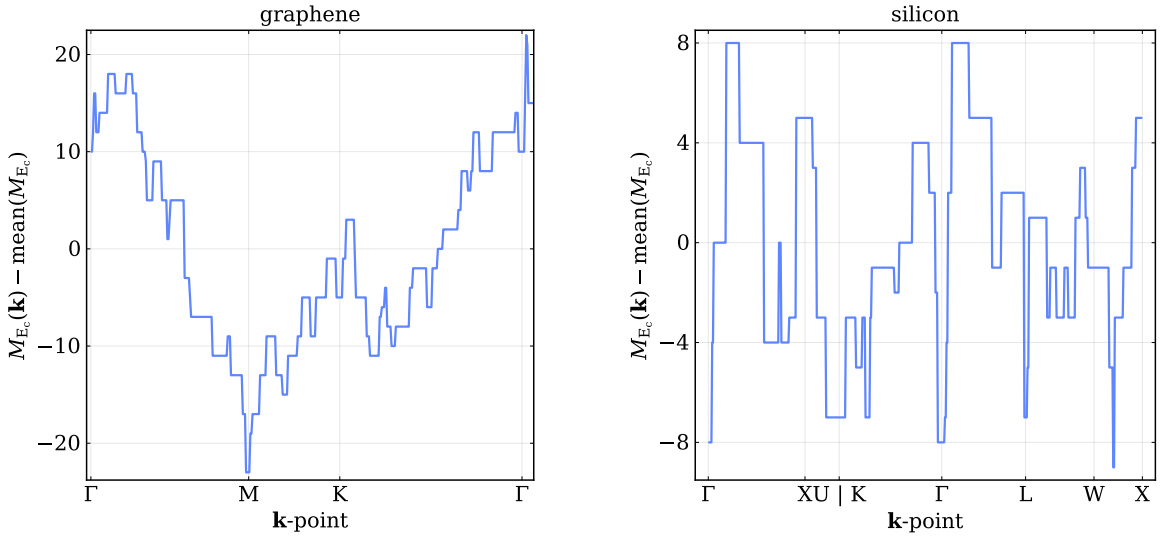


Figure 4.9 – Distance to the mean value of $M_{\mathrm{E_c}}(\cdot)$ along standard $\mathbf{k}$-path with a small $\mathrm{E_c} = 5$ Ha, showing the abrupt changes in the cardinality of the $\mathbf{k}$-dependent plane-wave basis set $\mathcal{B}_{\mathbf{k}}^{\mathrm{E_c}}$. The test cases are (left) a single layer of graphene and (right) FCC crystalline silicon. The $\mathbf{k}$-path is automatically generated by *DFTK.jl* using the *Brillouin.jl* package.

Returning to Inequality (4.7.1), we deduce that for all $\mathrm{E_c} \geqslant \widetilde{\mathrm{E_c}}$ and any $\mathbf{k} \in \Omega^*$ it holds that

$$\begin{aligned}
\widetilde{\varepsilon}_{n,\mathbf{k}}^{4\mathrm{E_c}} &\leqslant \max_{\substack{\Phi \in \mathrm{Y}_n^{\mathbf{k}} \\ \|\Pi_{\mathbf{k},\mathrm{E_c}}\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} \left(\Phi, \left(\Pi_{\mathbf{k},\mathrm{E_c}} \widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},4\mathrm{E_c}} \Pi_{\mathbf{k},\mathrm{E_c}}\right)\Phi\right)_{L^2_{\mathrm{per}}(\Omega)} \\
&\leqslant \max_{\substack{\Phi \in \mathrm{Y}_n^{\mathbf{k}} \\ \|\Pi_{\mathbf{k},\mathrm{E_c}}\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} \left(\Phi, \left(\Pi_{\mathbf{k},\mathrm{E_c}} \widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},4\mathrm{E_c}} \Pi_{\mathbf{k},\mathrm{E_c}} - \mathrm{H}_{\mathbf{k}}\right)\Phi\right)_{L^2_{\mathrm{per}}(\Omega)} \\
&\quad + \max_{\substack{\Phi \in \mathrm{Y}_n^{\mathbf{k}} \\ \|\Pi_{\mathbf{k},\mathrm{E_c}}\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} \Big(\underbrace{(\Phi, \mathrm{H}_{\mathbf{k}}\Phi)_{L^2_{\mathrm{per}}(\Omega)} - \varepsilon_n(\mathbf{k})\|\Phi\|^2_{L^2_{\mathrm{per}}(\Omega)}}_{\leqslant 0} + \varepsilon_n(\mathbf{k})\|\Phi\|^2_{L^2_{\mathrm{per}}(\Omega)} \Big).
\end{aligned}$$

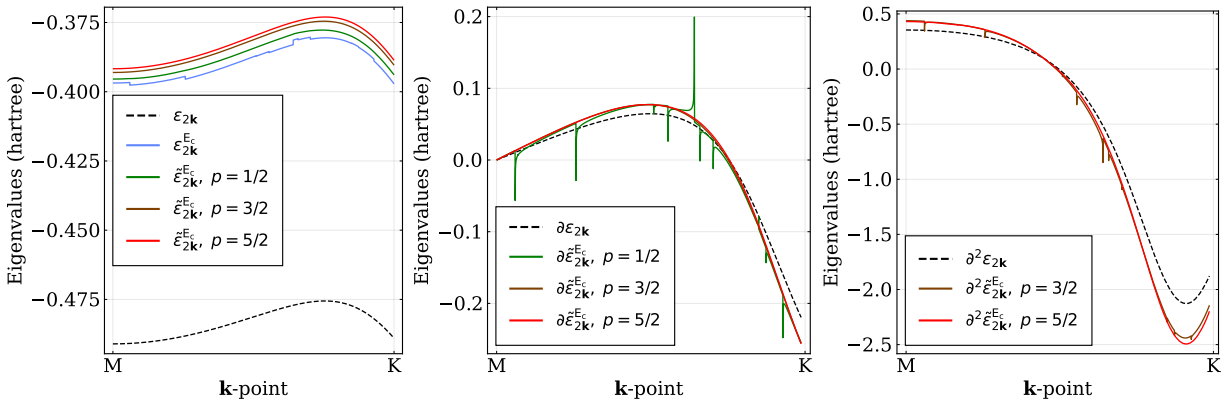
It therefore follows that for all $\mathrm{E_c} \geqslant \widetilde{\mathrm{E_c}}$ and for any $\mathbf{k} \in \Omega^*$ we have

$$\begin{aligned}
\widetilde{\varepsilon}_{n,\mathbf{k}}^{4\mathrm{E_c}} - \varepsilon_{n,\mathbf{k}} \leqslant & \max_{\substack{\Phi \in \mathrm{Y}_n^{\mathbf{k}} \\ \|\Pi_{\mathbf{k},\mathrm{E_c}}\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} \underbrace{\left(\Phi, \left(\Pi_{\mathbf{k},\mathrm{E_c}} \widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},4\mathrm{E_c}} \Pi_{\mathbf{k},\mathrm{E_c}} - \mathrm{H}_{\mathbf{k}}\right)\Phi\right)_{L^2_{\mathrm{per}}(\Omega)}}_{:=(\mathrm{I})} \\
& + \max_{\substack{\Phi \in \mathrm{Y}_n^{\mathbf{k}} \\ \|\Pi_{\mathbf{k},\mathrm{E_c}}\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} \varepsilon_n(\mathbf{k}) \underbrace{\|\Pi^{\perp}_{\mathbf{k},\mathrm{E_c}}\Phi\|^2_{L^2_{\mathrm{per}}(\Omega)}}_{:=(\Pi)}.
\end{aligned} \tag{4.7.2}$$

Let us first simplify the term (I) for an arbitrary $\Phi \in \mathrm{Y}_n^{\mathbf{k}}$ and $\mathbf{k} \in \Omega^*$. We begin by rewriting the term (I) as

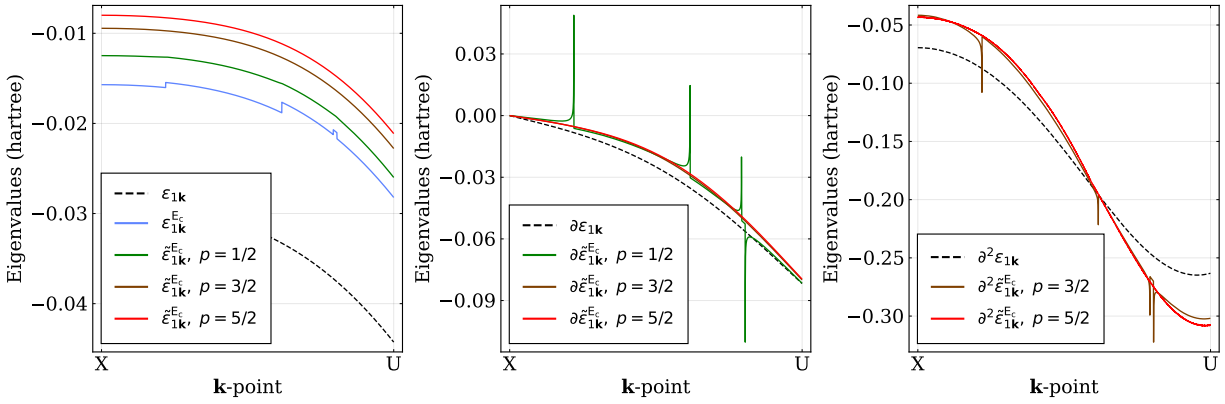
$$(\mathrm{I}) = \left(\Phi, \left(\Pi_{\mathbf{k},\mathrm{E_c}} \widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},4\mathrm{E_c}} \Pi_{\mathbf{k},\mathrm{E_c}} - \mathrm{H}_{\mathbf{k}}^{\mathrm{E_c}}\right)\Phi\right)_{L^2_{\mathrm{per}}(\Omega)} + \left(\Phi, \left(\mathrm{H}_{\mathbf{k}}^{\mathrm{E_c}} - \mathrm{H}_{\mathbf{k}}\right)\Phi\right)_{L^2_{\mathrm{per}}(\Omega)}. \tag{4.7.3}$$

We claim that the first term on the right hand side is identically zero. Indeed, recalling Definition 4.3.2 of the basis $\mathcal{B}_{\mathbf{k}}^{\mathrm{E_c}}$ as well as the definitions of the exact fiber and modified Hamiltonian matrix given by Equations (4.2.3) and (4.4.2) respectively, and using the fact that $\Pi_{\mathbf{k},4\mathrm{E_c}}\Pi_{\mathbf{k},\mathrm{E_c}} = \Pi_{\mathbf{k},\mathrm{E_c}} = \Pi_{\mathbf{k},\mathrm{E_c}}\Pi_{\mathbf{k},4\mathrm{E_c}}$ we



(A) Energy band (B) First derivative (C) Second derivative

Figure 4.10 – Comparison of the first and second derivatives of the second band of graphene between points **M** and **K** of the band-structure for the **k**-dependent and modified discretization schemes.



(A) Energy band (B) First derivative (C) Second derivative

Figure 4.11 – Comparison of the first and second derivatives of the first band of silicon between points **X** and **U** of the band-structure for the **k**-dependent and modified discretization schemes.

deduce that

$$\left(\Pi_{\mathbf{k},\mathrm{E_c}}\widetilde{\mathrm{H}}^{\mathscr{G},4\mathrm{E_c}}_{\mathbf{k}}\Pi_{\mathbf{k},\mathrm{E_c}} - \mathrm{H}^{\mathrm{E_c}}_{\mathbf{k}}\right)\Phi$$
$$= \Pi_{\mathbf{k},\mathrm{E_c}}\left(\sum_{\substack{\mathbf{G}\in\mathbb{L}^*\\ \frac{1}{2}|\mathbf{k}+\mathbf{G}|^2<\mathrm{E_c}}} \widehat{\Phi}_{\mathbf{G}}\left(4\mathrm{E_c}\mathscr{G}\left(\frac{|\mathbf{k}+\mathbf{G}|}{\sqrt{8\mathrm{E_c}}}\right) - \frac{1}{2}|\mathbf{G}+\mathbf{k}|^2\right)e_{\mathbf{G}}\right).$$

Since $\frac{1}{2}|\mathbf{k}+\mathbf{G}|^2 < \mathrm{E_c} \implies \frac{|\mathbf{k}+\mathbf{G}|}{\sqrt{8\mathrm{E_c}}} < \frac{1}{2}$, we obtain from Definition 4.4.1 of $\mathscr{G}$ that

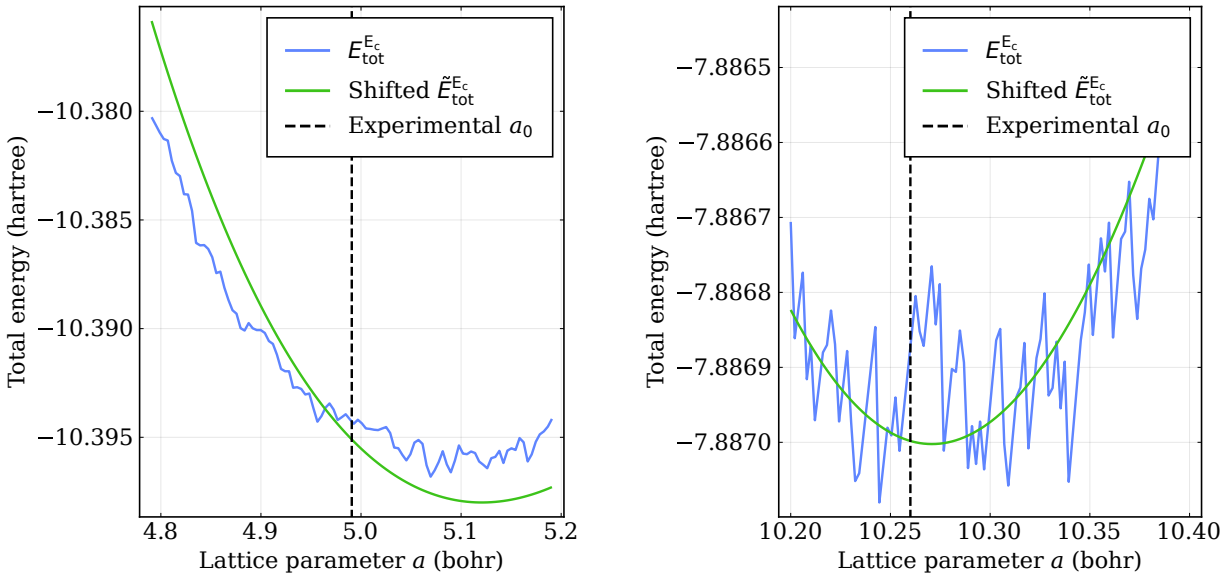
$$\left(\Pi_{\mathbf{k},\mathrm{E_c}}\widetilde{\mathrm{H}}^{\mathscr{G},4\mathrm{E_c}}_{\mathbf{k}}\Pi_{\mathbf{k},\mathrm{E_c}} - \mathrm{H}^{\mathrm{E_c}}_{\mathbf{k}}\right)\Phi = 0, \tag{4.7.4}$$

which implies the claimed result.

The simplification of the second term on the right is classical but we perform it for the sake of completeness. For ease of exposition, in the sequel we will use $\mathrm{C} > 0$ to denote a generic constant whose value may change from step to step but that remains independent of $\mathbf{k} \in \mathbb{R}^d$, $\mathrm{E_c} > 0$, and $\Phi \in \mathrm{Y}^{\mathbf{k}}_{\mathrm{n}}$.

Let us begin by noting that since $\mathrm{H}^{\mathrm{E_c}}_{\mathbf{k}} = \Pi_{\mathbf{k},\mathrm{E_c}}\mathrm{H}_{\mathbf{k}}\Pi_{\mathbf{k},\mathrm{E_c}}$ by definition, Equation (4.7.3) implies that

$$\begin{aligned}(\mathrm{I}) &= -2\left(\Phi, \left(\Pi^{\perp}_{\mathbf{k},\mathrm{E_c}}\mathrm{H}_{\mathbf{k}}\Pi_{\mathbf{k},\mathrm{E_c}}\right)\Phi\right)_{L^2_{\mathrm{per}}(\Omega)} - \left(\Phi, \left(\Pi^{\perp}_{\mathbf{k},\mathrm{E_c}}\mathrm{H}_{\mathbf{k}}\Pi^{\perp}_{\mathbf{k},\mathrm{E_c}}\right)\Phi\right)_{L^2_{\mathrm{per}}(\Omega)}\\ &\leqslant 2\|\Pi^{\perp}_{\mathbf{k},\mathrm{E_c}}V\Pi_{\mathbf{k},\mathrm{E_c}}\Phi\|_{L^2_{\mathrm{per}}(\Omega)}\|\Pi^{\perp}_{\mathbf{k},\mathrm{E_c}}\Phi\|_{L^2_{\mathrm{per}}(\Omega)} - \min\{\varepsilon_1(\mathbf{k}),0\}\|\Pi^{\perp}_{\mathbf{k},\mathrm{E_c}}\Phi\|^2_{L^2_{\mathrm{per}}(\Omega)},\end{aligned} \tag{4.7.5}$$



(A) Graphene

(B) FCC crystalline silicon

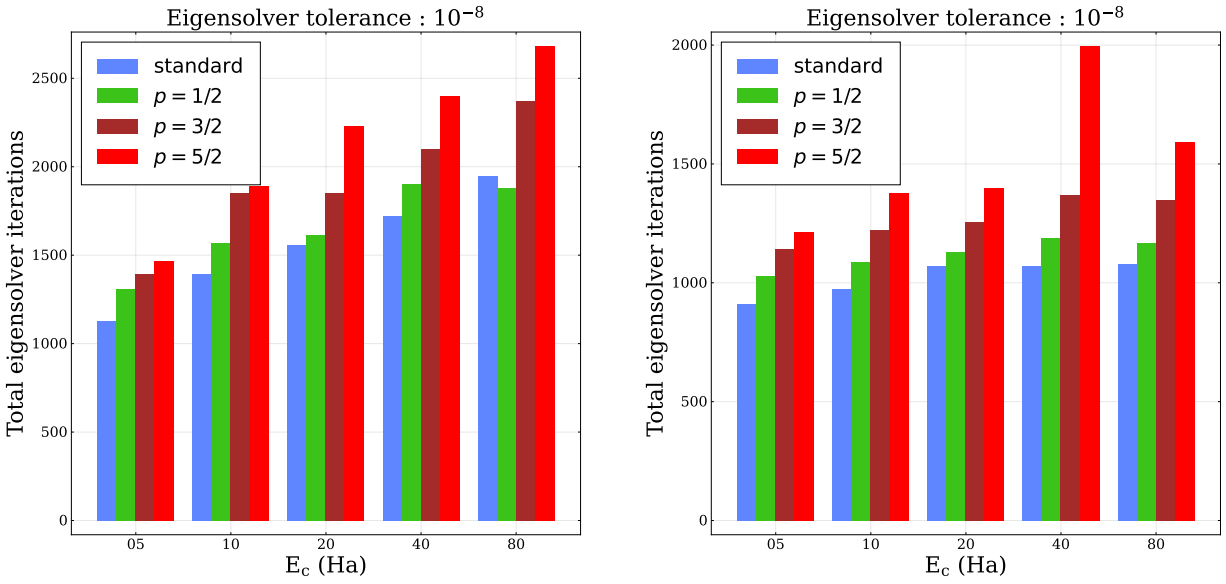
Figure 4.12 – Energy per unit volume of graphene and FCC silicon as a function of the lattice parameter a for the $\mathbf{k}$-dependent and modified discretization schemes (Equations (4.3.2) and (4.4.3) respectively). The blow-up function $\mathscr{G}$ has a blow-up of order $|\cdot|^{-\frac{3}{2}}$ and the cutoff energy is set to $E_c = 5$ Ha. For the sake of legibility, the total energy of the system for the modified Hamiltonian is shifted to the mean value of the standard Hamiltonian total energy over the sample of parameters a. The empirical value a_0 of the equilibrium lattice parameter is also indicated.

We simplify this last estimate term-by-term. Since $V \in H^r_{\text{per}}(\Omega)$, we deduce that each exact eigenfunction $u_{n,\mathbf{k}}$, $n \in \mathbb{N}^*$ is an element of $H^{r+2}_{\text{per}}(\Omega)$ (see, e.g., [CCM10]). It follows that $\Phi \in H^{r+2}_{\text{per}}(\Omega)$ so that we can write

$$\begin{aligned}
\|\Pi^{\perp}_{\mathbf{k},\mathrm{E_c}}\Phi\|^2_{L^2_{\text{per}}(\Omega)} = \sum_{\substack{\mathbf{G}\in\mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}|^2 \geqslant \mathrm{E_c}}} |\widehat{\Phi}_{\mathbf{G}}|^2 &\leqslant \left(\frac{1}{2\mathrm{E_c}}\right)^{r+2} \sum_{\substack{\mathbf{G}\in\mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}|^2 \geqslant \mathrm{E_c}}} \left(1+|\mathbf{k}+\mathbf{G}|^2\right)^{r+2} |\widehat{\Phi}_{\mathbf{G}}|^2 \\
&\leqslant \left(\frac{\mathrm{C}}{\mathrm{E_c}}\right)^{r+2} \sum_{\substack{\mathbf{G}\in\mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}|^2 \geqslant \mathrm{E_c}}} \left(1+|\mathbf{G}|^2\right)^{r+2} |\widehat{\Phi}_{\mathbf{G}}|^2 \qquad (4.7.6) \\
&\leqslant \left(\frac{\mathrm{C}}{\mathrm{E_c}}\right)^{r+2} \|\Phi\|^2_{H^{r+2}_{\text{per}}(\Omega)},
\end{aligned}$$

where the second line follows from the fact that $\mathbf{k} \in \Omega^*$ so that, in particular, $|\mathbf{k}+\mathbf{G}| < \operatorname{diam}(\Omega^*) + |\mathbf{G}|$ $\forall \mathbf{G} \in \mathbb{L}^*$.

In order to simplify the term in Inequality (4.7.5) involving the potential V, we will use similar tactics.



(A) Graphene

(B) FCC crystaline silicon

Figure 4.13 – The total number of eigensolver iterations required to solve the unmodified eigenvalue problem (4.3.2) and 'modified' eigenvalue problem (4.4.3) for all $\mathbf{k}$-points on a Monkhorst-Pack grid with a fixed Kohn-Sham reference potential. Results for different choices of blow-up function $\mathscr{G}$ and values of E_c are displayed for both graphene and FCC crystalline silicon. The eigensolver is an LOPBCG algorithm [HL06] with a simple version of the diagonal Tetter-Payne-Allan preconditioner [TPA89]. Both are default choices in the *DFTK.jl* package, and can be found in the *DFTK.jl/src/eigen* folder. Note that the eigensolver algorithm is considered converged when the Frobenius distance between consecutive one-body density matrices falls bellow 10^{-8}.

As a first step, we make use of the fact that $V \in L^\infty_{\mathrm{per}}(\Omega)$ is a multiplicative operator so that we may write

$$\begin{aligned}
\|\Pi^\perp_{\mathbf{k},E_c} V \Pi_{\mathbf{k},E_c} \Phi\|^2_{L^2_{\mathrm{per}}(\Omega)} &= \sum_{\substack{\mathbf{G}' \in \mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}'|^2 \geqslant E_c}} \left| \sum_{\substack{\mathbf{G} \in \mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}|^2 < E_c}} \widehat{V}_{\mathbf{G}'-\mathbf{G}} \widehat{\Phi}_{\mathbf{G}} \right|^2 \\
&\leqslant \sum_{\substack{\mathbf{G}' \in \mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}'|^2 \geqslant E_c}} \left| \sum_{\substack{\mathbf{G} \in \mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}|^2 < E_c}} 1 \sum_{\substack{\mathbf{G} \in \mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}|^2 < E_c}} |\widehat{V}_{\mathbf{G}'-\mathbf{G}}|^2 |\widehat{\Phi}_{\mathbf{G}}|^2 \right|, \\
&\leqslant \sum_{\substack{\mathbf{G}' \in \mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}'|^2 | \geqslant E_c}} C E_c^{\frac{d}{2}} \sum_{\substack{\mathbf{G} \in \mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}|^2 < E_c}} |\widehat{V}_{\mathbf{G}'-\mathbf{G}}|^2 |\widehat{\Phi}_{\mathbf{G}}|^2 \\
&= C E_c^{\frac{d}{2}} \sum_{\substack{\mathbf{G} \in \mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}|^2 < E_c}} |\widehat{\Phi}_{\mathbf{G}}|^2 \sum_{\substack{\mathbf{G}' \in \mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}'|^2 \geqslant E_c}} |\widehat{V}_{\mathbf{G}'-\mathbf{G}}|^2 \\
&= C E_c^{\frac{d}{2}} \sum_{\substack{\mathbf{G} \in \mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}|^2 < E_c}} |\widehat{\Phi}_{\mathbf{G}}|^2 \sum_{\substack{\mathbf{R} \in \mathbb{L}^* \\ \frac{1}{2}|\mathbf{k}+\mathbf{G}+\mathbf{R}|^2 \geqslant E_c}} |\widehat{V}_{\mathbf{R}}|^2, \qquad (4.7.7)
\end{aligned}$$

where the second step follows from the Cauchy-Schwarz inequality and the third step follows by bounding the number of lattice points inside a d-dimensional ball of radius $\sqrt{2E_c}$ centered at $\mathbf{k} \in \mathbb{R}^d$. Using now a

similar calculation to the one carried out to arrive at Inequality (4.7.6), we deduce that

$$\begin{aligned}\sum_{\substack{\mathbf{R}\in\mathbb{L}^*\\ \frac{1}{2}|\mathbf{k}+\mathbf{G}+\mathbf{R}|^2\geqslant \mathrm{E_c}}} |\widehat{V}_{\mathbf{R}}|^2 &\leqslant \left(\frac{1}{2\mathrm{E_c}}\right)^r \sum_{\substack{\mathbf{R}\in\mathbb{L}^*\\ \frac{1}{2}|\mathbf{k}+\mathbf{G}+\mathbf{R}|^2\geqslant \mathrm{E_c}}} \left(1+|\mathbf{k}+\mathbf{G}+\mathbf{R}|^2\right)^r |\widehat{V}_{\mathbf{R}}|^2\\ &\leqslant \left(\frac{1}{2\mathrm{E_c}}\right)^r \sum_{\substack{\mathbf{R}\in\mathbb{L}^*\\ \frac{1}{2}|\mathbf{k}+\mathbf{G}+\mathbf{R}|^2\geqslant \mathrm{E_c}}} \left(1+2|\mathbf{k}+\mathbf{R}|^2+2|\mathbf{G}|^2\right)^r |\widehat{V}_{\mathbf{R}}|^2\\ &\leqslant \left(\frac{\mathrm{C}}{\mathrm{E_c}}\right)^r \sum_{\substack{\mathbf{R}\in\mathbb{L}^*\\ \frac{1}{2}|\mathbf{k}+\mathbf{G}+\mathbf{R}|^2\geqslant \mathrm{E_c}}} \left(1+|\mathbf{k}+\mathbf{R}|^2\right)^r |\widehat{V}_{\mathbf{R}}|^2 + \left(1+|\mathbf{G}|^2\right)^r |\widehat{V}_{\mathbf{R}}|^2\\ &\leqslant \left(\frac{\mathrm{C}}{\mathrm{E_c}}\right)^r \left(\|V\|^2_{H^r_{\mathrm{per}}(\Omega)} + \left(1+|\mathbf{G}|^2\right)^r \|V\|^2_{L^2_{\mathrm{per}}(\Omega)}\right).\end{aligned}$$

Plugging in this last expression in Inequality (4.7.7) easily allows us to deduce that

$$\|\Pi^{\perp}_{\mathbf{k},\mathrm{E_c}} V \Pi_{\mathbf{k},\mathrm{E_c}} \Phi\|^2_{L^2_{\mathrm{per}}(\Omega)} \leqslant \left(\frac{\mathrm{C}}{\mathrm{E_c}}\right)^{r-\frac{d}{2}} \|V\|^2_{H^r_{\mathrm{per}}(\Omega)} \|\Phi\|^2_{H^{r+2}_{\mathrm{per}}(\Omega)}. \tag{4.7.8}$$

Finally, an identical calculation to the one used to obtain Inequality (4.7.6) can be used to simplify the term (II) in Inequality (4.7.2). Combining now Estimates (4.7.6)-(4.7.8) with Inequalities (4.7.2)-(4.7.5), we deduce that for all $\mathrm{E_c} \geqslant \widetilde{\mathrm{E_c}}$ and any $\mathbf{k} \in \Omega^*$ it holds that

$$\begin{aligned}\widetilde{\varepsilon}^{4\mathrm{E_c}}_{n,\mathbf{k}} - \varepsilon_{n,\mathbf{k}} \leqslant \max_{\substack{\Phi\in \mathrm{Y}^{\mathbf{k}}_n\\ \|\Pi_{\mathbf{k},\mathrm{E_c}}\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} &\left(\left(\frac{\mathrm{C}}{\mathrm{E_c}}\right)^{r+1-\frac{d}{4}} \|V\|_{H^r_{\mathrm{per}}(\Omega)} \|\Phi\|^2_{H^{r+2}_{\mathrm{per}}(\Omega)}\right.\\ &\left. - \left(\frac{\mathrm{C}}{\mathrm{E_c}}\right)^{r+2} \min\{\varepsilon_1(\mathbf{k}),0\} \|\Phi\|^2_{H^{r+2}_{\mathrm{per}}(\Omega)}\right)\\ + \max_{\substack{\Phi\in \mathrm{Y}^{\mathbf{k}}_n\\ \|\Pi_{\mathbf{k},\mathrm{E_c}}\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} &\left(\frac{\mathrm{C}}{\mathrm{E_c}}\right)^{r+2} \varepsilon_n(\mathbf{k}) \|\Phi\|^2_{H^{r+2}_{\mathrm{per}}(\Omega)}.\end{aligned}$$

Collecting terms, we obtain an estimate of the form

$$\widetilde{\varepsilon}^{4\mathrm{E_c}}_{n,\mathbf{k}} - \varepsilon_{n,\mathbf{k}} \leqslant \left(\frac{\mathrm{C}}{\mathrm{E_c}}\right)^{r+1-\frac{d}{4}} \max_{\substack{\Phi\in \mathrm{Y}^{\mathbf{k}}_n\\ \|\Pi_{\mathbf{k},\mathrm{E_c}}\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} \|\Phi\|^2_{H^{r+2}_{\mathrm{per}}(\Omega)}.$$

To conclude, we notice that we may write

$$\max_{\substack{\Phi\in \mathrm{Y}^{\mathbf{k}}_n\\ \|\Pi_{\mathbf{k},\mathrm{E_c}}\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} \|\Phi\|^2_{H^{r+2}_{\mathrm{per}}(\Omega)} \leqslant \max_{\substack{\Phi\in \mathrm{Y}^{\mathbf{k}}_n\\ \|\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} \|\Phi\|^2_{H^{r+2}_{\mathrm{per}}(\Omega)} \underbrace{\max_{\substack{\Phi\in \mathrm{Y}^{\mathbf{k}}_n\\ \|\Pi_{\mathbf{k},\mathrm{E_c}}\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} \|\Phi\|^2_{L^2_{\mathrm{per}}(\Omega)}}_{:=(\mathrm{III})},$$

and it is well-known that there exists an upper bound $\mathrm{C}_{M,\mathrm{E_c}} \geqslant 0$ for the term (III), provided that the basis $\mathcal{B}^{\mathrm{E_c}}_{\mathbf{k}}$ satisfies the so-called approximation property (which it does) and that $\mathrm{E_c}$ is larger than some threshold $\widehat{\mathrm{E_c}} \geqslant 0$. Defining appropriate constants and setting the discretization cutoff $\mathrm{E_c^*}$ large enough thus completes the proof. □

We can now turn our attention to the more technical proof of Theorem 4.5.2 which characterizes precisely the regularity of the modified energy bands $\{\widetilde{\varepsilon}^{\mathrm{E_c}}_n\}_{n\in\mathbb{N}^*}$. As stated in Section 4.5, the proof of Theorem 4.5.2 will require the use of Lemma 1 which we now prove.

Proof of Lemma 1.

We begin by claiming that thanks to the assumptions on the matrix C_n, for every $\Upsilon > 0$ there exists a natural number $N(\Upsilon) \in \mathbb{N}$ such that for all $n \geqslant N(\Upsilon)$ and any $\lambda \in \mathbb{C}$ such that $|\lambda| < \Upsilon$, the inverse matrix $(\mathrm{C}_n - \lambda)^{-1}$ exists. Indeed, this is simply a consequence of the fact that for any invertible matrix E, all its eigenvalues are lower bounded in magnitude by $\|E^{-1}\|_2^{-1}$.

Consequently, for any $\Upsilon > 0$ and any $n \geqslant N(\Upsilon)$, we can introduce the open disk $\mathbb{B}_\Upsilon(0) := \{z \in \mathbb{C} \colon |z| < \Upsilon\}$, and define the non-linear function $g_{\Upsilon,n} \colon \mathbb{B}_\Upsilon(0) \to \mathbb{C}$ given by

$$g_{\Upsilon,n}(\lambda) := \det\left(\mathrm{A}_n - \lambda - \mathrm{B}_n \left(\mathrm{C}_n - \lambda\right)^{-1} \widetilde{\mathrm{B}_n}\right) \qquad \forall \lambda \in \mathbb{B}_\Upsilon(0).$$

Using a well-known determinant identity for block matrices we deduce that for all $\Upsilon > 0$, all $n \geqslant N(\Upsilon)$ and any $\lambda \in \mathbb{B}_\Upsilon(0)$ it holds that

$$\det\left(\mathrm{H}_n - \lambda\right) = \det\left(\mathrm{C}_n - \lambda\right) \; g_{\Upsilon,n}(\lambda). \tag{4.7.9}$$

Equation (4.7.9) implies that for any $\Upsilon > 0$ sufficiently large there exists a natural number $N(\Upsilon) \in \mathbb{N}$ such that for all $n \geqslant N(\Upsilon)$

$$\lambda \in \mathbb{B}_\Upsilon(0) \text{ is an eigenvalue of } \mathrm{H}_n \iff g_{\Upsilon,n}(\lambda) = 0. \tag{4.7.10}$$

We are now interested in studying the zeros of the function $g_{\Upsilon,n}$ in the asymptotic regime $n \to \infty$. Our goal is to show that the sequence of functions $\{g_{\Upsilon,n}\}_{n\in\mathbb{N}}$ satisfy the hypotheses of Hurwitz's theorem from complex analysis (see, e.g., [Con78, Chapter VII, Theorem 2.5]).

We begin by establishing that for any $\Upsilon > 0$ and any $n \geqslant N(\Upsilon)$, the non-linear function $g_{\Upsilon,n}$ is holomorphic on the open disk $\mathbb{B}_\Upsilon(0)$. To do so, observe that for any $\Upsilon > 0$, all $n \geqslant N(\Upsilon)$ and any $\lambda \in \mathbb{B}_\Upsilon(0)$ we can define the matrix

$$\mathrm{Z}_n(\lambda) := \mathrm{A}_n - \lambda - \mathrm{B}_n \left(\mathrm{C}_n - \lambda\right)^{-1} \widetilde{\mathrm{B}_n}.$$

Recall that the natural number $N(\Upsilon)$ was chosen so that all eigenvalues of C_n are strictly larger in magnitude than Υ and consequently, $(\mathrm{C}_n - \lambda)^{-1}$ exists for all $n \geqslant N(\Upsilon)$ and all $\lambda \in \mathbb{B}_\Upsilon(0) \subset \mathbb{C}$. In view of the assumptions on the sub-matrices $\mathrm{A}_n, \mathrm{B}_n, \widetilde{\mathrm{B}_n}$ and C_n, it therefore follows that the matrix $Z_n(\lambda)$ exists and is bounded for all $n \geqslant N(\Upsilon)$, and has a power series expansion in the disk $\mathbb{B}_\Upsilon(0)$ of the form

$$\mathrm{Z}_n(\lambda) = \mathrm{A}_n - \lambda - \sum_{q=0}^{\infty} \mathrm{B}_n \lambda^q \left(\mathrm{C}_n\right)^{-q-1} \widetilde{\mathrm{B}_n} = \sum_{q=0}^{\infty} \lambda^q \, M_{q,n}, \tag{4.7.11}$$

where each $M_{q,n}$ is a square matrix of dimension $p = \dim \mathrm{A}$.

Equation (4.7.11) implies that each entry of the matrix $\mathrm{Z}_n(\lambda)$ is itself a holomorphic function of λ with a power series expansion valid in $\mathbb{B}_\Upsilon(0)$. Moreover, since $g_{\Upsilon,n}(\lambda) = \det\left(\mathrm{Z}_n(\lambda)\right)$ for each $\lambda \in \mathbb{B}_\Upsilon(0)$, and the determinant is a polynomial of the entries of the underlying matrix, we deduce that for any $\Upsilon > 0$ and all $n \geqslant N(\Upsilon)$ the non-linear function $g_{\Upsilon,n}$ is indeed holomorphic on $\mathbb{B}_\Upsilon(0) \subset \mathbb{C}$.

Next, we claim that on any compact set $K \subset \mathbb{B}_\Upsilon(0)$, the sequence of non-linear functions $\{g_{\Upsilon,n}\}_{\ell\in\mathbb{N}}$ converges uniformly to the characteristic polynomial of A_n, i.e.,

$$\lim_{n\to\infty} \sup_{\lambda\in K} \left|\det\left(\mathrm{A}_n - \lambda\right) - g_{\Upsilon,n}(\lambda)\right| = 0. \tag{4.7.12}$$

To prove that Equation (4.7.12) indeed holds, we appeal to a known determinant inequality for differences of matrices (see [IR08, Theorem 2.12]): For any two matrices $E, F \in \mathbb{C}^{p\times p}$, it holds that

$$\left|\det(E) - \det(E+F)\right| \leqslant p\|F\|_2 \; \max\left\{\|E\|_2, \|E+F\|_2\right\}^{p-1}. \tag{4.7.13}$$

Applying Inequality (4.7.13) to our situation yields that for any $\Upsilon > 0$, any compact set $K \subset \mathbb{B}_\Upsilon(0)$, all $\lambda \in K$ and all $n \geqslant N(\Upsilon)$ it holds that

$$\left|\det\left(\mathrm{A}_n - \lambda\right) - g_{\Upsilon,n}(\lambda)\right| \leqslant \; p\left\|\mathrm{B}_n \left(\mathrm{C}_n - \lambda\right)^{-1} \widetilde{\mathrm{B}_n}\right\|_2 \max\left\{\left\|\mathrm{A}_n - \lambda\right\|_2, \left\|Z_n(\lambda)\right\|_2\right\}^{p-1}. \tag{4.7.14}$$

To simplify this last estimate, we first use the definition of the parametrized matrix $\mathrm{Z_n}(\lambda)$ to deduce that for all $n \geqslant N(\Upsilon)$ it holds that

$$\begin{aligned}\max\left\{ \|\mathrm{A}_n - \lambda\|_2 , \|Z_n(\lambda)\|_2 \right\}^{p-1} &\leqslant \left(\|\mathrm{A}_n - \lambda\|_2 + \|\mathrm{B}_n \left(\mathrm{C}_n - \lambda\right)^{-1} \widetilde{\mathrm{B}_n}\|_2 \right)^{p-1} \\ &\leqslant 2^{p-2} \left(\|\mathrm{A}_n - \lambda\|_2^{p-1} + \|\mathrm{B}_n \left(\mathrm{C}_n - \lambda\right)^{-1} \widetilde{\mathrm{B}_n}\|_2^{p-1} \right).\end{aligned}$$

Additionally, recalling the assumptions on the sub-matrix C_n, we see that for all $n \geqslant N(\Upsilon)$ we have that

$$\|\mathrm{B}_n \left(\mathrm{C}_n - \lambda\right)^{-1} \widetilde{\mathrm{B}_n}\|_2 \leqslant \|\mathrm{B}_n\|_2 \, \|\mathrm{C}_n^{-1}\|_2 \, \| \left(I - \lambda \mathrm{C}_n^{-1}\right)^{-1} \|_2 \, \|\widetilde{\mathrm{B}_n}\|_2$$

Using now the estimates derived above together with Inequality (4.7.14) and recalling the boundedness assumptions on the sub-matrices $\mathrm{B}_n, \widetilde{\mathrm{B}_n}$ as well as the convergence result for the sub-matrices $\mathrm{A}_n, \mathrm{C}_n$, we deduce that for any $\Upsilon > 0$ and any compact set $K \subset \mathbb{B}_\Upsilon(0)$, there exists a constant $\Lambda_{\Upsilon,K}$ (which depends also on p) such that for all $n \geqslant N(\Upsilon)$ and all $\lambda \in K$ it holds that

$$|\det\left(\mathrm{A}_n - \lambda\right) - g_{\Upsilon,n}(\lambda)| \leqslant \Lambda_{\Upsilon,K} \, \|\mathrm{C}_n^{-1}\|_2,$$

from which Equation (4.7.12) now readily follows. Let us also emphasize here that since $\lim_{n\to\infty} \mathrm{A}_n = \mathrm{A}$, we have in fact shown that the sequence of non-linear functions $\{g_{\Upsilon,\ell}\}_{\ell\in\mathbb{N}}$ converges uniformly to the characteristic polynomial of A on any compact set $K \subset \mathbb{B}_\Upsilon(0)$.

In order to complete our analysis, we observe that the characteristic polynomial of A is an *entire* function which is not identically zero on any open subset of $\mathbb{C}$. As a consequence, Hurwitz's theorem can be applied: For any $\Upsilon > 0$ and any non-empty open, connected set U such that $\overline{U} \subset \mathbb{B}_\Upsilon(0)$ and $\det(\mathrm{A} - \lambda) \neq 0 \; \forall \lambda \in \partial U$, there exists $N(\Upsilon) \in \mathbb{N}$ such that for all $n \geqslant N(\Upsilon)$, the non-linear function $g_{\Upsilon,n}$ and the characteristic polynomial $\det(\mathrm{A} - \bullet)$ of the matrix A have the same number of zeros in U counting multiplicity. In particular,

(i) for all $\Upsilon > 0$ sufficiently large, there exists $\widetilde{N}(\Upsilon) \in \mathbb{N}$ such that for all $n \geqslant \widetilde{N}(\Upsilon)$ the non-linear function $g_{\Upsilon,n}$ has <u>exactly</u> $p = \dim \mathrm{A}$ zeros counting multiplicity with magnitude strictly smaller than Υ;

(ii) picking a fixed Υ large enough so that $\mathbb{B}_\Upsilon(0)$ contains all roots of the characteristic polynomial of A, for every $\rho > 0$ sufficiently small, there exists $N(\rho) \in \mathbb{N}$ such that for any eigenvalue λ^{A} of the matrix A with algebraic multiplicity $Q \in \mathbb{N}^*$, and all $n \geqslant N(\rho)$, the non-linear function $g_{\Upsilon,n}$ has <u>exactly</u> Q zeros (counting multiplicity) in the open disk $\mathbb{B}_\rho\left(\lambda^{\mathrm{A}}\right) \subset \mathbb{C}$.

The proof now follows easily by making use of Relation (4.7.10).

□

Proof of Theorem 4.5.2.

For clarity of exposition, we will divide this proof into three portions: We will first consider the regularity of the approximate energy bands away from crossings and away from changes in the cardinality of the $\mathbf{k}$-dependent basis sets. This will allows us to deduce, as a corollary, the regularity of the approximate energy bands at crossings but under the assumption that the cardinality of the $\mathbf{k}$-dependent basis set does not change. Lastly, we will consider the regularity of the approximate energy bands in the neighborhood of points where the cardinality of the $\mathbf{k}$-dependent basis sets may change.

For the remainder of this proof we recall the setting of the $\mathbf{k}$-dependent modified Galerkin discretization (4.4.3), we select $\mathrm{E_c} > 0$ such that $M_{\mathrm{E_c}}^- > 0$, and we pick some index $n \in \{1, \ldots, M_{\mathrm{E_c}}^-\}$ and some point $\mathbf{k}_0 \in \mathbb{R}^d$.

Case one: we assume that $\mathbf{k}_0 \in \mathbb{R}^d$ satisfies

$$\text{For all } \mathbf{G} \in \mathbb{L}^* \text{ it holds that } |\mathbf{k}_0 + \mathbf{G}|^2 \neq 2\mathrm{E_c} \quad \text{and}$$

$$\widetilde{\varepsilon}_n^{\mathrm{E_c}}(\mathbf{k}_0) \neq \widetilde{\varepsilon}_{\widetilde{n}}^{\mathrm{E_c}}(\mathbf{k}_0) \quad \forall \widetilde{n} \in \{1, \ldots, M_{\mathrm{E_c}}^-\} \quad \text{with} \quad \widetilde{n} \neq n.$$

In other words, we assume that there is no change in the cardinality of the basis set at $\mathbf{k}_0$ and that there are no band crossings at $\big(\mathbf{k}_0, \widetilde{\varepsilon}_n^{\mathrm{E_c}}(\mathbf{k}_0)\big)$.

We claim that in this case, the approximate energy band $\widetilde{\varepsilon}_n^{\mathrm{E_c}}$ is of class $\mathscr{C}^m$ in a neighborhood of $\mathbf{k}_0$, i.e., $\widetilde{\varepsilon}_n^{\mathrm{E_c}}$ has the same local regularity at $\mathbf{k}_0$ as the blow-up function $\mathscr{G}$ does on the interval $(0,1) \subset \mathbb{R}$. Indeed, this is a straightforward application of the implicit function theorem: We notice that the dimensions of the modified-Hamiltonian matrices $\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}$ do not change in a sufficiently small neighborhood of $\mathbf{k}_0$ and the dependence of this matrix on $\mathbf{k}$ in such a neighborhood is of class $\mathscr{C}^m$ (thanks to the regularity properties of $\mathscr{G}$). Since, additionally, the eigenvalue $\widetilde{\varepsilon}_n^{\mathrm{E_c}}(\mathbf{k}_0)$ is simple, it can be shown that the assumptions of the implicit function theorem hold, and therefore by a classical argument (see, e.g.,[Ser10, Theorem 5.3]) it follows that the approximate energy band $\widetilde{\varepsilon}_n^{\mathrm{E_c}}$ is indeed of class $\mathscr{C}^m$ in a neighborhood of $\mathbf{k}_0$ as claimed.

Let us remark here that a similar argument involving the implicit function theorem yields $\mathscr{C}^m$ regularity, as a function of $\mathbf{k} \in \mathbb{R}^d$, of the (normalized) eigenfunction $\widetilde{u}_{n,\mathbf{k}}^{\mathrm{E_c}}$ associated with the eigenvalue $\widetilde{\varepsilon}_{n,\mathbf{k}}^{\mathrm{E_c}}$ at $\mathbf{k} = \mathbf{k}_0$. A detailed argument can, for instance, be found in [Lax07, Chapter 9, Theorem 8]. This additional fact will of use in the sequel.

Case two: we assume that $\mathbf{k}_0 \in \mathbb{R}^d$ satisfies

$$\text{For all } \mathbf{G} \in \mathbb{L}^* \text{ it holds that } |\mathbf{k}_0 + \mathbf{G}|^2 \neq 2\mathrm{E_c} \quad \text{and}$$

$$\exists\, \widetilde{n} \in \{1, \dots, M_{\mathrm{E_c}}^-\} \text{ with } \widetilde{n} \neq n\colon \quad \widetilde{\varepsilon}_n^{\mathrm{E_c}}(\mathbf{k}_0) = \widetilde{\varepsilon}_{\widetilde{n}}^{\mathrm{E_c}}(\mathbf{k}_0).$$

In other words, we assume that there is no change in the cardinality of the basis set at $\mathbf{k}_0$ but there is a band crossing at $\big(\mathbf{k}_0, \widetilde{\varepsilon}_n^{\mathrm{E_c}}(\mathbf{k}_0)\big)$.

We claim that in this case, the approximate energy band $\widetilde{\varepsilon}_n^{\mathrm{E_c}}$ is either Lipschitz continuous in a neighborhood of $\mathbf{k}_0$ if $m \geqslant 1$ or of class $\mathscr{C}^0$ otherwise. To this end, we notice that the dimensions of the matrix $\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}$ do not change in a sufficiently small neighborhood of $\mathbf{k}_0$ and the dependence of this matrix on $\mathbf{k}$ in such a neighborhood is of class $\mathscr{C}^m$, $m \geqslant 0$ thanks to the regularity properties of $\mathscr{G}$. If $m = 0$, then it follows from a classical argument (see [Whi72, Appendix V, Page 363]) that all approximate energy bands $\{\widetilde{\varepsilon}_n^{\mathrm{E_c}}\}_{n=1}^{M_{\mathrm{E_c}}^-}$ are continuous at $\mathbf{k}_0$. If, on the other hand, $m \geqslant 1$, then the claimed Lipschitz continuity follows from the min-max theorem. Indeed, for any $\mathbf{k} \in \mathbb{R}^d$ let $\widetilde{\mathrm{Y}}_n^{\mathbf{k},\mathrm{E_c}}$ be the span of the first n eigenvectors of the modified Hamiltonian matrix $\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}$, let $\delta > 0$ be a sufficiently small constant such that the basis set $\mathcal{B}_{\mathbf{k}}^{\mathrm{E_c}}$ remains unchanged for all $\mathbf{k}$ in the open ball $\mathbb{B}_\delta(\mathbf{k}_0) \subset \mathbb{R}^d$, and let $\mathbf{k}_1, \mathbf{k}_2 \in \mathbb{B}_\delta(\mathbf{k}_0)$. It then follows from the min-max theorem that

$$\begin{aligned}
\widetilde{\varepsilon}_n^{\mathrm{E_c}}(\mathbf{k}_1) - \widetilde{\varepsilon}_n^{\mathrm{E_c}}(\mathbf{k}_2) &\leqslant \max_{\substack{\Phi \in \widetilde{\mathrm{Y}}_n^{\mathbf{k}_2,\mathrm{E_c}} \\ \|\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} \Big(\Phi, \widetilde{\mathrm{H}}_{\mathbf{k}_1}^{\mathscr{G},\mathrm{E_c}}\Phi\Big)_{L^2_{\mathrm{per}}(\Omega)} - \max_{\substack{\Phi \in \widetilde{\mathrm{Y}}_n^{\mathbf{k}_2,\mathrm{E_c}} \\ \|\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} \Big(\Phi, \widetilde{\mathrm{H}}_{\mathbf{k}_2}^{\mathscr{G},\mathrm{E_c}}\Phi\Big)_{L^2_{\mathrm{per}}(\Omega)} \\
&\leqslant \max_{\substack{\Phi \in \widetilde{\mathrm{Y}}_n^{\mathbf{k}_2,\mathrm{E_c}} \\ \|\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} \Big|\Big(\Phi, (\widetilde{\mathrm{H}}_{\mathbf{k}_1}^{\mathscr{G},\mathrm{E_c}} - \widetilde{\mathrm{H}}_{\mathbf{k}_2}^{\mathscr{G},\mathrm{E_c}})\Phi\Big)_{L^2_{\mathrm{per}}(\Omega)}\Big| \\
&\leqslant \max_{\substack{\Phi \in \widetilde{\mathrm{Y}}_n^{\mathbf{k}_2,\mathrm{E_c}} \\ \|\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} \sum_{\mathbf{G} \in \mathcal{B}_{\mathbf{k}_0}^{\mathrm{E_c}}} |\widehat{\Phi}_{\mathbf{G}}|^2 \mathrm{E_c} \left|\mathscr{G}\left(\frac{|\mathbf{G}+\mathbf{k}_1|}{\sqrt{2\mathrm{E_c}}}\right) - \mathscr{G}\left(\frac{|\mathbf{G}+\mathbf{k}_2|}{\sqrt{2\mathrm{E_c}}}\right)\right| \\
&\leqslant \max_{\mathbf{G} \in \mathcal{B}_{\mathbf{k}_0}^{\mathrm{E_c}}} \mathrm{E_c} \left|\mathscr{G}\left(\frac{|\mathbf{G}+\mathbf{k}_1|}{\sqrt{2\mathrm{E_c}}}\right) - \mathscr{G}\left(\frac{|\mathbf{G}+\mathbf{k}_2|}{\sqrt{2\mathrm{E_c}}}\right)\right| \max_{\substack{\Phi \in \widetilde{\mathrm{Y}}_n^{\mathbf{k}_2,\mathrm{E_c}} \\ \|\Phi\|_{L^2_{\mathrm{per}}(\Omega)}=1}} \|\Phi\|^2_{L^2_{\mathrm{per}}(\Omega)}.
\end{aligned}$$

Here, the third inequality follows directly from Definition (4.4.2) of the modified Hamiltonian matrices together with the fact that $\mathcal{B}_{\mathbf{k}_0}^{\mathrm{E_c}} = \mathcal{B}_{\mathbf{k}_1}^{\mathrm{E_c}} = \mathcal{B}_{\mathbf{k}_2}^{\mathrm{E_c}}$, i.e., the $\mathbf{k}$-dependent basis sets in the open ball $\mathbb{B}_\delta(\mathbf{k}_0)$ are identical.

Using now the fact that $\mathscr{G}$ is of class $\mathscr{C}^1$ and thus locally Lipschitz continuous on the open interval $(-1,1)$ while $\frac{|\mathbf{G}+\widetilde{\mathbf{k}}|}{\sqrt{2\mathrm{E_c}}} < 1$ for any $\widetilde{\mathbf{k}} \in \mathbb{B}_\delta(\mathbf{k}_0)$ and all $\mathbf{G} \in \mathcal{B}_{\mathbf{k}_0}^{\mathrm{E_c}}$, we deduce the existence of a constant $C_{\mathscr{G}}$

such that that

$$\widetilde{\varepsilon}_n^{\mathrm{E_c}}(\mathbf{k}_1) - \widetilde{\varepsilon}_n^{\mathrm{E_c}}(\mathbf{k}_2) \leqslant \sqrt{\frac{\mathrm{E_c}}{2}} C_{\mathscr{G}} \max_{\mathbf{G} \in \mathcal{B}_{\mathbf{k}_0}^{\mathrm{E_c}}} \Big| |\mathbf{G} + \mathbf{k}_1| - |\mathbf{G} + \mathbf{k}_2| \Big| \leqslant \sqrt{\frac{\mathrm{E_c}}{2}} C_{\mathscr{G}} |\mathbf{k}_1 - \mathbf{k}_2|,$$

where the last inequality is a consequence of the reverse triangle inequality. The Lipschitz continuity of the approximate energy bands now readily follows.

Case three: we assume that $\mathbf{k}_0 \in \mathbb{R}^d$ satisfies

$$\text{There exists } \mathbf{G} \in \mathbb{L}^* \text{ such that } |\mathbf{k}_0 + \mathbf{G}|^2 = 2\mathrm{E_c}.$$

In other words, we assume that there *is* a change in the cardinality of the basis set at $\mathbf{k}_0$.

We claim that in this case, there are two possibilities: if there is no band crossing at $(\mathbf{k}_0, \widetilde{\varepsilon}_n^{\mathrm{E_c}}(\mathbf{k}_0))$, then the approximate energy band $\widetilde{\varepsilon}_n^{\mathrm{E_c}}$ is of class $\mathscr{C}^m$ in a neighborhood of $\mathbf{k}_0$, i.e., $\widetilde{\varepsilon}_n^{\mathrm{E_c}}$ has the same local regularity at $\mathbf{k}_0$ as the rate of blow-up of the function $\mathscr{G}(x)$ in the limit $x \to 1$. If, on the other hand, there is a band crossing at $(\mathbf{k}_0, \widetilde{\varepsilon}_n(\mathbf{k}_0))$, then the approximate energy band $\widetilde{\varepsilon}_n^{\mathrm{E_c}}$ is Lipschitz continuous at $\mathbf{k}_0$ if $m \geqslant 1$ and of class $\mathscr{C}^0$ otherwise.

We begin by defining the non-empty sets

$$\begin{aligned} \mathrm{S}_{\mathbf{k}_0}^{\mathrm{E_c^-}} &:= \left\{ \mathbf{G} \in \mathbb{L}^* \colon \frac{1}{2}|\mathbf{k}_0 + \mathbf{G}|^2 < \mathrm{E_c} \right\} \quad \text{and} \\ \mathrm{S}_{\mathbf{k}_0}^{\mathrm{E_c}} &:= \left\{ \mathbf{G} \in \mathbb{L}^* \colon \frac{1}{2}|\mathbf{k}_0 + \mathbf{G}|^2 = \mathrm{E_c} \right\} \quad \text{with} \quad M := \dim \mathrm{S}_{\mathbf{k}_0}^{\mathrm{E_c}} < \infty. \end{aligned}$$

Clearly there exists $\overline{\delta} > 0$ sufficiently small such that $\forall\ \mathbf{k} \in \mathbb{R}^d$ with $|\mathbf{k} - \mathbf{k}_0| < \overline{\delta}$, the $\mathbf{k}$-dependent basis sets satisfy

$$\left\{ e_{\mathbf{G}} \colon \mathbf{G} \in \mathrm{S}_{\mathbf{k}_0}^{\mathrm{E_c^-}} \right\} \subseteq \ \mathcal{B}_{\mathbf{k}}^{\mathrm{E_c}} \ \subseteq \left\{ e_{\mathbf{G}} \colon \mathbf{G} \in \mathrm{S}_{\mathbf{k}_0}^{\mathrm{E_c^-}} \cup \mathrm{S}_{\mathbf{k}_0}^{\mathrm{E_c}} \right\}.$$

Let us therefore fix some $\delta \leqslant \overline{\delta}$ and consider the open ball $\mathbb{B}_\delta(\mathbf{k}_0)$ of radius δ centered at $\mathbf{k}_0$. We will study the behavior of sequences of $\mathbf{k}$-points in this open ball that converge to $\mathbf{k}_0$.

To do so, we consider a specific decomposition of the open ball $\mathbb{B}_\delta(\mathbf{k}_0)$ into sectors $\{\Omega_j\}_{j=1}^M$ defined as follows: First, for every $\widetilde{\mathbf{G}} \in \mathrm{S}_{\mathbf{k}_0}^{\mathrm{E_c}}$ we define the open set

$$\mathrm{S}_{\widetilde{\mathbf{G}}} = \left\{ \mathbf{k} \in \mathbb{B}_\delta(\mathbf{k}_0) \colon \ \frac{1}{2}|\mathbf{k} + \widetilde{\mathbf{G}}|^2 < \mathrm{E_c} \right\}. \tag{4.7.15}$$

It is now easy to see that there are exactly two cases:

1. For all $\mathbf{k} \in \mathrm{S}_{\widetilde{\mathbf{G}}}$, the Fourier mode $e_{\widetilde{\mathbf{G}}}$ is an element of the $\mathbf{k}$-dependent basis set $\mathcal{B}_{\mathbf{k}}^{\mathrm{E_c}}$.
2. For all $\mathbf{k} \in \mathbb{B}_\delta(\mathbf{k}_0) \setminus \mathrm{S}_{\widetilde{\mathbf{G}}}$, the Fourier mode $e_{\widetilde{\mathbf{G}}}$ is <u>not</u> an element of the $\mathbf{k}$-dependent basis set $\mathcal{B}_{\mathbf{k}}^{\mathrm{E_c}}$.

Next, we label the elements of $\mathrm{S}_{\mathbf{k}_0}^{\mathrm{E_c}}$ as $\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_M$. It follows that there exist sets $\Omega_j,\ j \in \{1, \ldots, 2^M\} \subset \mathbb{B}_\delta(\mathbf{k}_0)$ such that $\mathbb{B}_\delta(\mathbf{k}_0) = \cup_{j=1}^{2^M} \Omega_j$ with

$$\begin{aligned} &\Omega_1 := \mathbb{B}_\delta(\mathbf{k}_0) \setminus \Big(\bigcup_{\mathbf{G} \in \mathrm{S}_{\mathbf{k}_0}^{\mathrm{E_c}}} \mathrm{S}_{\mathbf{G}} \Big) \quad \text{and} \\ &\forall j \in \{2, \ldots, 2^M\},\ \exists\ L \leqslant M,\ J := \{j_1, j_2, \ldots, j_L\} \subset \{1, \ldots, M\} \text{ such that} \\ &\qquad \Omega_j = \left(\bigcap_{\ell \in J} \mathrm{S}_{\mathbf{G}_\ell} \right) \setminus \left(\bigcup_{\ell \in \{1,\ldots,M\} \setminus J} \mathrm{S}_{\mathbf{G}_\ell} \right). \end{aligned} \tag{4.7.16}$$

A visual example of the above decomposition is displayed in Figure 4.14. Note that we allow for the possibility of some $\Omega_j,\ j \in \{1, \ldots, 2^M\}$ to be empty. Two observations should now be made.

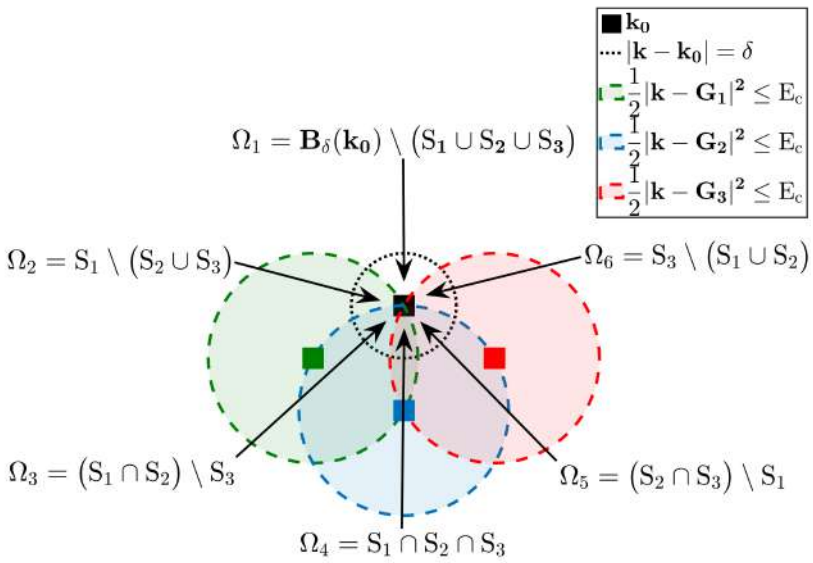


Figure 4.14 – An example of Decomposition (4.7.16) introduced above for a hexagonal lattice. For clarity, we show only the decomposition corresponding to three elements of $\mathrm{S}^{\mathrm{E_c}}_{\mathbf{k}_0}$.

Observation one: the set Ω_1 has the following property: For all $\mathbf{k} \in \Omega_1$ it holds that $\mathcal{B}^{\mathrm{E_c}}_{\mathbf{k}} = \mathcal{B}^{\mathrm{E_c}}_{\mathbf{k}_0}$.

Observation two: each set $\Omega_j,\ j > 1$ has the following property: For all $\mathbf{k} \in \Omega_j$ it holds that

$$\mathcal{B}^{\mathrm{E_c}}_{\mathbf{k}} = \mathcal{B}^{\mathrm{E_c}}_{\mathbf{k}_0} \cup \{e_{\mathbf{G}_{j_1}}, e_{\mathbf{G}_{j_2}}, \ldots, e_{\mathbf{G}_{j_L}}\}.$$

Let us now fix some $j \in \{1, \ldots, 2^M\}$ and consider the set Ω_j. Our goal is to study the convergence of the approximate, modified energy band $\widetilde{\varepsilon}^{\mathrm{E_c}}_n(\cdot)$ (and its derivatives) along sequences $\{\mathbf{k}_\ell\}_{\ell \in \mathbb{N}^*}$ contained in Ω_j that converge to $\mathbf{k}_0$. The reason we have introduced the decomposition $\{\Omega_j\}_{j=1}^{2^M}$ and we restrict ourselves, as a first step, to sequences contained in a fixed Ω_j is because for each fixed Ω_j and all $\mathbf{k} \in \Omega_j$, we have precise knowledge of the $\mathbf{k}$-dependent basis set thanks to **Observation two**.

Obviously, if Ω_j is an empty set, then there is nothing to study so we assume without loss of generality that Ω_j is non-empty. Additionally, thanks to **Observation one** above, the choice $\Omega_j = \Omega_1$ is already covered by the proof of **Case one** of our proof so we may assume that $j > 1$. It now follows from **Observation two** that for all $\mathbf{k} \in \Omega_j$, we have the decomposition

$$\mathcal{B}^{\mathrm{E_c}}_{\mathbf{k}} = \mathcal{B}^{\mathrm{E_c}}_{\mathbf{k}_0} \cup \widetilde{\mathcal{B}_{\Omega_j}}^{\mathrm{E_c}} \quad \text{and} \quad \mathcal{B}^{\mathrm{E_c}}_{\mathbf{k}_0} \cap \widetilde{\mathcal{B}_{\Omega_j}}^{\mathrm{E_c}} = \emptyset, \qquad \text{where} \tag{4.7.17}$$

$$\widetilde{\mathcal{B}_{\Omega_j}}^{\mathrm{E_c}} := \{e_{\mathbf{G}} :\ \mathbf{G} \in \{\mathbf{G}_{j_1}, \mathbf{G}_{j_2}, \ldots, \mathbf{G}_{j_L}\}\} \quad \text{and} \quad \widetilde{\mathrm{X}}^{\mathrm{E_c}}_{\mathbf{k}_0,\Omega_j} := \operatorname{span} \widetilde{\mathcal{B}_{\Omega_j}}^{\mathrm{E_c}}. \tag{4.7.18}$$

Let us remark here that in Equation (4.7.17), the fact that the set intersection is empty follows from the fact that if the Fourier mode $e_{\mathbf{G}_k} \in \widetilde{\mathcal{B}_{\Omega_j}}^{\mathrm{E_c}}$, then $\mathbf{G}_k \in \mathrm{S}^{\mathrm{E_c}}_{\mathbf{k}_0}$ so that $e_{\mathbf{G}_k} \notin \mathcal{B}^{\mathrm{E_c}}_{\mathbf{k}_0}$ by Definition 4.3.2.

From Equations (4.7.17) and (4.7.18) we can now deduce that $\forall \mathbf{k} \in \Omega_j$ it holds that

$$\mathrm{X}^{\mathrm{E_c}}_{\mathbf{k}} := \operatorname{span} \mathcal{B}^{\mathrm{E_c}}_{\mathbf{k}} = \operatorname{span} \mathcal{B}^{\mathrm{E_c}}_{\mathbf{k}_0} \oplus \operatorname{span} \widetilde{\mathcal{B}_{\Omega_j}}^{\mathrm{E_c}} = \mathrm{X}^{\mathrm{E_c}}_{\mathbf{k}_0} \oplus \widetilde{\mathrm{X}}^{\mathrm{E_c}}_{\mathbf{k}_0,\Omega_j}.$$

We denote by $\widetilde{\Pi}_{\Omega_j,\mathrm{E_c}} \colon L^2_{\mathrm{per}}(\Omega) \to L^2_{\mathrm{per}}(\Omega)$ the L^2_{per}-orthogonal projection operator onto $\widetilde{\mathrm{X}}^{\mathrm{E_c}}_{\mathbf{k}_0,\Omega_j}$. It follows that for all $\mathbf{k} \in \Omega_j$, the modified Hamiltonian matrix $\widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}}$ admits the following block representation:

$$\widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}} = \left[\begin{array}{c|c} \Pi_{\mathbf{k}_0,\mathrm{E_c}}\left(\widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}}\right)\Pi_{\mathbf{k}_0,\mathrm{E_c}} & \Pi_{\mathbf{k}_0,\mathrm{E_c}}\left(\widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}}\right)\widetilde{\Pi}_{\Omega_j,\mathrm{E_c}} \\ \hline \widetilde{\Pi}_{\Omega_j,\mathrm{E_c}}\left(\widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}}\right)\Pi_{\mathbf{k}_0,\mathrm{E_c}} & \widetilde{\Pi}_{\Omega_j,\mathrm{E_c}}\left(\widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}}\right)\widetilde{\Pi}_{\Omega_j,\mathrm{E_c}} \end{array}\right] = \left[\begin{array}{c|c} \mathcal{A}_{\mathbf{k}} & \mathcal{B}_{\mathbf{k}} \\ \hline \mathcal{B}^*_{\mathbf{k}} & \mathcal{C}_{\mathbf{k}} \end{array}\right]. \tag{4.7.19}$$

We will now consider the convergence properties of each of the sub-matrices appearing in Equation (4.7.19).

Convergence properties of sub-matrix $\mathcal{A}_{\mathbf{k}}$.

For the sub-matrix $\mathcal{A}_{\mathbf{k}}$, using Equation (4.4.2) we have that for all $\mathbf{k}\in\Omega_j$ and any $\Phi\in \mathrm{X}^{\mathrm{E_c}}_{\mathbf{k}_0}$ it holds that

$$\mathcal{A}_{\mathbf{k}}\Phi = \Pi_{\mathbf{k}_0,\mathrm{E_c}}\left(\widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}}\right)\Phi = \Pi_{\mathbf{k}_0,\mathrm{E_c}}\sum_{\substack{\mathbf{G}\in\mathbb{L}^*\\ |\mathbf{k}_0+\mathbf{G}|^2<2\mathrm{E_c}}}\widehat{\Phi}_{\mathbf{G}}\left(\mathrm{E_c}\mathscr{G}\left(\frac{|\mathbf{G}+\mathbf{k}|}{\sqrt{2\mathrm{E_c}}}\right)+V\right)e_{\mathbf{G}}. \tag{4.7.20}$$

Since the blow-up function $\mathscr{G}$ is continuous on the interval $(0,1)$, we see immediately that for any sequence $\{\mathbf{k}_\ell\}_{\ell\in\mathbb{N}^*}\subset\Omega_j$ such that $\lim_{\ell\to\infty}\mathbf{k}_\ell=\mathbf{k}_0$ it holds that

$$\lim_{\ell\to\infty}\left\|\mathcal{A}_{\mathbf{k}_\ell}\Phi-\widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}_0}\Phi\right\|_{L^2_{\mathrm{per}}(\Omega)}=0 \qquad \text{and thus} \qquad \lim_{\ell\to\infty}\left\|\mathcal{A}_{\mathbf{k}_\ell}-\widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}_0}\right\|_2=0 \tag{4.7.21}$$

Additionally, if the blow-up function $\mathscr{G}$ is of class $\mathscr{C}^m$ on $(0,1)$ for $m>0$, then Equation (4.7.20) also allows us to deduce that the sub-matrix $\mathcal{A}_{\mathbf{k}},\ \mathbf{k}\in\mathbb{R}^d$ is continuously differentiable up to order m at $\mathbf{k}=\mathbf{k}_0$.

Convergence properties of sub-matrices $\mathcal{B}_{\mathbf{k}}$ and $\mathcal{B}^*_{\mathbf{k}}$.

Notice that the blow-up function $\mathscr{G}$ does not appear in the off-diagonal blocks $\mathcal{B}$ and $\mathcal{B}^*$ and that the effective potential V is independent of $\mathbf{k}$ and $\mathbf{k}_0$. Consequently, a similar argument as the one used for the sub-matrix $\mathcal{A}_{\mathbf{k}}$ yields that for any sequence $\{\mathbf{k}_\ell\}_{\ell\in\mathbb{N}^*}\subset\Omega_j$ such that $\lim_{\ell\to\infty}\mathbf{k}_\ell=\mathbf{k}_0$ we have

$$\lim_{\ell\to\infty}\|\mathcal{B}_{\mathbf{k}_\ell}-\mathcal{B}_{\mathbf{k}_0}\|_2=0 \qquad \text{and} \qquad \lim_{\ell\to\infty}\left\|\mathcal{B}^*_{\mathbf{k}_\ell}-\mathcal{B}^*_{\mathbf{k}_0}\right\|_2=0. \tag{4.7.22}$$

where $\mathcal{B}_{\mathbf{k}_0},\mathcal{B}^*_{\mathbf{k}_0}$ are fixed, rectangular matrices that are independent of the specific choice of sequence $\{\mathbf{k}_\ell\}$, although they depend of course on the chosen sector Ω_j.

Convergence properties of sub-matrix $\mathcal{C}_{\mathbf{k}}$.

We use once again Equation (4.4.2) to deduce that for all $\mathbf{k}\in\Omega_j$ and any $\Phi\in\widetilde{\mathrm{X}}^{\mathrm{E_c}}_{\mathbf{k}_0,\Omega_j}$ it holds that

$$\mathcal{C}_{\mathbf{k}}\Phi = \widetilde{\Pi}_{\Omega_j,\mathrm{E_c}}\left(\widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}}\right)\Phi = \widetilde{\Pi}_{\Omega_j,\mathrm{E_c}}\sum_{\mathbf{G}\in\{\mathbf{G}_{j_1},\ldots,\mathbf{G}_{j_\ell}\}\subset\mathbb{L}^*}\widehat{\Phi}_{\mathbf{G}}\left(\mathrm{E_c}\mathscr{G}\left(\frac{|\mathbf{G}+\mathbf{k}|}{\sqrt{2\mathrm{E_c}}}\right)+V\right)e_{\mathbf{G}}. \tag{4.7.23}$$

Consider now a sequence $\{\mathbf{k}_\ell\}_{j\in\mathbb{N}^*}\subset\Omega_j$ such that $\lim_{\ell\to\infty}\mathbf{k}_\ell=\mathbf{k}_0$. Recalling that $\{\mathbf{G}_{j_1},\ldots,\mathbf{G}_{j_\ell}\}\subset\mathrm{S}^{\mathrm{E_c}}_{\mathbf{k}_0}$ and the corresponding Fourier modes $e_{\mathbf{G}_{j_1}},\ldots,e_{\mathbf{G}_{j_\ell}}$ are elements of $\widetilde{\mathcal{B}_{\Omega_j}}^{\mathrm{E_c}}$, and using Equation (4.7.15) we see that

$$\begin{aligned} |\mathbf{G}_{j_k}+\mathbf{k}_\ell| &\leqslant \sqrt{2\mathrm{E_c}} \quad \text{for all } \ell\in\mathbb{N}^*,\ j_1,\ldots,j_\ell \quad \text{and} \\ \lim_{\ell\to\infty}|\mathbf{G}_{j_k}+\mathbf{k}_\ell| &= \sqrt{2\mathrm{E_c}} \quad \text{for all } \ j_1,\ldots,j_\ell. \end{aligned}$$

Since, on the one hand the blow-up function $\mathscr{G}(x)$ has a singularity at $x=1$, and on the other hand the effective potential V is independent of $\{\mathbf{k}_\ell\}_{\ell\in\mathbb{N}^*}\subset\Omega_j$ we infer that for all $\ell\in\mathbb{N}^*$, we can write the matrix $\mathcal{C}_{\mathbf{k}_\ell}$ in the form

$$\mathcal{C}_{\mathbf{k}_\ell}=\mathcal{D}_{\mathbf{k}_\ell}+\mathcal{N}_{\mathbf{k}_0}, \tag{4.7.24}$$

where $\mathscr{D}_{\mathbf{k}_\ell}$ and $\mathscr{N}_{\mathbf{k}_0}$ are both square matrices of dimension $\dim \widetilde{\mathrm{X}}^{\mathrm{E_c}}_{\mathbf{k}_0,\Omega_j}$, and the matrix $\mathscr{D}_{\mathbf{k}_\ell}$ is diagonal with entries that all diverge to $+\infty$ in the limit $\ell \to \infty$ while the entries of $\mathscr{N}_{\mathbf{k}_0}$ are independent of ℓ. A particular consequence of this is that the matrix $\mathscr{D}_{\mathbf{k}_\ell}$ is invertible for ℓ sufficiently large.

We now claim that for all $p \in \{0, \dots, m\}$ it holds that

$$\lim_{\ell \to \infty} \frac{\left\| \mathscr{D}^{-1}_{\mathbf{k}_\ell} \right\|_2}{|\mathbf{k}_0 - \mathbf{k}_\ell|^p} = 0.$$

To see this, we recall Equation (4.7.23) and the fact that $\mathscr{D}_{\mathbf{k}_\ell}$ is diagonal so that it suffices to show that $\forall \mathbf{G}_{j_k} \in \{\mathbf{G}_{j_1}, \dots, \mathbf{G}_{j_\ell}\}$ and any $p \in \{0, \dots, m\}$ it holds that

$$\lim_{\ell \to \infty} \frac{1}{|\mathbf{k}_0 - \mathbf{k}_\ell|^p} \cdot \frac{1}{\mathscr{G}\left(\frac{|\mathbf{G}_{j_k} + \mathbf{k}_\ell|}{\sqrt{2\mathrm{E_c}}}\right)} = 0 \quad \text{or equivalently}$$

$$\lim_{\ell \to \infty} |\mathbf{k}_0 - \mathbf{k}_\ell|^p \cdot \mathscr{G}\left(\frac{|\mathbf{G}_{j_k} + \mathbf{k}_\ell|}{\sqrt{2\mathrm{E_c}}}\right) = +\infty.$$

Using simple algebra, one can show that for all $\mathbf{G}_{j_k} \in \{\mathbf{G}_{j_1}, \dots, \mathbf{G}_{j_\ell}\}$ and any $p \in \{0, \dots, m\}$ we have

$$\lim_{\ell \to \infty} |\mathbf{k}_0 - \mathbf{k}_\ell|^p \cdot \mathscr{G}\left(\frac{|\mathbf{G}_{j_k} + \mathbf{k}_\ell|}{\sqrt{2\mathrm{E_c}}}\right) = +\infty \quad \iff \quad \lim_{x \to 1^-} (1-x)^j \cdot \mathscr{G}(x) = +\infty.$$

But this latter condition is satisfied by the blow-up function $\mathscr{G}$ by assumption (see Definition 4.4.1). We therefore conclude that for all $p \in \{0, \dots, m\}$ it holds that

$$\lim_{\ell \to \infty} \frac{\mathscr{D}^{-1}_{\mathbf{k}_\ell}}{|\mathbf{k}_0 - \mathbf{k}_\ell|^p} = 0 = \lim_{\ell \to \infty} \frac{\mathcal{C}^{-1}_{\mathbf{k}_\ell}}{|\mathbf{k}_0 - \mathbf{k}_\ell|^p} \qquad \text{in the matrix 2-norm topology.} \tag{4.7.25}$$

Consider again a sequence $\{\mathbf{k}_\ell\}_{\ell \in \mathbb{N}^*} \subset \Omega_j$ such that $\lim_{\ell \to \infty} \mathbf{k}_\ell = \mathbf{k}_0$. Having understood the convergence properties of the sub-blocks of the modified Hamiltonian matrix $\widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}}$, we will now study the convergence of the approximate energy band $\widetilde{\varepsilon}^{\mathrm{E_c}}_n$ and its derivatives up to order m as functions of the sequence $\{\mathbf{k}_\ell\}_{\ell \in \mathbb{N}^*}$.

Continuity of energy bands.

Recall that $M_{\mathrm{E_c}}(\mathbf{k})$ denotes the dimension of the matrix $\widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}}$ at $\mathbf{k} \in \mathbb{R}^d$. Thanks to the definition of the set Ω_j, we see that for each element of the sequence $\{\mathbf{k}_\ell\}_{\ell \in \mathbb{N}^*}$, the dimension of $\widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}_\ell}$ remains constant, i.e.,$M_{\mathrm{E_c}}(\mathbf{k}_\ell) = M \in \mathbb{N}^*$. Consequently, we may apply Lemma 1 to the modified Hamiltonian matrices $\left\{\widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}_\ell}\right\}_{\ell \in \mathbb{N}^*}$. Indeed, thanks to the convergence properties of the sub-matrices $\{\mathcal{A}_{\mathbf{k}_\ell}\}_{\ell \in \mathbb{N}^*}, \{\mathcal{B}_{\mathbf{k}_\ell}\}_{\ell \in \mathbb{N}^*}, \{\mathcal{B}^*_{\mathbf{k}_\ell}\}_{\ell \in \mathbb{N}^*}$ and $\{\mathcal{C}_{\mathbf{k}_\ell}\}_{\ell \in \mathbb{N}^*}$ established above (and taking a subsequence, if necessary, to ensure the invertibility of all $\mathcal{C}_{\mathbf{k}_\ell}$), we see that the assumptions of Lemma 1 are satisfied. Denoting therefore, $p = \dim \mathcal{A}_{\mathbf{k}_0}$ and recalling that $\widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}_0} = \mathcal{A}_{\mathbf{k}_0}$, we deduce that for each $q \in \{1, \dots, p\}$ it holds that

$$\begin{aligned} &\lim_{\ell \to \infty} \widetilde{\varepsilon}^{\mathrm{E_c}}_q(\mathbf{k}_\ell) = \widetilde{\varepsilon}^{\mathrm{E_c}}_q(\mathbf{k}_0) \quad \text{and} \\ &\lim_{\ell \to \infty} \widetilde{\varepsilon}^{\mathrm{E_c}}_{p+1}(\mathbf{k}_\ell) = \lim_{\ell \to \infty} \widetilde{\varepsilon}^{\mathrm{E_c}}_{p+2}(\mathbf{k}_\ell) = \cdots = \lim_{\ell \to \infty} \widetilde{\varepsilon}^{\mathrm{E_c}}_{M}(\mathbf{k}_\ell) = \infty. \end{aligned} \tag{4.7.26}$$

In order to conclude the continuity of the bounded energy bands $\{\widetilde{\varepsilon}^{\mathrm{E_c}}_q\}_{q \in \{1,\dots,p\}}$, it suffices to recall that we have considered a sequence $\{\mathbf{k}_\ell\}_{\ell \in \mathbb{N}^*} \subset \Omega_j$ for some $j \in \{2, \dots, 2^M\}$. But since Ω_j was chosen arbitrarily and there are only a finite number of possible choices for Ω_j, we conclude that Equation (4.7.26) holds for <u>any</u> sequence $\{\mathbf{k}_\ell\}_{\ell \in \mathbb{N}^*} \subset \mathbb{B}_\delta(\mathbf{k})$. It follows that all bounded, modified energy bands $\{\widetilde{\varepsilon}^{\mathrm{E_c}}_q\}_{q=1}^p$ are continuous at $\mathbf{k} = \mathbf{k}_0$ as claimed. Noting that $p = \dim \mathcal{A}_{\mathbf{k}_0} \leqslant M^-_{\mathrm{E_c}}$ completes the proof of continuity.

If the blow-up function $\mathscr{G}$ satisfies Properties (1)-(4) from Definition 4.4.1 only for $m = 0$, then we are done. Hence, we may assume that $m \geqslant 1$ and that all eigenvalues of $\mathcal{A}_{\mathbf{k}_0} = \widetilde{\mathrm{H}}^{\mathscr{G},\mathrm{E_c}}_{\mathbf{k}_0}$ are simple.

We study next the regularity of the derivatives of the bounded, modified energy bands.

First order differentiability of energy bands.

We begin by studying the convergence of the eigenvectors associated with the bounded energy bands $\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)$, $q \in \{1, \dots, p\}$. The primary tool we will use for this study will be the Schur complement associated with the block decomposition (4.7.19) of the modified Hamiltonian matrix $\widetilde{\mathrm{H}}_{\mathbf{k}_\ell}^{\mathscr{G},\mathrm{E_c}}$.

Let $q \in \{1, \dots, p\}$ be the index of a bounded energy band. A straightforward calculation using the block decomposition (4.7.19) shows that for any $\mathbf{k}_\ell \in \Omega_j$ it holds that

$$\Pi_{\mathbf{k}_0,\mathrm{E_c}} \left(\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\, \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} \right) = \mathscr{A}_{\mathbf{k}_\ell} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} - \mathscr{B}_{\mathbf{k}_\ell} \left(\mathcal{C}_{\mathbf{k}_\ell} - \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell) \right)^{-1} \mathscr{B}_{\mathbf{k}_\ell}^* \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}}, \tag{4.7.27}$$

where $\widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}}$ denotes the q^{th} normalized eigenfunction of the modified Hamiltonian matrix $\widetilde{\mathrm{H}}_{\mathbf{k}_\ell}^{\mathscr{G},\mathrm{E_c}}$. Additionally, thanks to Equations (4.7.22) and (4.7.25), we deduce from Equation (4.7.27) that

$$\lim_{\ell \to \infty} \left(\Pi_{\mathbf{k}_0,\mathrm{E_c}} \left(\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell) \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} \right) - \mathscr{A}_{\mathbf{k}_\ell} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} \right) = 0. \tag{4.7.28}$$

Next, observe that since the sequence $\left\{ \Pi_{\mathbf{k}_0,\mathrm{E_c}}\, \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} \right\}_{\ell \in \mathbb{N}^*}$ is bounded, it possesses a convergent subsequence, which we also write as $\left\{ \Pi_{\mathbf{k}_0,\mathrm{E_c}}\, \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} \right\}_{\ell \in \mathbb{N}^*}$. We can then deduce from Equations (4.7.21), (4.7.26), and (4.7.28) that

$$\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_0) \left(\lim_{\ell \to \infty} \Pi_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} \right) = \mathscr{A}_{\mathbf{k}_0} \left(\lim_{\ell \to \infty} \Pi_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} \right).$$

But this implies that either $\lim_{\ell \to \infty} \Pi_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}}$ is (up to normalization) the eigenvector $\widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}}$ associated with the eigenvalue $\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_0)$ or $\lim_{\ell \to \infty} \Pi_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} = 0$. Suppose on the contrary that $\lim_{\ell \to \infty} \Pi_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} = 0$ and note that the block decomposition (4.7.19) implies that for all $\mathbf{k}_\ell \in \Omega_j$ we have

$$\Pi_{\mathbf{k}_0,\mathrm{E_c}}^{\perp} \left(\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\, \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} \right) = \mathscr{B}_{\mathbf{k}_\ell}^* \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} + \mathcal{C}_{\mathbf{k}_\ell} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}}. \tag{4.7.29}$$

We can now take the limit $\ell \to \infty$ on both sides of Equation (4.7.29). But $\lim_{\ell \to \infty} \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell) = \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_0) < \infty$ and $\lim_{\ell \to \infty} \| \Pi_{\mathbf{k}_0,\mathrm{E_c}}^{\perp}\, \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} \| = 1$ while all eigenvalues of the matrix $\mathcal{C}_{\mathbf{k}_\ell}$ diverge to $+\infty$ in the limit $\ell \to \infty$. Consequently, we must have

$$\lim_{\ell \to \infty} \Pi_{\mathbf{k}_0,\mathrm{E_c}}\, \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} = \lim_{\ell \to \infty} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} = \widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}}. \tag{4.7.30}$$

Moreover, similar to the argument for sequential continuity of the approximate, modified energy bands, we conclude from the fact that $\{\mathbf{k}_\ell\}_{\ell \in \mathbb{N}^*} \subset \Omega_j$ and only a finite number of possibilities exist for the choice of $j \in \{2, \dots, 2^M\}$, that Equation (4.7.30) also holds for <u>any</u> sequence $\{\mathbf{k}_\ell\}_{\ell \in \mathbb{N}^*} \subset \mathbb{B}_\delta(\mathbf{k})$, which proves sequential continuity of the normalized eigenfunctions associated with all bounded energy bands.

Equipped with the convergence properties of the eigenvectors associated with the bounded energy bands, we can now consider the first order derivatives of the bounded, modified energy band $\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k})$, $q \in \{1, \dots, p\}$ at $\mathbf{k} = \mathbf{k}_0$. Thanks once again to Equations (4.7.22) and (4.7.25), we deduce from the Schur-type decomposition (4.7.27) that

$$\lim_{\ell \to \infty} \frac{\left(\Pi_{\mathbf{k}_0,\mathrm{E_c}} \left(\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\, \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} \right) - \mathscr{A}_{\mathbf{k}_\ell}\, \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} \right)}{|\mathbf{k}_\ell - \mathbf{k}_0|} = 0.$$

Adding and subtracting the term $\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_0)\, \Pi_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}}$, using the fact that $\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_0) \widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}} = \mathscr{A}_{\mathbf{k}_0} \widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}}$, and taking the inner product with the eigenfunction $\widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}}$ then yields

$$\lim_{\ell \to \infty} \frac{\left(\left(\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell) - \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_0) + \mathscr{A}_{\mathbf{k}_0} - \mathscr{A}_{\mathbf{k}_\ell} \right) \widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}}, \Pi_{\mathbf{k}_0,\mathrm{E_c}}\, \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} \right)_{L^2_{\mathrm{per}}(\Omega)}}{|\mathbf{k}_\ell - \mathbf{k}_0|} = 0. \tag{4.7.31}$$

Next, recall that $\mathscr{A}_{\mathbf{k}}$ is m-times continuously differentiable at $\mathbf{k}=\mathbf{k}_0$ and denote by $\mathrm{d}\mathscr{A}_{\mathbf{k}_0}:\mathbb{R}^d\to\mathbb{R}^{p\times p}$ the total derivative of $\mathscr{A}_{\mathbf{k}}$ at $\mathbf{k}=\mathbf{k}_0$. Adding and subtracting the term $\mathrm{d}\mathscr{A}_{\mathbf{k}_0}[\mathbf{k}_\ell-\mathbf{k}_0]$, i.e., $\mathrm{d}\mathscr{A}_{\mathbf{k}_0}$ acting on the vector $\mathbf{k}_\ell-\mathbf{k}_0$, we can deduce from Equation (4.7.31) that

$$\lim_{\ell\to\infty}\frac{\left(\left(\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)-\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_0)-\mathrm{d}\mathscr{A}_{\mathbf{k}_0}[\mathbf{k}_\ell-\mathbf{k}_0]\right)\widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}},\Pi_{\mathbf{k}_0,\mathrm{E_c}}\,\widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}}\right)_{L^2_{\mathrm{per}}(\Omega)}}{|\mathbf{k}_\ell-\mathbf{k}_0|}=0. \tag{4.7.32}$$

Adding and subtracting the term $\left(\mathrm{d}\mathscr{A}_{\mathbf{k}_0}[\mathbf{k}_\ell-\mathbf{k}_0]\widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}},\widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}}\right)_{L^2_{\mathrm{per}}(\Omega)}$ and using simple algebra, it can be deduced that

$$\lim_{\ell\to\infty}\frac{\left(\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)-\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_0)\right)-\left(\mathrm{d}\mathscr{A}_{\mathbf{k}_0}[\mathbf{k}_\ell-\mathbf{k}_0]\widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}},\widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}}\right)_{L^2_{\mathrm{per}}(\Omega)}}{|\mathbf{k}_\ell-\mathbf{k}_0|}=0. \tag{4.7.33}$$

Similar to the argument for sequential continuity of the approximate, modified energy bands, we conclude from the fact that $\{\mathbf{k}_\ell\}_{\ell\in\mathbb{N}^*}\subset\Omega_j$ and only a finite number of possibilities exist for the choice of $j\in\{2,\dots,2^M\}$, that Equation (4.7.33) also holds for any sequence $\{\mathbf{k}_\ell\}_{\ell\in\mathbb{N}^*}\subset\mathbb{B}_\delta(\mathbf{k})$. Thus, the total derivative of the approximate, modified energy bands exists at $\mathbf{k}_0$ and is given by

$$\mathrm{d}\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_0)=\left(\mathrm{d}\mathscr{A}_{\mathbf{k}_0}\widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}},\widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}}\right)_{L^2_{\mathrm{per}}(\Omega)}. \tag{4.7.34}$$

As a consequence, $\widetilde{\varepsilon}_q^{\mathrm{E_c}}$ is of class $\mathscr{C}^1$ at $\mathbf{k}_0$ as claimed. Noting that we picked an arbitrary $q\in\{1\dots,p\}$ where $p=\dim\mathscr{A}_{\mathbf{k}_0}\leqslant M_{\mathrm{E_c}}^-$ completes the proof of differentiability of order one.

If the blow-up function $\mathscr{G}$ satisfies Properties (1)-(4) from Definition 4.4.1 only for $m\leqslant 1$, then we are done. Hence, we may assume that $m\geqslant 2$.

Higher order differentiability of energy bands.

Imitating the procedure carried out for the case $m=1$, we will first make use of order one differentiability of the energy band $\widetilde{\varepsilon}_{q,\mathbf{k}_0}^{\mathrm{E_c}}$, $q\in\{1,\dots,p\}$, to establish order one differentiability of the corresponding eigenfunction $\widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}}$.

Let $q\in\{1,\dots,p\}$ be the index of a bounded energy band. As a first remark, let us recall that by construction, the approximation space $\mathrm{X}_{\mathbf{k}}^{\mathrm{E_c}}$ is identical for all $\mathbf{k}\in\Omega_j$. Since the sequence $\{\mathbf{k}_\ell\}_{\ell\in\mathbb{N}^*}\subset\Omega_j$, it can readily be deduced from the definition of the modified Hamiltonian matrix $\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}$ given by Equation (4.4.2) and the regularity properties of the blow-up function $\mathscr{G}$ that the modified Hamiltonian matrix $\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}$ is m-times continuously differentiable at any $\mathbf{k}=\mathbf{k}_\ell$. Moreover, we have assumed that all eigenvalues of $\mathscr{A}_{\mathbf{k}_0}=\widetilde{\mathrm{H}}_{\mathbf{k}_0}^{\mathscr{G},\mathrm{E_c}}$ are simple. Therefore, as discussed in **Case one** of the current proof, the implicit function theorem can be used to prove that for ℓ sufficiently large, the energy band $\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k})$ and the associated (normalized) eigenfunction $\widetilde{u}_{q,\mathbf{k}}^{\mathrm{E_c}}$ are m-times continuously differentiable (as a function of $\mathbf{k}\in\mathbb{R}^d$) at any $\mathbf{k}=\mathbf{k}_\ell$. Without loss of generality, we may assume that this is the case for all $\ell\in\mathbb{N}^*$.

Next, let us recall the Schur-type decomposition (4.7.27) which offers an expression for the eigenvalue $\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k})$ in terms of the block decomposition and Schur complement of the modified Hamiltonian matrix $\widetilde{\mathrm{H}}_{\mathbf{k}}^{\mathscr{G},\mathrm{E_c}}$. Taking partial derivatives with respect to the i^{th} component of $\mathbf{k}=(\mathbf{k}_1,\dots,\mathbf{k}_d)\in\mathbb{R}^d$ of this equation yields that for any $\ell\in\mathbb{N}^*$ it holds that

$$\begin{aligned}&\left(\mathscr{A}_{\mathbf{k}_\ell}-\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)-\mathscr{B}_{\mathbf{k}_\ell}\left(\mathcal{C}_{\mathbf{k}_\ell}-\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\right)^{-1}\mathscr{B}_{\mathbf{k}_\ell}^*\right)\,\partial_i\Pi_{\mathbf{k}_0,\mathrm{E_c}}\widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}}\\ =-\Bigg(&\partial_i\mathscr{A}_{\mathbf{k}_\ell}-\partial_i\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\\ &+\mathscr{B}_{\mathbf{k}_\ell}\left(\mathcal{C}_{\mathbf{k}_\ell}-\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\right)^{-1}\left(\partial_i\mathcal{C}_{\mathbf{k}_\ell}-\partial_i\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\right)\left(\mathcal{C}_{\mathbf{k}_\ell}-\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\right)^{-1}\mathscr{B}_{\mathbf{k}_\ell}^*\Bigg)\,\Pi_{\mathbf{k}_0,\mathrm{E_c}}\widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}},\end{aligned} \tag{4.7.35}$$

where we have used the fact that the sub-matrices $\mathscr{B}_{\mathbf{k}}$ and $\mathscr{B}_{\mathbf{k}}^*$ do not change for different choices of $\mathbf{k}\in\Omega_j$ while the sub-matrices $\mathscr{A}_{\mathbf{k}}$ and $\mathcal{C}_{\mathbf{k}}$ are m-times continuously differentiable by construction for all $\mathbf{k}=\mathbf{k}_\ell\in\Omega_j$.

Let now $\left(\lambda_{\mathscr{A}_{\mathbf{k}_0}}, v_{\mathbf{k}_0}\right) = \left(\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_0), \widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}}\right)$ denote the unique, normalized eigenpair of the matrix $\mathscr{A}_{\mathbf{k}_0} \in \mathbb{R}^{p\times p}$ such that $\lim_{\ell\to\infty} \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell) = \lambda_{\mathscr{A}_{\mathbf{k}_0}}$ and $\lim_{\ell\to\infty} \Pi_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} = v_{\mathbf{k}_0}$. Thanks to the regularity properties of the sub-matrix $\mathscr{A}_{\mathbf{k}_0}$, we can once again deduce that both $\lambda_{\mathscr{A}_{\mathbf{k}}}$, and $v_{\mathbf{k}}$ are m-times continuously differentiable at $\mathbf{k} = \mathbf{k}_0$. Our goal now is to use Equation (4.7.35) to demonstrate that $\lim_{\ell\to\infty} \partial_i \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} = \partial_i v_{\mathbf{k}_0}$.

To this end, we claim that in fact

$$\lim_{\ell\to\infty} \mathscr{B}_{\mathbf{k}_\ell} \left(\mathcal{C}_{\mathbf{k}_\ell} - \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\right)^{-1} \left(\partial_i \mathcal{C}_{\mathbf{k}_\ell} - \partial_i \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\right) \left(\mathcal{C}_{\mathbf{k}_\ell} - \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\right)^{-1} \mathscr{B}_{\mathbf{k}_\ell}^* = 0. \tag{4.7.36}$$

Indeed, Equation (4.7.36) is a consequence of the properties of the blow-up function $\mathscr{G}\colon \mathbb{R} \to \mathbb{R}$ given by Definition 4.4.1 as can easily be verified using a similar calculation as the one used to arrive at Equation (4.7.25).

Taking limits on both sides of Equation (4.7.35) and using the convergence properties we have proven thus far, we obtain that

$$\begin{aligned}\lim_{\ell\to\infty} \left(\mathscr{A}_{\mathbf{k}_\ell} - \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell) - \mathscr{B}_{\mathbf{k}_\ell} \left(\mathcal{C}_{\mathbf{k}_\ell} - \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\right)^{-1} \mathscr{B}_{\mathbf{k}_\ell}^*\right) \partial_i \Pi_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} \\ = -\left(\partial_i \mathscr{A}_{\mathbf{k}_0} - \partial_i \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_0)\right) \widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}}.\end{aligned} \tag{4.7.37}$$

A direct calculation now yields that the right hand side of the above equation is L^2_{per}-orthogonal to $\mathrm{span}\{\widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}}\}$. Moreover, thanks to the convergence properties of the sub-matrices $\mathscr{A}_{\mathbf{k}_\ell}, \mathscr{B}_{\mathbf{k}_\ell}, \mathscr{B}_{\mathbf{k}_\ell}^*$ and $\mathcal{C}_{\mathbf{k}_\ell}$, we also deduce that for ℓ sufficiently large it holds that

$$\left(\mathscr{A}_{\mathbf{k}_\ell} - \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell) - \mathscr{B}_{\mathbf{k}_\ell} \left(\mathcal{C}_{\mathbf{k}_\ell} - \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\right)^{-1} \mathscr{B}_{\mathbf{k}_\ell}^*\right) \text{ is invertible on } \left\{\widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}}\right\}^{\perp} \subset \mathrm{X}_{\mathbf{k}_0}^{\mathrm{E_c}}.$$

Consequently, Equation (4.7.37) yields that

$$\begin{aligned}&\Pi^{\perp}_{\widetilde{u}_q^{\mathrm{E_c}}(\mathbf{k}_0)} \left(\lim_{\ell\to\infty} \partial_i \Pi_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}}\right) \\ &\quad = -\left(\Pi^{\perp}_{\widetilde{u}_q^{\mathrm{E_c}}(\mathbf{k}_0)} \left(\mathscr{A}_{\mathbf{k}_0} - \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_0)\right) \Pi^{\perp}_{\widetilde{u}_q^{\mathrm{E_c}}(\mathbf{k}_0)}\right)^{-1} \left(\partial_i \mathscr{A}_{\mathbf{k}_0} - \partial_i \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_0)\right) \widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}},\end{aligned} \tag{4.7.38}$$

where $\Pi^{\perp}_{\widetilde{u}_q^{\mathrm{E_c}}(\mathbf{k}_0)}$ is the L^2_{per}-orthogonal projection operator onto $\{\widetilde{u}_{q,\mathbf{k}_0}^{\mathrm{E_c}}\}^{\perp} \subset \mathrm{X}_{\mathbf{k}_0}^{\mathrm{E_c}}$.

To proceed to the conclusion, we need to demonstrate that

$$\left(\mathbb{I} - \Pi^{\perp}_{\widetilde{u}_q^{\mathrm{E_c}}(\mathbf{k}_0)}\right) \left(\lim_{\ell\to\infty} \partial_i \Pi_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}}\right) = 0 \qquad \text{and} \qquad \lim_{\ell\to\infty} \partial_i \Pi^{\perp}_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} = 0.$$

We first focus on the latter limit. To this end, we recall Equation (4.7.29), which yields that for all $\ell \in \mathbb{N}^*$ it holds that

$$\Pi^{\perp}_{\mathbf{k}_0,\mathrm{E_c}} \left(\widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\, \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}}\right) = \mathscr{B}_{\mathbf{k}_\ell}^* \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} + \mathcal{C}_{\mathbf{k}_\ell} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}}. \tag{4.7.39}$$

We have already demonstrated that $\lim_{\ell\to\infty} \Pi^{\perp}_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} = 0$. Recalling therefore the decomposition $\mathcal{C}_{\mathbf{k}_\ell} = \mathcal{D}_{\mathbf{k}_\ell} + \mathcal{N}_{\mathbf{k}_0}$ introduced priori to Equation (4.7.25), we see that we must have

$$\lim_{\ell\to\infty} \|\mathcal{D}_{\mathbf{k}_\ell} \Pi^{\perp}_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}}\|_{L^2_{\mathrm{per}}(\Omega)} = 0. \tag{4.7.40}$$

Taking now partial derivatives with respect to the i^{th} component of $\mathbf{k} = (\mathbf{k}_1, \ldots, \mathbf{k}_d) \in \mathbb{R}^d$ of Equation (4.7.39), taking limits, and simplifying terms that obviously goes to zero now, we obtain that

$$\lim_{\ell\to\infty} \partial_i \Pi^{\perp}_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} = -\lim_{\ell\to\infty} \left(\mathcal{C}_{\mathbf{k}_\ell} - \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\right)^{-1} \left(\partial_i \mathcal{D}_{\mathbf{k}_\ell} - \partial_i \widetilde{\varepsilon}_q^{\mathrm{E_c}}(\mathbf{k}_\ell)\right) \Pi^{\perp}_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} = 0,$$

where the last equality follows again from the properties of the blow-up function $\mathscr{G}\colon \mathbb{R} \to \mathbb{R}$ defined through Definition 4.4.1. We conclude that $\lim_{\ell\to\infty} \partial_i \Pi^{\perp}_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}_{q,\mathbf{k}_\ell}^{\mathrm{E_c}} = 0$ as claimed.

It remains to prove that $\left(\mathbb{I} - \Pi^{\perp}_{\widetilde{u}^{\mathrm{E_c}}_{q}(\mathbf{k}_0)}\right)\left(\lim_{\ell\to\infty} \partial_i \Pi_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}^{\mathrm{E_c}}_{q,\mathbf{k}_\ell}\right) = 0$. To this end, we note that due to the normalization of $\widetilde{u}^{\mathrm{E_c}}_{q,\mathbf{k}_\ell}$, for all $\ell \in \mathbb{N}^*$ it holds that

$$\left(\Pi_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}^{\mathrm{E_c}}_{q,\mathbf{k}_\ell}, \partial_i \Pi_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}^{\mathrm{E_c}}_{q,\mathbf{k}_\ell}\right)_{L^2_{\mathrm{per}}(\Omega)} = \frac{1}{2}\partial_i \left\| \Pi^{\perp}_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}^{\mathrm{E_c}}_{q,\mathbf{k}_\ell} \right\|^2_{L^2_{\mathrm{per}}(\Omega)}$$

$$\implies \left(\mathbb{I} - \Pi^{\perp}_{\widetilde{u}^{\mathrm{E_c}}_{q}(\mathbf{k}_0)}\right)\left(\lim_{\ell\to\infty} \partial_i \Pi_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}^{\mathrm{E_c}}_{q,\mathbf{k}_\ell}\right) = \frac{1}{2}\lim_{\ell\to\infty} \partial_i \left\| \Pi^{\perp}_{\mathbf{k}_0,\mathrm{E_c}} \widetilde{u}^{\mathrm{E_c}}_{q,\mathbf{k}_\ell} \right\|^2_{L^2_{\mathrm{per}}(\Omega)} = 0.$$

Collecting these convergence results and recalling Equation (4.7.38), we see that we have in fact shown that

$$\lim_{\ell\to\infty} \partial_i \widetilde{u}^{\mathrm{E_c}}_{q,\mathbf{k}_\ell} = -\left(\Pi^{\perp}_{\widetilde{u}^{\mathrm{E_c}}_{q}(\mathbf{k}_0)} \left(\mathcal{A}_{\mathbf{k}_0} - \widetilde{\varepsilon}^{\mathrm{E_c}}_{q}(\mathbf{k}_0)\right) \Pi^{\perp}_{\widetilde{u}^{\mathrm{E_c}}_{q}(\mathbf{k}_0)}\right)^{-1} \left(\partial_i \mathcal{A}_{\mathbf{k}_0} - \partial_i \widetilde{\varepsilon}^{\mathrm{E_c}}_{q}(\mathbf{k}_0)\right)\ \widetilde{u}^{\mathrm{E_c}}_{q,\mathbf{k}_0}. \tag{4.7.41}$$

Similar to the argument for sequential continuity of the approximate, modified energy bands, we conclude from the fact that $\{\mathbf{k}_\ell\}_{\ell\in\mathbb{N}^*} \subset \Omega_j$ and only a finite number of possibilities exist for the choice of $j \in \{2,\ldots,2^M\}$, that Equation (4.7.41) also holds for any sequence $\{\mathbf{k}_\ell\}_{\ell\in\mathbb{N}^*} \subset \mathbb{B}_\delta(\mathbf{k})$. It is now straightforward to conclude from Equation (4.7.41) that $\lim_{\ell\to\infty} \partial_i \widetilde{u}^{\mathrm{E_c}}_{q,\mathbf{k}_\ell} = \partial_i v_{\mathbf{k}_0}$ as claimed since $\partial_i v_{\mathbf{k}_0}$ by definition also satisfies the equation

$$\left(\mathcal{A}_{\mathbf{k}_0} - \widetilde{\varepsilon}^{\mathrm{E_c}}_{q}(\mathbf{k}_0)\right) \partial_i v_{\mathbf{k}_0} = -\left(\partial_i \mathcal{A}_{\mathbf{k}_0} - \partial_i \widetilde{\varepsilon}^{\mathrm{E_c}}_{q}(\mathbf{k}_0)\right)\ \widetilde{u}^{\mathrm{E_c}}_{q,\mathbf{k}_0}.$$

As a consequence, $\widetilde{u}^{\mathrm{E_c}}_{q,\mathbf{k}}$ is of class $\mathscr{C}^1$ at $\mathbf{k} = \mathbf{k}_0$ as claimed. Noting that we picked an arbitrary $q \in \{1\ \ldots, p\}$ where $p = \dim \mathcal{A}_{\mathbf{k}_0} \leqslant M^-_{\mathrm{E_c}}$ completes the proof of differentiability of order one of the bounded energy band eigenfunctions.

By making use of this first order differentiability, we can perform a similar demonstration involving limits of finite-difference approximations of second order derivatives in order to establish $\mathscr{C}^2$ regularity of the modified energy band $\widetilde{\varepsilon}^{\mathrm{E_c}}_{q}$ at $\mathbf{k}_0$. For the sake of brevity, we omit the details of these (and higher order differentiability) demonstrations.

□

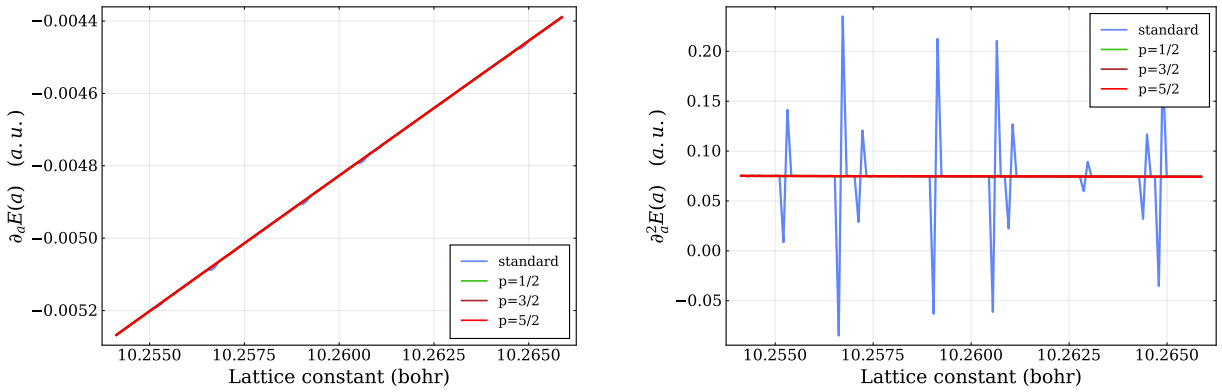
4.8 Perspectives

In the preceding section, our results were presented for a low cut-off energy $\mathrm{E_c} = 5$ Ha. It is a priori not clear from these results that our method offers advantages for practical applications with a higher $\mathrm{E_c}$. Indeed, if we compute the total energy of FCC crystalline silicon as a function of the lattice parameter for $\mathrm{E_c} = 80$ Ha rather than with $\mathrm{E_c} = 5$ Ha, as in Figure 4.12, the standard calculation gives a smooth curve for the eye-norm. Actually, more accurate numerical investigation shows that the first derivative of the total energy seems continuous and mostly agrees with the one obtained with the modified operator method, but that the second derivative is discontinuous. This is illustrated in Figure 4.15, where the first (left plot) and second (right plot) derivatives computed by finite differences are plotted for lattice parameters in a narrow range around equilibrium lattice constant $a_0 = 10.26$ bohr and energy cutoff $\mathrm{E_c} = 80$ Ha.

These preliminary results point toward possible applications of our method. In a subsequent study, we will investigate the effect of the modified Galerkin method on the computation of integrals over the Brillouin zone.

Acknowledgements

The authors thank Antoine Levitt and François Gygi for useful comments and fruitful discussions. The first two authors also thank IPAM where portions of this work were completed (March-June 2022). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 810367).



(A) First derivative

(B) Second derivative

Figure 4.15 – First and second derivatives of the energy per unit volume of FCC silicon as a function of the lattice parameter a for the unmodified and modified Galerkin discretizations for different value of blow-up parameters and for a high cut-off energy $\mathrm{E_c} = 80$ Ha. Derivatives are computed using a two-point centered finite difference approximation with step size $\Delta a = 10^{-4}$ bohr. The second derivative obtained with the standard discretization is discontinuous.

Chapter 5

Contributions to the Julia electronic structure eco-system: models for Twitsted-Bilayer Graphene

The work described in Section 5.3 has been done in collaboration with Étienne Polack. We are thankful to Éric Cancès, Michael Herbst, Antoine Levitt and Louis Garrigue for useful discussions.

Abstract This chapter describes two numerical contribution related to the simulation of multilayer 2D materials in Julia language [Bez+17], with an emphasis on twisted-bilayer graphene. The first contribution is a package, built as an overlay to the DFTK [HLC21] code, that implements the effective models for twisted-bilayer graphene presented in [CGG23]. The second is an application of [Bak+18] in the computation of tight-binding matrix elements for twisted-bilayer graphene.

Contents

5.1 Introduction

In this chapter, we describe two numerical contributions in Julia language [Bez+17] for the simulation of Twisted-Bilayer Graphene (TBG). Note that these contributions could also be used for the simultation of other multilayer 2D materials. This moiré material is made of two layers of atomically thin graphene sheets, stacked on each other with relative twist angle θ. Among all moiré materials, TBG stands out by its simplicity (it is only made of carbon atoms) and experimental accessibility (starting from the seminal "scotch-tape" exfoliation and isolation of graphene in 2004 [Net+09]). In addition, this material showcased exotic quantum phenomena including correlated insulating states and unconventional superconductivity [Cao+18] for some small twist angles, now referred to as *magic angles*. The understanding of *Magic Angle Twisted Bilayer Graphene* (MATBG) at experimental and theoretical level holds great promises toward the understanding of strong electronic correlation.

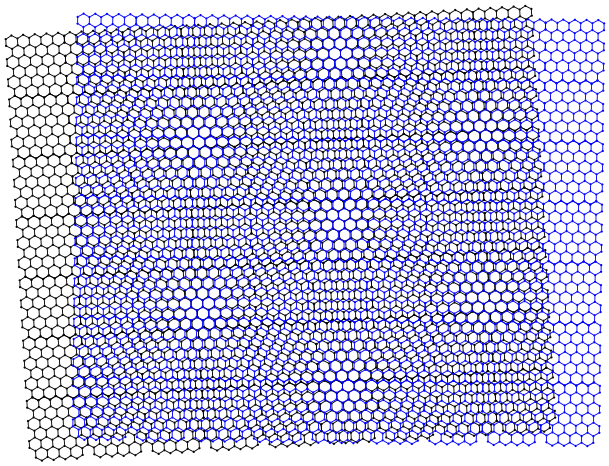


Figure 5.1 – A sample of twisted-bilayer graphene (TBG). The twist angle between the two graphene layers (respectively black and blue on the picture) creates a characteristic moiré pattern. *Source: adapted from Wikipedia Commons.* The corresponding *moiré lattice* seems periodic at mesoscopic scale.

When it comes to theory, moiré materials introduce new challenges. While the theoretical and computational study of solid materials is generally simplified by the use of translational symmetry and Bloch theory, moiré materials such as TBG are generally non-periodic. One way around this problem is to see that, at mesoscopic scale, TBG looks almost like a periodic crystal, with the associated lattice known as the *moiré lattice*. The smaller the angle, the more atoms in the moiré cell, with the typical MATBG cell containing of the order of 11,000 carbon atoms. Based on this *quasi-periodicity* property, several models have brought some insights about the electronic structure of TBG, mainly focusing on small twist angles and low-excitation energies. In that regime, the physics of interest seem to emerge from the interaction between the electronic states located in the $\mathbf{K}$ and $\mathbf{K}'$ valleys of each constituent layer. Those $\mathbf{k}$-points, also known as *Dirac points*, are quasi-momenta in the Brillouin zone of graphene for which the conduction and valence bands intersect conically (see Figure 5.2).

Introduced in 2011, the Bistritzer-MacDonald (BM) [BM11] is the most standard model for TBG. It is a continuous model, which treats TBG as a smooth system with a periodic potential determined by the moiré pattern. The BM model additionally assumes that the electrons are independent. It does not couple the spin-components and neglects the inter-valley coupling (the interaction between $\mathbf{K}$ and $\mathbf{K}'$ type states). In [BM11], the authors notably predict the appearance of a flat band in the TBG band diagram for a series of magic angles, which might be related to the observed exotic correlation phenomena [Po+18].

The BM model is an effective model which depends on three parameters, fine-tuned to match experimental results. A natural question follows, whether the BM model can be deduced from first-principle. In [CGG23], the authors answer that question by deriving a continuous model of TBG which is similar to the BM model in certain regimes, but contains additional terms. In contrast to other approaches, the CGG model is derived directly from an approximate Kohn-Sham Hamiltonian. While the authors of [CGG23] provide numerical simulations of their models, their research code was built as a proof of concept, not

necessarily flexible nor sustainable in the long term.

In this chapter, our first contribution is to lay the fundations for a Julia code, that offers a user-friendly playground for the simulation of 2D materials, starting with the implementation of the BM and CGG models for TBG.

Continuous models are computationally efficient, while being able to capture the essential physics of moiré materials. However, they neglect the possible strong correlation between electrons and the precise atomic arrangement of atoms in the moiré cell. As a result, they are by nature unable to predict the atomic-scale details that might affect their electronic properties. In condensed matter physics, a common approach to compute many-body interactions in strongly correlated materials is to use a tight-binding approximation with a two-body electron-electron repulsion term. These models are typically parameterized using Wannier functions of monolayer graphene corresponding to the electronic bands in the energy window of interest. Once obtained by an *ab initio* computation, Wannier functions are used to compute the matrix elements corresponding to the one-body and two-body interactions of the tight-binding Hamiltonian. In the case of TBG however, the large number of atoms in the moiré cell makes this approach difficult to use in practice.

One way around this problem is provided by [Bak+18], where the authors propose a systematic procedure to expand a Wannier function in a basis of Gaussian Type Orbitals (GTOs), for which the matrix elements of the tight-binding Hamiltonian can be computed analytically. This kind of basis set is widely used in quantum chemistry for molecular systems, and several libraries are available that already handle the computation of GTO integrals.

In the second contribution of this chapter, we describe a small numerical experiment where we applied the routine proposed in [Bak+18] to a Wannier function of monolayer graphene, corresponding to a targeted valence band.

This chapter is organized as follows. In Section 5.2, we describe the mathematical formalism for the modelization of monolayer and bilayer graphene. We recall in particular the important properties of graphene used throughout this chapter. Section 5.3 is concerned with our first numerical contribution. After recalling the formulation of the rescaled moiré-periodic BM and CGG models, we compute their discretization in a moiré plane-wave basis. Though mostly calculatory, this section serves as the basis for a public documentation for our package. We then present our code `TwistedBilayerGraphene` that runs BM and CGG computations on TBG, and discuss the perspectives of development of this package.

In Section 5.4, we discuss the compression of Wannier functions on symmetry adapted Gaussian basis sets. We start by finding and symmetry adapted basis for the Wannier function of graphene corresponding to a p_z-like valence band. We then describe some preliminary results.

5.2 The mathematical description of graphene systems

There exist several conventions for the mathematical description of monolayer and twisted-bilayer graphene. In order to ease the presentation of the effective CGG model, we adopt the same notations as in [CGG23], which we briefly recall bellow.

5.2.1 Monolayer graphene

Monolayer graphene is a two-dimensional material made of an atomically thin layer of carbon atoms arranged on a honeycomb lattice, with minimal inter-atomic distance $a \simeq 2.68$ Bohr. Mathematically, it is described by the 2D Bravais lattice (Figure 5.2)

$$\mathcal{R}_{\mathbf{x}} := \mathbf{a}_1\mathbb{Z} + \mathbf{a}_2\mathbb{Z}, \quad \mathbf{a}_1 := a_0 \begin{pmatrix} 1/2 \\ -\sqrt{3}/2 \end{pmatrix}, \quad \mathbf{a}_2 := a_0 \begin{pmatrix} 1/2 \\ \sqrt{3}/2 \end{pmatrix}, \tag{5.2.1}$$

which depends on the lattice constant $a_0 = \sqrt{3}a$. The unit cell Ω contains two atoms, often labeled "A" and "B", at respective positions

$$R_{\mathrm{A}} = \frac{1}{3}(\mathbf{a}_1 - \mathbf{a}_2) \quad \text{and} \quad R_{\mathrm{B}} = \frac{1}{3}(\mathbf{a}_2 - \mathbf{a}_1).$$

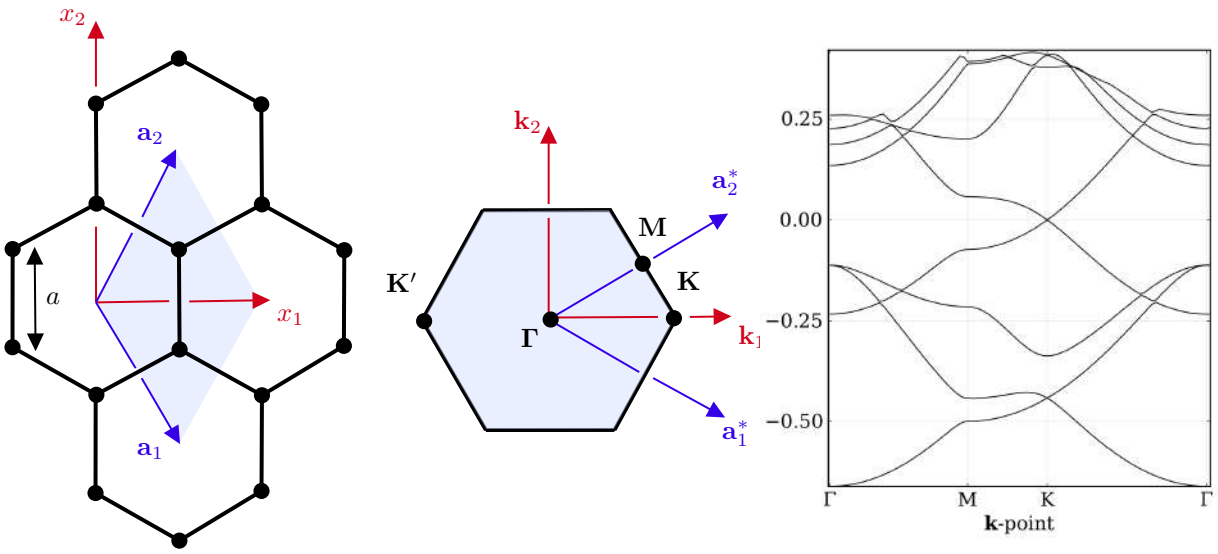


Figure 5.2 – (Left) Graphene layer and associated Bravais lattice. The red (resp. blue) arrows display the Cartesian x_1 and x_2 directions (resp. the lattice vectors). The unit cell Ω appears in blue. The minimum inter-atomic distance a is shown on the left. (Middle) The corresponding Brillouin zone. The color convention is the same as (Left). The plot also show some high-symmetry $\mathbf{k}$-points of interest, among which the $\mathbf{K}$ and $\mathbf{K}'$ Dirac points. (Right) Band diagram of graphene along the path $\mathbf{\Gamma} \to \mathbf{M} \to \mathbf{K} \to \mathbf{\Gamma}$. The vertical axis displays the energies of the bands in Hartree. The energies are shifted so that the Fermi level appears at zero Hartree on the graph. The conduction and valence bands cross conically at point $\mathbf{K}$.

As usual, we denote $\mathcal{R}_{\mathbf{x}}^* := \mathbf{a}_1^*\mathbb{Z} + \mathbf{a}_2^*\mathbb{Z}$ its reciprocal lattice with unit cell Ω^*. We denote the position variable by $\mathbf{r} = (\mathbf{x}, z) \in \mathbb{R}^3$ where $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$ and $z \in \mathbb{R}$ are respectively the longitudinal (in-plane) and transverse (out-of-plane) position variables. We may also denote $\mathbf{0} = (0, 0)$ the in-plane origin.

The Bloch fibers $\mathrm{H}_{\mathbf{k}}$ of the graphene Hamiltonian are labeled by a 2D quasi-momentum $\mathbf{k} \in \mathbb{R}^2$ and read as

$$\mathrm{H}_{\mathbf{k}} = \frac{1}{2}(-i\nabla_{\mathbf{x}} + \mathbf{k})^2 - \frac{1}{2}\partial_z^2 + V \tag{5.2.2}$$

where the $\mathcal{R}_{\mathbf{x}}$-periodic potential V is typically obtained via Kohn-Sham DFT. The operators $\mathrm{H}_{\mathbf{k}}$ act on $L^2_{\mathrm{per}}(\Omega \times \mathbb{R})$. While they do not have compact resolvent, the Bloch fibers have discrete eigenvalues below the bottom of their essential spectrum forming the so-called valence bands and low-energy conduction bands, pictured in band diagram of Figure 5.2. The bands exhibits two characteristic conical intersections at Fermi level and *Dirac* points:

$$\mathbf{K} = \frac{1}{3}(\mathbf{a}_1^* + \mathbf{a}_2^*) \quad \text{and} \quad \mathbf{K}' = -\mathbf{K}. \tag{5.2.3}$$

As in [CGG23], we define the following operators which are useful to describe the graphene and TBG symmetries. For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$, $\theta \in \mathbb{R}$ and $f : \mathbb{R}^3 \to \mathbb{C}$ we denote:

$$\begin{aligned}
&(\tau_{\mathbf{y}} f)(\mathbf{x}, z) := f(\mathbf{x} - \mathbf{y}, z) && \text{(horizontal translation of vector } \mathbf{y}) \\
&(\mathcal{R}_\theta f)(\mathbf{x}, z) := f(\mathcal{R}_{-\theta}\mathbf{x}, z) && \text{(rotation of angle } \theta \text{ around the } z\text{-axis}) \\
&(\mathfrak{R} f)(x, y, z) := f(x, -y, -z) && \text{(rotation of angle } \pi \text{ around the } x\text{-axis}) \\
&(\mathscr{P} f)(\mathbf{x}, z) := (\mathcal{R}_\pi f)(\mathbf{x}, z) = f(-\mathbf{x}, z) && \text{(in-plane parity operator)} \\
&(\mathscr{C} f)(\mathbf{x}, z) := \overline{f(\mathbf{x}, z)} && \text{(complex conjugation)} \\
&(\mathscr{S} f)(\mathbf{x}, z) := f(\mathbf{x}, -z) && \text{(mirror symmetry w.r.t. the plane } z = 0)
\end{aligned}$$

The space group of monolayer graphene is given by the semi-direct product

$$\mathrm{Dg80} = \mathrm{D}_{6h} \ltimes \mathcal{R}_{\mathbf{x}} \tag{5.2.4}$$

where D_{6h} is the group generated by $\mathcal{R}_{\frac{\pi}{3}}$, $\mathscr{S}$, and $\mathfrak{R}$.

At Fermi level, the respective eigenspaces of the Bloch Hamiltonians $\mathrm{H}_{\mathbf{K}}$ and $\mathrm{H}_{\mathbf{K}'}$ are two dimensional. By a symmetry argument detailed in [FW12], we can always choose for any of these eigenspaces a basis of Bloch waves $\Phi_1(\mathbf{x}) := u_1(\mathbf{x})e^{i\mathbf{K}\cdot x}$ and $\Phi_2(\mathbf{x}) = u_2(\mathbf{x})e^{i\mathbf{K}\cdot x}$ such that

$$\mathcal{R}_{2\frac{\pi}{3}}\Phi_1 = e^{i\frac{2\pi}{3}}\Phi_1 \quad \text{and} \quad \mathcal{R}_{2\frac{\pi}{3}}\Phi_2 = e^{-i\frac{2\pi}{3}}\Phi_2. \tag{5.2.5}$$

We call (Φ_1, Φ_2) a symmetry adapted basis. In particular, it is such that the Fermi energy

$$v_{\mathrm{F}} := \langle \Phi_1, (-i\partial_{x_1})\Phi_2 \rangle$$

is a real number.

5.2.2 Bilayer graphene

We now focus on systems built from two parallel layers of graphene, separated by a constant distance $d > 0$. As in [CGG23], the layers are placed at $z = \frac{d}{2}$ and $z = -\frac{d}{2}$. From that configuration, the first system of interest in this chapter is (untwisted) bilayer graphene (UBG), built by introducing a small disregistry $\mathbf{y} \in \Omega$ of the top layer. The Kohn-Sham potential of the top layer is therefore obtained with

$$V^{\mathrm{top}}_{d,\mathbf{y}}(\mathbf{x}, z) = \tau_{\mathbf{y}} V\left(\mathbf{x}, \mathbf{z} - \frac{d}{2}\right) = V\left(\mathbf{x} - \mathbf{y}, \mathbf{z} - \frac{d}{2}\right). \tag{5.2.6}$$

The cases $\mathbf{y} = \mathbf{0}$ and $\mathbf{y} = \frac{1}{2}(\mathbf{a}_1 + \mathbf{a}_2)$ correspond respectively to the so-called "AA" and "AB" stacking. Untwisted bilayer graphene has a crystalline structure with four atoms in its unit cell. We can therefore easily compute Kohn-Sham potential, denoted $V^{(2)}_{d,\mathbf{y}}$, for example with plane-wave DFT.

On the other hand, twisted bilayer graphene is constructed by rotating the top layer counterclockwise by $-\frac{\theta}{2}$ and the bottom layer by $\frac{\theta}{2}$ around the z axis. For a given angle $\theta \in \mathbb{R}$, let $c_\theta := \cos\frac{\theta}{2}$ and $\varepsilon_\theta := 2\sin\frac{\theta}{2}$, and introduce the twisting unitary operator

$$(U_{d,\theta} f)(\mathbf{x}, z) = f\left(\mathcal{R}^*_{-\theta}\mathbf{x},\, z - \frac{d}{2}\right) = f\left(c_\theta \mathbf{x} - \frac{1}{2}\varepsilon_\theta J\mathbf{x},\, z - \frac{d}{2}\right) \tag{5.2.7}$$

with

$$J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{5.2.8}$$

The Kohn-Sham potentials of each individual layers of TBG are obtained by

$$V^{\mathrm{top}}_{d,\theta} = U_{d,\theta} V \quad \text{and} \quad V^{\mathrm{bottom}}_{d,\theta} = U^{-1}_{d,\theta} V.$$

The TBG is not periodic, except for a countable set of twist angles, but is approximated as a moiré periodic system, with the associated rescaled moiré lattice and unit cell

$$\mathcal{R}_{\mathrm{M}} := J\mathcal{R}_{\mathbf{x}} \quad \text{and} \quad \Omega_{\mathrm{M}} := J\Omega.$$

In Table 5.1, we gather all the notations for the description of TBG, used in the following section. The $\mathbf{q}$-points and moiré Dirac points $\mathbf{K}_1$ and $\mathbf{K}_2$ are quasi-momenta of interest in the moiré reciprocal space.

5.3 Effective models for the electronic structure of Twisted-Bilayer Graphene

Let us now turn to our first contribution. For the sake of completeness, we report briefly below the notations from [CGG23], used throughout this chapter. However, providing a proper definition and physical meaning for all objects would inevitably result in a fully fledged re-writing of the first sections of that paper. We therefore refer the interested reader to the original work [CGG23] for further information on the CGG model. Our more concise exposition should in turn serve as a guide for the practical implementation of the models.

Table 5.1 – Notations and points of interest regarding the definition of monolayer and twisted bilayer graphene. The index M denotes the quantities of the moiré lattices $\mathcal{R}_{\mathrm{M}}$ and $\mathcal{R}^*_{\mathrm{M}}$.

Common	
J matrix	$J := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$
Rescalling	$c_\theta := \cos\left(\frac{\theta}{2}\right)$ $\quad \varepsilon_\theta := 2\sin\left(\frac{\theta}{2}\right)$
Others	$\mathbf{G}^{\mathbf{k}} := \mathbf{G} + \mathbf{k}$ $\quad k_D := 4\pi/(3a_0)$
Monolayer Graphene	
Lattice	$\mathbf{a}_1 := a_0 \begin{pmatrix} 1/2 \\ -\sqrt{3}/2 \end{pmatrix}$ $\quad \mathbf{a}_2 := a_0 \begin{pmatrix} 1/2 \\ \sqrt{3}/2 \end{pmatrix}$
Reciprocal Lattice	$\mathbf{a}_1^* := \sqrt{3}k_D \begin{pmatrix} \sqrt{3}/2 \\ -1/2 \end{pmatrix}$ $\quad \mathbf{a}_2^* := \sqrt{3}k_D \begin{pmatrix} \sqrt{3}/2 \\ 1/2 \end{pmatrix}$
Dirac points	$\mathbf{K} := \frac{1}{3}(\mathbf{a}_1^* + \mathbf{a}_2^*) = k_D \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ $\quad \mathbf{K}' := -\mathbf{K}$
Twisted bilayer graphene	
Lattice	$\mathbf{a}_{1,\mathrm{M}} := J\mathbf{a}_1 := a_0 \begin{pmatrix} -\sqrt{3}/2 \\ -1/2 \end{pmatrix}$ $\quad \mathbf{a}_{2,\mathrm{M}} := J\mathbf{a}_2 := a_0 \begin{pmatrix} \sqrt{3}/2 \\ -1/2 \end{pmatrix}$
q–points	$\mathbf{q}_1 := J\mathbf{K}$ $\quad \mathbf{q}_2 := \frac{1}{3}(-2\mathbf{a}^*_{1,\mathrm{M}} + \mathbf{a}^*_{2,\mathrm{M}})$ $\quad \mathbf{q}_3 := \frac{1}{3}(\mathbf{a}^*_{1,\mathrm{M}} - 2\mathbf{a}^*_{2,\mathrm{M}})$
Moiré Dirac points	$\mathbf{K}_1 := -\mathbf{q}_2$ $\quad \mathbf{K}_2 := \mathbf{q}_3$

5.3.1 Notations and conventions

These notations and remarks are useful for the exposition of BM and CGG.

Multicomponent operators

The effective models studied below are four-component models, which means that a given state α is not represented by a scalar wave-function but by a four-component wave functions

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix}.$$

In addition, BM and CGG approximate the real TBG as a mesoscale 2D system. The BM and CGG quantum states are thus functions of $L^2(\mathbb{R}^2;\mathbb{C}^4)$. For multicomponent Hamiltonian, some operations apply to each component individually, while others mix several components. In physics, the distinction between the two kinds of operations is often implicit, which might render the formulation of the effective TBG models difficult for the unfamiliar reader. For the sake of clarity, we detail three specific operations that might cause some misunderstanding in following exposition.

1. In the section bellow, the gradient operator always writes $(-i\nabla_{\mathbf{x}})(\bullet) = \begin{pmatrix} -i\partial_{x_1}(\bullet) \\ -i\partial_{x_2}(\bullet) \end{pmatrix}$, whether applied to a single or multicomponent function. The partial derivatives then act component-wise. For example, we have for all 2-component vector $\alpha \in L^2(\mathbb{R}^2, \mathbb{C}^2)$

$$(-i\nabla_{\mathbf{x}}) \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} -i\partial_{x_1} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \\ -i\partial_{x_2} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} -i \begin{pmatrix} \partial_{x_1}\alpha_1 \\ \partial_{x_1}\alpha_2 \end{pmatrix} \\ -i \begin{pmatrix} \partial_{x_2}\alpha_1 \\ \partial_{x_2}\alpha_2 \end{pmatrix} \end{pmatrix}. \tag{5.3.1}$$

2. A given 2×2 (or 4×4) matrix valued function $M\in L^2(\mathbb{R}^2;\mathbb{C}^{2\times 2})$ acts on a vector of components with the usual matrix vector product $M\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}=\begin{pmatrix}M_{11}\alpha_1+M_{12}\alpha_2\\ M_{21}\alpha_1+M_{22}\alpha_2\end{pmatrix}$. Note the association with the gradient

$$M(-i\nabla)\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}=\begin{pmatrix}-iM\begin{pmatrix}\partial_{x_1}\alpha_1\\ \partial_{x_1}\alpha_2\end{pmatrix}\\ -iM\begin{pmatrix}\partial_{x_2}\alpha_1\\ \partial_{x_2}\alpha_2\end{pmatrix}\end{pmatrix}. \tag{5.3.2}$$

3. For all $f,g\in L^2(\mathbb{R}^2;\mathbb{C})$ and $M\in L^\infty(\mathbb{R}^2;\mathbb{C}^{d\times d})$, the notation $\langle f,Mg\rangle_{L^2(\mathbb{R}:\mathbb{C})}$ refers to the complex valued $d\times d$ matrix with entries

$$\langle f,Mg\rangle_{L^2(\mathbb{R}^2,\mathbb{C})}\in\mathbb{C}^{2\times 2}\quad\text{and}\quad[\langle f,Mg\rangle_{L^2(\mathbb{R}^2,\mathbb{C})}]_{ij}=\langle f,[M]_{ij}g\rangle_{L^2(\mathbb{R}^2);\mathbb{C}}\quad\forall\,1\leqslant i,j\leqslant d. \tag{5.3.3}$$

Rotated Pauli matrices

The rotated Pauli matrices are operators involved in the kinetic term of the BM and CGG models. We recall that standard Pauli matrices are defined by

$$\sigma_1:=\begin{pmatrix}0&1\\1&0\end{pmatrix},\quad \sigma_2:=\begin{pmatrix}0&-i\\i&0\end{pmatrix},\quad\text{and}\quad \sigma_3:=\begin{pmatrix}1&0\\0&-1\end{pmatrix}.$$

For all $\theta\in\mathbb{R}$ and sign $\eta\in\{-1,1\}$, we define the $\eta\frac{\theta}{2}$ rotated Pauli matrix as

$$\boldsymbol{\sigma}_{\eta\theta/2}:=e^{-\eta i\frac{\theta}{4}\sigma_3}(\sigma_1,\sigma_2)e^{\eta i\frac{\theta}{4}\sigma_3}=\left(\begin{pmatrix}0&e^{-i\eta\frac{\theta}{2}}\\e^{i\eta\frac{\theta}{2}}&0\end{pmatrix},\begin{pmatrix}0&-ie^{-i\eta\frac{\theta}{2}}\\ie^{i\eta\frac{\theta}{2}}&0\end{pmatrix}\right). \tag{5.3.4}$$

This operator mixes components. Its action on $L^2(\mathbb{R}^2;\mathbb{C}^2)$ reads for all 2-component vector $\begin{pmatrix}a\\b\end{pmatrix}\in\mathbb{C}^2$

$$\boldsymbol{\sigma}_{\eta\theta/2}\cdot\begin{pmatrix}a\\b\end{pmatrix}=\begin{pmatrix}0&e^{-i\eta\frac{\theta}{2}}(a-ib)\\e^{i\eta\frac{\theta}{2}}(a+ib)&0\end{pmatrix}.$$

5.3.2 The BM and CGG eigenvalue problems

In this section, we fix $\mathbf{x}\in\mathbb{R}^2$ and an angle $\theta\in\mathbb{R}$.

5.3.2.1 Rescaled moiré-periodic BM Hamiltonian

The rescaled moiré-periodic formulation of the BM Hamiltonian is the self-adjoint operator

$$\mathrm{H}^{\mathrm{BM}}_\theta:=P\begin{pmatrix}v_{\mathrm{F}}\boldsymbol{\sigma}_{-\theta/2}\cdot(-i\nabla_{\mathbf{x}})&\varepsilon_\theta^{-1}\mathbf{V}\\ \varepsilon_\theta^{-1}\mathbf{V}^*&v_{\mathrm{F}}\boldsymbol{\sigma}_{\theta/2}\cdot(-i\nabla_{\mathbf{x}})\end{pmatrix}P^* \tag{5.3.5}$$

on $L^2(\mathbb{R}^2;\mathbb{C}^4)$ with domain $H^1(\mathbb{R}^2;\mathbb{C}^4)$. The Bistritzer-MacDonald potential $\mathbf{V}:\mathbb{R}^2\to\mathbb{C}^{2\times 2}$ is the matrix valued function

$$\mathbf{V}(\mathbf{x}):=\sum_{j=1}^3 V_j e^{-i\mathbf{q}_j\cdot\mathbf{x}},\quad\text{where}\quad V_j:=\begin{pmatrix}w_{\mathrm{AA}}&w_{\mathrm{AB}}\overline{\omega}^{j-1}\\ w_{\mathrm{AB}}\omega^{j-1}&w_{\mathrm{AA}}\end{pmatrix}.$$

In the above expression, $\omega=e^{i2\pi/3}$, w_{AA} and w_{AB} are the two real parameters describing the inter-layer coupling in AA and AB stackings and

$$P(\mathbf{x}):=\begin{pmatrix}e^{i\mathbf{K}_1\cdot\mathbf{x}}\mathbb{I}_2&0\\0&e^{i\mathbf{K}_2\cdot\mathbf{x}}\mathbb{I}_2\end{pmatrix}.$$

Note that since $\omega^2=\overline{\omega}$ and $\overline{\omega}^2=\omega$, one simply has $V_3=V_2^*$. The operator P is a gauge transformation that allows the BM Hamiltonian to be $\mathcal{R}_{\mathrm{M}}^*$ periodic.

Let us give a simpler formulation of (5.3.5). For the sake of clarity, we use the block matrix notation

$$P(\mathbf{x}) =: \begin{pmatrix} P_{\mathbf{K}_1}(\mathbf{x}) & 0 \\ 0 & P_{\mathbf{K}_2}(\mathbf{x}) \end{pmatrix}.$$

Doing the full matrix multiplication in Equation 5.3.5 we obtain

$$\mathrm{H}_\theta^{\mathrm{BM}} = \begin{pmatrix} P_{\mathbf{K}_1}(\mathbf{x}) v_{\mathrm{F}} \left[\boldsymbol{\sigma}_{-\theta/2} \cdot (-i\nabla_{\mathbf{x}})\right] P_{\mathbf{K}_1}(\mathbf{x})^* & P_{\mathbf{K}_1}(\mathbf{x}) \varepsilon_\theta^{-1} \mathbf{V}(\mathbf{x}) P_{\mathbf{K}_2}(\mathbf{x})^* \\ P_{\mathbf{K}_2}(\mathbf{x}) \varepsilon_\theta^{-1} \mathbf{V}(\mathbf{x})^* P_{\mathbf{K}_1}(\mathbf{x})^* & P_{\mathbf{K}_2}(x) \left[v_{\mathrm{F}} \boldsymbol{\sigma}_{\theta/2} \cdot (-i\nabla_{\mathbf{x}})\right] P_{\mathbf{K}_2}(x)^* \end{pmatrix}.$$

For the diagonal terms, one has for all sign $\eta \in \{-1, 1\}$

$$v_{\mathrm{F}} \boldsymbol{\sigma}_{\eta\theta/2} \cdot (-i\nabla_{\mathbf{x}}) = v_{\mathrm{F}} \begin{pmatrix} 0 & e^{-\eta i \frac{\theta}{2}} (-\partial_y - i\partial_x) \\ e^{\eta i \frac{\theta}{2}} (\partial_y - i\partial_x) & 0 \end{pmatrix},$$

so that for $j = 1, 2$

$$P_{\mathbf{K}_j}(\mathbf{x}) v_{\mathrm{F}} [\boldsymbol{\sigma}_{\eta\theta/2} \cdot (-i\nabla_{\mathbf{x}})] P_{\mathbf{K}_j}(x)^* = v_{\mathrm{F}} \boldsymbol{\sigma}_{\eta\theta/2} \cdot (-i\nabla - \mathbf{K}_j).$$

For the off-diagonal term, we use the fact that $\mathbf{V}$ is a multiplicative potential to write

$$P_{\mathbf{K}_j}(\mathbf{x}) \varepsilon_\theta^{-1} \mathbf{V}(\mathbf{x}) P_{\mathbf{K}_{j'}}(\mathbf{x})^* = e^{i(\mathbf{K}_j - \mathbf{K}_{j'}) \cdot \mathbf{x}} \varepsilon_\theta^{-1} \mathbf{V}(\mathbf{x}).$$

As a result we obtain

$$\boxed{\mathrm{H}_\theta^{\mathrm{BM}} = \begin{pmatrix} v_{\mathrm{F}} \boldsymbol{\sigma}_{-\theta/2} \cdot (-i\nabla - \mathbf{K}_1) & e^{i(\mathbf{K}_1 - \mathbf{K}_2) \cdot \mathbf{x}} \varepsilon_\theta^{-1} \mathbf{V}(\mathbf{x}) \\ e^{-i(\mathbf{K}_1 - \mathbf{K}_2) \cdot \mathbf{x}} \varepsilon_\theta^{-1} \mathbf{V}(\mathbf{x})^* & v_{\mathrm{F}} \boldsymbol{\sigma}_{\theta/2} \cdot (-i\nabla - \mathbf{K}_2) \end{pmatrix}.} \tag{5.3.6}$$

5.3.2.2 Rescaled moiré-periodic CGG Hamiltonian

The writing of the CGG Hamiltonian requires two additional definitions. Let V be the monolayer graphene Kohn-Sham Hamiltonian. The first feature of CGG is do define for all $z \in \mathbb{R}$ a TBG potential correction term

$$V_{\mathrm{int},d}(z) = \fint_\Omega V_{\mathrm{int},d,\mathbf{y}}(z) \mathrm{d}\mathbf{y}. \tag{5.3.7}$$

For all disregistry $\mathbf{y} \in \Omega$, $V_{\mathrm{int},d,\mathbf{y}}$ is defined by

$$V_{\mathrm{int},d,\mathbf{y}}(z) = \fint_\Omega \left(V_{d,\mathbf{y}}^{(2)}(\mathbf{x}, z) - V\left(\mathbf{x}, z + \frac{d}{2}\right) - V\left(\mathbf{x} - \mathbf{y}, z - \frac{d}{2}\right) \right) \mathrm{d}\mathbf{x}, \tag{5.3.8}$$

where we recall that $V_{d,\mathbf{y}}^{(2)}$ is the Kohn-Sham potential of untwisted bilayer graphene with disregistry $\mathbf{y}$. The correction term $V_{\mathrm{int},d}$ has been introduced in [Tri+16] to define an approximate Kohn-Sham potential from bilayer 2D materials. Second let $f, g \in L^2_{\mathrm{loc}}(\mathbb{R}^3; \mathbb{C})$ be $\mathcal{R}_{\mathbf{x}}$-periodic in their in-plane variable. For all signs $\eta, \eta' \in \{0, 1\}$, we define the bilayer scalar product

$$(\!(f, g)\!)_d^{\eta\eta'}(\mathbf{X}) = \int_{\Omega \times \mathbb{R}} \overline{f\left(\mathbf{x} - \eta \frac{1}{2} J\mathbf{X}, z - \eta \frac{d}{2}\right)} g\left(\mathbf{x} - \eta' \frac{1}{2} J\mathbf{X}, z - \eta' \frac{d}{2}\right) \mathrm{d}\mathbf{x}\mathrm{d}z.$$

We can now write the effective rescaled moiré-periodic CGG Hamiltonian. It expresses for all $\mathbf{x} \in \mathbb{R}^2$ as the sum of three terms of order -1, 0 and 1 in ε_θ:

$$\boxed{\mathrm{H}_{d,\theta} := \varepsilon_\theta^{-1} \begin{pmatrix} \widetilde{\mathbb{W}}_d^+(\mathbf{x}) & \widetilde{\mathbb{V}}_d(\mathbf{x}) \\ \widetilde{\mathbb{V}}_d(\mathbf{x})^* & \widetilde{\mathbb{W}}_d^-(\mathbf{x}) \end{pmatrix} + \begin{pmatrix} v_{\mathrm{F}} \boldsymbol{\sigma}_{-\theta/2} \cdot (-i\nabla_{\mathbf{K}_1}) & c_\theta J \widetilde{A}_d(\mathbf{x}) \cdot (-i\nabla_{\boldsymbol{K}_2}) \\ c_\theta J \widetilde{A}_d^*(\mathbf{x}) \cdot (-i\nabla_{\boldsymbol{K}_1}) & v_{\mathrm{F}} \boldsymbol{\sigma}_{\theta/2} \cdot (-i\nabla_{\boldsymbol{K}_2}) \end{pmatrix} - \frac{\varepsilon_\theta}{2} \nabla \cdot (\mathbb{S}_d(\mathbf{x}) \nabla \bullet)}$$

(5.3.9)

Like BM, it is an unbouded self-adjoint operator with domain $H^1(\mathbb{R}^2; \mathbb{C}^4)$. The above terms are defined as follows:

1. the potentials $\widetilde{\mathbb{V}}_d$ and $\widetilde{\mathbb{W}}_d$ are 2×2 matrix-valued defined for all $1 \leqslant j, j' \leqslant 2$ as

$$[\widetilde{\mathbb{V}}_d]_{j,j'} := \left(\!\!\left(\left(V + V_{\mathrm{int},d}(\cdot + \tfrac{d}{2})\right) u_j, u'_j\right)\!\!\right)$$
$$[\widetilde{\mathbb{W}}_d]_{j,j'} := \left(\!\!\left(u_j \overline{u_{j'}}, V\right)\!\!\right)_d^{\pm\mp}(\mathbf{x}) + \left(W^{\pm}_{\mathrm{int},d}\right)_{j,j'}$$

with $\left(W^{\pm}_{\mathrm{int},d}\right)_{j,j'} := \displaystyle\int_{\Omega\times\mathbb{R}} (\overline{u_j} u'_j)\left(\mathbf{x}, z \mp \tfrac{d}{2}\right) V_{\mathrm{int},d}(z)\mathrm{d}\mathbf{x}\mathrm{d}z.$

Note that u_1 and u_2 are the symmetry adapted Bloch waves introduced in Section 5.2.1.

2. the overlap matrix $\mathbb{S}_d$ writes

$$\mathbb{S}_d(\mathbf{x}) := \begin{pmatrix} \mathbb{I}_2 & \widetilde{\Sigma}_d(\mathbf{x}) \\ \widetilde{\Sigma}_d^*(\mathbf{x}) & \mathbb{I}_2 \end{pmatrix}$$

where $\widetilde{\Sigma}_d$ is the 2×2 matrix-valued function $[\widetilde{\Sigma}_d]_{j,j'}(\mathbf{x}) = \left(\!\!\left(u_j, u'_j\right)\!\!\right)_d^{+-}(\mathbf{x})$.

3. $\widetilde{A}_d := (-i\nabla - \mathbf{q}_1)\widetilde{\Sigma}_d$, where $\mathbf{q}_1$ is as in Table 5.1.

5.3.2.3 The BM and CGG eigenvalue problems

After application of the Bloch theorem, the BM and CGG problems consist in solving the respective eigenvalue and generalized eigenvalue problems

$$\mathrm{H}^{\mathrm{BM}}_{\theta,\mathbf{k}}\alpha_{\mathbf{k}} = \varepsilon_{n,\mathbf{k}}\alpha_{\mathbf{k}} \quad (\mathrm{BM}) \qquad \mathrm{H}_{d,\theta,\mathbf{k}}\alpha_{n,\mathbf{k}} = \varepsilon_{n,\mathbf{k}} S_d \alpha_{n,\mathbf{k}} \quad (\mathrm{CGG}) \tag{5.3.10}$$

where for all n and $\mathbf{k} \in \Omega^*_{\mathrm{M}}$: $(\varepsilon_{n,\mathbf{k}}, \alpha_{n,\mathbf{k}}) \in \mathbb{R} \times H^1(\Omega_{\mathrm{M}}; \mathbb{C}^4)$.

Let us now remark that, in contrast with the other Hamiltonian operators encountered in this manuscript, the BM and CGG Hamiltonians are not bounded bellow. Indeed, the evolution of the quasi particles in the $\mathbf{K}$ and $\mathbf{K}'$ valleys close to the Fermi energy is mediated by a Dirac Hamiltonian, this coming from the specific conical crossing of the bands. The BM and CGG effective models therefore contain a momentum operator $\hat{\mathrm{P}} = -i\nabla$ whose spectrum is not bounded below. For that reason, the eigenvalues of interest belong to the bulk of the spectrum, near the Fermi energy $\mu_{\mathrm{F}} = 0$.

This has very practical consequences. Traditional eigensolvers are typically designed for problems with isolated eigenvalues at the bottom (or top) of the spectrum. They often only need matrix-vector product (application of the Hamiltonina on a trial state) and scalar product operations to compute these eigenvalues. This kind of matrix-free implementation saves a lot of memory, especially in Fourier discretization where the number of basis functions can be very large. Computing eigenvalues in the bulk of the spectrum comes with additional difficulties, such as spectral pollution [LS10], and extra care has to be taken in that case.

To get around this problem, our code works by assembling the full matrix of the BM and CGG Hamiltonians, and by selecting the spectrum around Fermi level. Still, we provide bellow the formulae for the BM and CGG terms as matrix-vector product, to serve for future developments.

5.3.3 Plane-wave discretization conventions

To start with, let us detail our discretization conventions. In order to match the conventions of DFTK, we define for all $\mathbf{G} \in \mathcal{R}^*$ the plane-wave of momentum $\mathbf{G}$ by

$$e_{\mathbf{G}}(\mathbf{r}) := \frac{1}{\sqrt{|\Omega|}} e^{i\mathbf{G}\cdot\mathbf{r}}, \qquad \forall \mathbf{r} \in \mathbb{R}^3. \tag{5.3.11}$$

It is such that for all $\mathbf{G}, \mathbf{G}' \in \mathcal{R}^*$, $\langle e_{\mathbf{G}}, e_{\mathbf{G}'}\rangle_{L^2_{\mathrm{per}}(\Omega)} = \delta_{\mathbf{G},\mathbf{G}'}$. The Fourier expansion of any $\mathcal{R}$-periodic function $u \in L^2_{\mathrm{per}}(\Omega)$ then writes

$$u(\mathbf{r}) := \frac{1}{\sqrt{|\Omega|}} \sum_{\mathbf{G}\in\mathcal{R}^*} u[\mathbf{G}] e^{i\mathbf{G}\cdot\mathbf{r}}. \tag{5.3.12}$$

The computation of the bilayer scalar products, needed for the CGG potential, require to discretize an integral over $\Omega \times \mathbb{R}$. We choose to compute this integral in Fourier space, by approximating the 2D monolayer graphene as a 3D-periodic system with lattice

$$\mathcal{R}_h = \mathcal{R}_{\mathbf{x}} \times h\mathbb{Z}, \quad \text{and} \quad \mathcal{R}_h^* = \mathcal{R}_{\mathbf{x}}^* \times \frac{2\pi}{h}\mathbb{Z} := \mathcal{R}_{\mathbf{x}}^* \times \mathcal{R}_{z,h}^* \tag{5.3.13}$$

with unit cell $\Omega_h = \Omega \times [0, h)$, and where $h > 0$ is the height of the supercell in the transverse direction. When necessary, we will split the momenta $\mathbf{G}$ of $\mathcal{R}_h^*$ as one in-plane and one out-of plane momenta:

$$\mathbf{G} = (\mathbf{G}_{\mathbf{x}}, \mathbf{G}_z) \quad \text{with} \quad \mathbf{G}_{\mathbf{x}} \in \mathcal{R}_{\mathbf{x}}^* \quad \text{and} \quad \mathbf{G}_z \in \mathcal{R}_{z,h}^* \tag{5.3.14}$$

With that convention, the correction term $V_{\mathrm{int},d}$ is approached by a $\mathcal{R}_{z,h}$ periodic function, which is a reasonable approximation as long as $h \gg 1$.

5.3.4 Discretization of BM in a plane-wave basis

For all $\mathbf{k} \in \mathbb{R}^2$, the fibers of the BM Hamiltonian are the operators

$$\boxed{\mathrm{H}^{\mathrm{BM}}_{\theta,\mathbf{k}} = \begin{pmatrix} v_{\mathrm{F}} \boldsymbol{\sigma}_{-\theta/2} \cdot (-i\nabla + \mathbf{k} - \mathbf{K}_1) & e^{i(\mathbf{K}_1 - \mathbf{K}_2)\cdot \mathbf{x}} \varepsilon_\theta^{-1} \mathbf{V}(\mathbf{x}) \\ e^{-i(\mathbf{K}_1 - \mathbf{K}_2)\cdot \mathbf{x}} \varepsilon_\theta^{-1} \mathbf{V}(\mathbf{x})^* & v_{\mathrm{F}} \boldsymbol{\sigma}_{\theta/2} \cdot (-i\nabla + \mathbf{k} - \mathbf{K}_2) \end{pmatrix}} \tag{5.3.15}$$

acting on $L^2_{\mathrm{per}}(\Omega_{\mathrm{M}}; \mathbb{C}^4)$. Let $\alpha_{\mathbf{k}}$ be in $L^2_{\mathrm{per}}(\Omega_{\mathrm{M}}; \mathbb{C}^4)$ and let $\alpha_{j,\mathbf{k}} \in L^2_{\mathrm{per}}(\Omega_{\mathrm{M}}; \mathbb{C})$ such that for all $\mathbf{x} \in \mathbb{R}^2$

$$\alpha_{\mathbf{k}}(X) := [\alpha_{j,\mathbf{k}}(\mathbf{x})]_{1 \leqslant j \leqslant 4}.$$

Since $\alpha_{\mathbf{k}}$ is moiré-periodic, each of its component can be expressed as a Fourier series on the moiré reciprocal lattice $\mathcal{R}_{\mathrm{M}}^*$

$$\alpha_{j,\mathbf{k}}(\mathbf{x}) = \frac{1}{\sqrt{|\Omega_{\mathrm{M}}|}} \sum_{\mathbf{G} \in \mathcal{R}_{\mathrm{M}}^*} \alpha_{j,\mathbf{k}}[\mathbf{G}] e^{i\mathbf{G}\cdot\mathbf{x}}.$$

We can then compute the action of $\mathrm{H}^{\mathrm{BM}}_{\theta,\mathbf{k}}$ on the individual Fourier modes. In order to simplify the computations, we can split the fiber $\mathrm{H}^{\mathrm{BM}}_{\theta,\mathbf{k}}$ as the sum of a diagonal kinetic term and an off-diagonal potential term

$$\mathrm{H}^{\mathrm{BM}}_{\theta,\mathbf{k}} =: \begin{pmatrix} \mathbf{T}_{\mathrm{A},\mathbf{k},\theta} & \mathbf{V}_{\mathrm{AB},\theta}(\mathbf{x}) \\ \mathbf{V}_{\mathrm{AB},\theta}(\mathbf{x})^* & \mathbf{T}_{\mathrm{B},\mathbf{k},\theta} \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{T}_{\mathrm{A},\mathbf{k},\theta} & 0 \\ 0 & \mathbf{T}_{\mathrm{B},\mathbf{k},\theta} \end{pmatrix}}_{=: \mathbf{T}^{\mathrm{BM}}_{\mathbf{k},\theta}} + \underbrace{\begin{pmatrix} 0 & \mathbf{V}_{\mathrm{AB},\theta}(\mathbf{x}) \\ \mathbf{V}_{\mathrm{AB},\theta}(\mathbf{x})^* & 0 \end{pmatrix}}_{=: \mathbf{V}^{\mathrm{BM}}_{\theta}(\mathbf{x})}.$$

5.3.4.1 Kinetic term $\mathbf{T}^{\mathrm{BM}}_{\mathbf{k},\theta}$

Let us focus on the top layer kinetic term $\mathbf{T}_{A,\mathbf{k},\theta}$. By linearity

$$\mathbf{T}_{\mathrm{A},\mathbf{k}} \begin{pmatrix} \alpha_{1,\mathbf{k}}(\mathbf{x}) \\ \alpha_{2,\mathbf{k}}(\mathbf{x}) \end{pmatrix} = \frac{1}{\sqrt{|\Omega_{\mathrm{M}}|}} \sum_{\mathbf{G} \in \mathcal{R}^*_{\mathrm{M},h}} \mathbf{T}_{\mathrm{A},\mathbf{k}} \begin{pmatrix} \alpha_{1,\mathbf{k}}[\mathbf{G}] e^{i\mathbf{G}\cdot\mathbf{x}} \\ \alpha_{2,\mathbf{k}}[\mathbf{G}] e^{i\mathbf{G}\cdot\mathbf{x}} \end{pmatrix}.$$

Now, it follows from Equation 5.3.4 that for all $\mathbf{G} \in \mathcal{R}_{\mathrm{M}}^*$

$$\mathbf{T}_{\mathrm{A},\mathbf{k}} \begin{pmatrix} \alpha_{1,\mathbf{k}}[\mathbf{G}] e^{i\mathbf{G}\cdot\mathbf{x}} \\ \alpha_{2,\mathbf{k}}[\mathbf{G}] e^{i\mathbf{G}\cdot\mathbf{x}} \end{pmatrix} = v_{\mathrm{F}} \boldsymbol{\sigma}_{-\theta/2} \cdot (-i\nabla + \mathbf{k} - \mathbf{K}_1) \begin{pmatrix} \alpha_{1,\mathbf{k}}[\mathbf{G}] e^{i\mathbf{G}\cdot\mathbf{x}} \\ \alpha_{2,\mathbf{k}}[\mathbf{G}] e^{i\mathbf{G}\cdot\mathbf{x}} \end{pmatrix} = \begin{pmatrix} \beta_{1,\mathbf{k}}[\mathbf{G}] e^{i\mathbf{G}\cdot\mathbf{x}} \\ \beta_{2,\mathbf{k}}[\mathbf{G}] e^{i\mathbf{G}\cdot\mathbf{x}} \end{pmatrix},$$

where the β coefficients are given by

$$\begin{pmatrix} \beta_{1,\mathbf{k}}[\mathbf{G}] \\ \beta_{2,\mathbf{k}}[\mathbf{G}] \end{pmatrix} := v_{\mathrm{F}} \boldsymbol{\sigma}_{-\theta/2} \cdot (\mathbf{G} + \mathbf{k} - \mathbf{K}_1) \begin{pmatrix} \alpha_{1,\mathbf{k}}[\mathbf{G}] \\ \alpha_{2,\mathbf{k}}[\mathbf{G}] \end{pmatrix}.$$

The same holds *mutatis mutandis* for the bottom layer. Using the above expression, we identify the Fourier component of the vector $\mathbf{T}^{\mathrm{BM}}_{\mathbf{k},\theta}\alpha_k$ on the reciprocal lattice vector $\mathbf{G} \in \mathcal{R}^*_{\mathrm{M}}$

$$\boxed{(\mathbf{T}^{\mathrm{BM}}_{\mathbf{k},\theta}\alpha_k)[\mathbf{G}] = \begin{pmatrix} v_{\mathrm{F}}\boldsymbol{\sigma}_{-\theta/2}\cdot(\mathbf{G}+\mathbf{k}-\mathbf{K}_1)\begin{pmatrix}\alpha_{1,\mathbf{k}}[\mathbf{G}]\\ \alpha_{2,\mathbf{k}}[\mathbf{G}]\end{pmatrix} \\ v_{\mathrm{F}}\boldsymbol{\sigma}_{\theta/2}\cdot(\mathbf{G}+\mathbf{k}-\mathbf{K}_2)\begin{pmatrix}\alpha_{3,\mathbf{k}}[\mathbf{G}]\\ \alpha_{4,\mathbf{k}}[\mathbf{G}]\end{pmatrix}\end{pmatrix}.} \tag{5.3.16}$$

We also deduce the expression of the $\mathbf{G},\mathbf{G}'$ element of the Fourier matrix of $\mathbf{T}^{\mathrm{BM}}_{\mathbf{k},\theta}$

$$\boxed{[\mathbf{T}^{\mathrm{BM}}_{\mathbf{k},\theta}]_{\mathbf{G},\mathbf{G}'} = \delta_{\mathbf{G}\mathbf{G}'}\begin{pmatrix} v_{\mathrm{F}}\boldsymbol{\sigma}_{-\theta/2}\cdot(\mathbf{G}+\mathbf{k}-\mathbf{K}_1) & 0 \\ 0 & v_{\mathrm{F}}\boldsymbol{\sigma}_{\theta/2}\cdot(\mathbf{G}+\mathbf{k}-\mathbf{K}_2)\end{pmatrix}.} \tag{5.3.17}$$

One can use the respective expressions (5.3.16) and (5.3.17) to implement the BM kinetic term in a matrix-free or a full matrix fashion. Let us add that since the matrix in equation (5.3.17) is block diagonal, its spectral decomposition can be computed at low computational cost, with 2×2 diagonal blocks, making it a good candidate for diagonal preconditioning.

5.3.4.2 Potential term $\mathbf{V}^{\mathrm{BM}}_\theta$

A straightforward way to compute the action of $\hat{\mathbf{V}}^{\mathrm{BM}}_\theta$ in a matrix free fashion, is to compute the multiplication by $\hat{\mathbf{V}}^{\mathrm{BM}}_\theta$ in real space by the mean of discrete Fourier and inverse Fourier transforms. However, in the present case, it proves advantageous to directly compute the matrix of $\hat{\mathbf{V}}^{\mathrm{BM}}_\theta$ in Fourier space. By using the relation

$$\mathbf{K}_1 - \mathbf{K}_2 = -\mathbf{q}_1,$$

we compute for all $\mathbf{G},\mathbf{G}' \in \mathcal{R}^*_{\mathrm{M}}$

$$\langle e_{\mathbf{G}}, \mathbf{V}_{\mathrm{AB},\theta} e_{\mathbf{G}'}\rangle_{L^2(\mathbb{R}^2)} = \varepsilon_\theta^{-1}\sum_{j=1}^{3} V_j \delta_{(\mathbf{G}'-\mathbf{G}),(\mathbf{q}_j-\mathbf{q}_1)},$$

where we recall that the scalar product acts on each individual component of $\mathbf{V}_{\mathrm{AB},\theta}$. The Fourier matrix of the BM potential is given for all $\mathbf{G},\mathbf{G}' \in \mathcal{R}^*_{\mathrm{M}}$ by

$$[\hat{\mathbf{V}}^{\mathrm{BM}}_\theta]_{\mathbf{G},\mathbf{G}'} = \begin{pmatrix} 0 & \langle e_{\mathbf{G}}, \mathbf{V}_{\mathrm{AB},\theta} e_{\mathbf{G}'}\rangle_{L^2(\mathbb{R}^2)} \\ \left\langle e_{\mathbf{G}}, \mathbf{V}^*_{\mathrm{AB},\theta} e_{\mathbf{G}'}\right\rangle_{L^2(\mathbb{R}^2)} & 0\end{pmatrix}$$

which ultimately writes

$$\boxed{[\hat{\mathbf{V}}^{\mathrm{BM}}_\theta]'_{\mathbf{G},\mathbf{G}} = \varepsilon_\theta^{-1}\sum_{j=1}^{3}\begin{pmatrix} 0 & V_j\delta_{(\mathbf{G}'-\mathbf{G}),(\mathbf{q}_j-\mathbf{q}_1)} \\ V_j^*\delta_{(\mathbf{G}'-\mathbf{G}),(\mathbf{q}_1-\mathbf{q}_j)} & 0\end{pmatrix}.} \tag{5.3.18}$$

Let us point out the change of sign in the two Kronecker deltas in the above expression.

5.3.5 Discretization of the CGG Hamiltonian in a plane-wave basis

For clarity, we denote by $\mathbf{T}^{(-1)}_{d,\theta}$, $\mathbf{T}^{(0)}_{d,\theta}$ and $\mathbf{T}^{(1)}_{d,\theta}$ the CGG Hamiltonian terms, of respective order $-1, 0$ and 1 in ε_θ. We apply the same discretization procedure as for BM to the three CGG terms independently.

5.3.5.1 Discretization of the CGG potential $\mathbf{T}_{d,\theta}^{(-1)}$

As in the BM case, the term $\mathbf{T}_{d,\theta}^{(-1)}$ can simply be applied in real space, by means of discrete Fourier and inverse Fourier transforms, in matrix-free fashion. Otherwise, one needs to compute the matrix elements of $\mathbf{T}_{d,\theta}^{(-1)}$ in the moiré Fourier basis, which depends on three basic elements: the $\mathbf{K}$ Dirac point symmetry adapted basis (Φ_1, Φ_2), the discrete bilayer scalar products $\langle\!\langle \cdot, \cdot \rangle\!\rangle_d^{\eta\eta'}$ and the Kohn-Sham potential correction $V_{\mathrm{int},d}$.

Monolayer Dirac points natural basis

The goal of this section is compute the symmetry adapted basis (Φ_1, Φ_2) associated to the degenerate valence π-band of monolayer graphene at Dirac point $\mathbf{K}$ or $\mathbf{K}'$, as defined in Section 5.2.1. We recall that

$$\Phi_1 \in \mathrm{Ker}\,(\mathcal{R}_{\frac{2\pi}{3}} - \omega) \quad \text{and} \quad \Phi_2 \in \mathrm{Ker}\,(\mathcal{R}_{\frac{2\pi}{3}} - \omega^2). \tag{5.3.19}$$

Numerically, the diagonalization of the monolayer graphene Kohn-Sham DFT Hamiltonian at Dirac point $\mathbf{K}$ provides a basis

$$\varphi_a, \varphi_b \in \mathrm{Ker}\,\left(\mathcal{R}_{\frac{2\pi}{3}} - \omega\right) + \mathrm{Ker}\,\left(\mathcal{R}_{\frac{2\pi}{3}} - \omega^2\right). \tag{5.3.20}$$

Let u_1, u_2, u_a and u_b be such that

$$\begin{aligned} \Phi_1(\mathbf{r}) &=: e^{i\mathbf{K}\cdot\mathbf{r}} u_1(\mathbf{r}), \qquad & \Phi_2(\mathbf{r}) &=: e^{i\mathbf{K}\cdot\mathbf{r}} u_2(\mathbf{r}), \\ \varphi_a(\mathbf{r}) &=: e^{i\mathbf{K}\cdot\mathbf{r}} u_a(\mathbf{r}), \qquad & \varphi_b(\mathbf{r}) &=: e^{i\mathbf{K}\cdot\mathbf{r}} u_b(\mathbf{r}). \end{aligned}$$

Note that the basis (Φ_1, Φ_2) can be computed only with u_1 since $u_2 = \mathscr{C}\mathscr{P}(u_1)$. We remark that by equation (5.3.19), for all $j = 1, 2$

$$\begin{aligned} (\mathcal{R}_{\frac{2\pi}{3}}\Phi_j)(\mathbf{r}) = \omega^j \Phi_j(\mathbf{r}) &\iff e^{i\mathcal{R}_{\frac{2\pi}{3}}\mathbf{K}\cdot\mathbf{r}} (\mathcal{R}_{\frac{2\pi}{3}} u_j)(\mathbf{r}) = \omega^j e^{i\mathbf{K}\cdot\mathbf{r}} u_j(\mathbf{r}) \\ &\iff e^{i(\mathbf{K}+\mathbf{G}_s)\cdot\mathbf{r}} (\mathcal{R}_{\frac{2\pi}{3}} u_j)(\mathbf{r}) = \omega^j e^{i\mathbf{K}\cdot\mathbf{r}} u_j(\mathbf{r}) \\ &\iff e^{i\mathbf{G}_s\cdot\mathbf{r}} (\mathcal{R}_{\frac{2\pi}{3}} u_j)(\mathbf{r}) = \omega^j u_j(\mathbf{r}), \end{aligned}$$

where $\mathbf{G}_s$ is the reciprocal lattice vector such that $\mathcal{R}_{\frac{2\pi}{3}}\mathbf{K} = \mathbf{K} + \mathbf{G}_s$. Second, by Equation 5.3.20, there exist complex numbers c_a^1 and c_a^2 such that

$$u_a = c_a^1 u_1 + c_a^2 u_2.$$

Combining these equalities we obtain

$$e^{i\mathbf{G}_s\cdot\mathbf{r}} (\mathcal{R}_{\frac{2\pi}{3}} u_a) - \omega^2 u_a = c_a^1 (e^{i\mathbf{G}_s\cdot\mathbf{r}} \mathcal{R}_{\frac{2\pi}{3}} - \omega^2) u_1,$$

so that

$$u_1 \in \mathrm{Span}\,\left(e^{i\mathbf{G}_s\cdot\mathbf{r}} (\mathcal{R}_{\frac{2\pi}{3}} u_a) - \omega^2 u_a\right).$$

Let us set

$$\tilde{u} := \frac{e^{i\mathbf{G}_s\cdot\mathbf{r}} (\mathcal{R}_{\frac{2\pi}{3}} u_a) - \omega^2 u_a}{\left\| e^{i\mathbf{G}_s\cdot\mathbf{r}} (\mathcal{R}_{\frac{2\pi}{3}} u_a) - \omega^2 u_a \right\|_{L^2_{\mathrm{per}}(\Omega)}}.$$

Since u_1 is on the unit sphere, there exists α such that $u_1 = e^{i\alpha}\tilde{u}$. Then

$$u_2 = \mathscr{C}\mathscr{P}(u_1) = e^{-i\alpha}\mathscr{C}\mathscr{P}(\tilde{u}). \tag{5.3.21}$$

It only remains to choose α such that the Fermi velocity $v_{\mathrm{F}} := \langle \Phi_1, (-i\partial_{x_1})\Phi_2 \rangle_{L^2}$ is a real number. Let $\left\langle e^{i\mathbf{K}\cdot\mathbf{r}}\tilde{u}, (-i\partial_{x_1}) e^{i\mathbf{K}\cdot\mathbf{r}} \mathscr{C}\mathscr{P}(\tilde{u}) \right\rangle = a + ib$. Then

$$v_{\mathrm{F}} = e^{-i2\alpha}(a + ib).$$

Hence

$$\mathrm{Im}\,(v_{\mathrm{F}}) = 0 \iff \cos(2\alpha)b - \sin(2\alpha)a = 0 \iff \alpha = \frac{1}{2}\arctan\left(\frac{b}{a}\right).$$

Algorithm 4: Computing Dirac point natural basis

Given: any entry basis (u_a, u_b)

1. Compute $\tilde{u} := \dfrac{e^{i\mathbf{G}_s\cdot\mathbf{r}}(\mathcal{R}_{\frac{2\pi}{3}}u_a) - \omega^2 u_a}{\left\| e^{i\mathbf{G}_s\cdot\mathbf{r}}(\mathcal{R}_{\frac{2\pi}{3}}u_a) - \omega^2 u_a \right\|_{L^2_{\mathrm{per}}}}$
2. Compute $\left\langle e^{i\mathbf{K}\cdot\mathbf{r}}\tilde{u}, (-i\partial_{x_1})e^{i\mathbf{K}\cdot\mathbf{r}}\mathscr{C}\mathscr{P}(\tilde{u})\right\rangle_{L^2_{\mathrm{per}}} = a + ib$
3. Set $\alpha = \dfrac{1}{2}\arctan\left(\dfrac{b}{a}\right)$
4. Set $u_1 = e^{i\alpha}\tilde{u}$ and $u_2 = e^{-i\alpha}\mathscr{C}\mathscr{P}(\tilde{u})$

The global procedure to compute (Φ_1, Φ_2) is summarized in algorithm 4.

Let $\mathscr{O}$ be a unitary operator such that $\mathcal{O}(\mathcal{R}_h^*) \subset \mathcal{R}_h^*$. In algorithm 4, the action of $\mathscr{O}$ on u_a (resp. u_b) is computed using the Fourier decomposition of u_a (resp. u_b) in the truncated lattice $\mathcal{R}_h^*$ (see. (5.3.13)). Since

$$\mathscr{O}u_a(x) = \frac{1}{\sqrt{\Omega_h}}\sum_{\mathbf{G}\in\mathcal{R}_h^*} u_a[\mathbf{G}]e^{\langle \mathscr{O}^* x, \mathbf{G}\rangle} = \frac{1}{\sqrt{\Omega_h}}\sum_{\mathbf{G}\in\mathcal{R}_h^*} u_a[\mathbf{G}]e^{\langle x, \mathscr{O}\mathbf{G}\rangle} = \frac{1}{\sqrt{\Omega_h}}\sum_{\mathbf{G}\in\mathcal{R}_h^*} u_a[\mathcal{O}^*\mathbf{G}]e^{\langle x, \mathbf{G}\rangle}.$$

we find for all $\mathbf{G}\in\mathcal{R}_h^*$

$$(\mathscr{O}u_a)[\mathbf{G}] = u_a[\mathscr{O}^*\mathbf{G}]. \tag{5.3.22}$$

Bilayer potential correction

We now turn to the numerical evaluation of $V_{\mathrm{int},d}$. For all $d > 0$, $z \in [0, h)$ and for a given disregistry $\mathbf{y}\in\Omega$, we have

$$\begin{aligned}
V_{\mathrm{int},d,\mathbf{y}}(z) &= \frac{1}{|\Omega|}\int_\Omega V^{(2)}_{d,\mathbf{y}}(\mathbf{x}, z) - V(\mathbf{x}, z+\tfrac{d}{2}) - V(\mathbf{x}-\mathbf{y}, z-\tfrac{d}{2})\,\mathrm{d}\mathbf{x} \\
&= \frac{1}{|\Omega|}\int_\Omega V^{(2)}(\mathbf{x}, z) - V(\mathbf{x}, z+\tfrac{d}{2}) - V(\mathbf{x}, z-\tfrac{d}{2})\mathrm{d}\mathbf{x} \\
&\overset{\text{discretization}}{\simeq} \frac{1}{|\Omega|^{3/2}\sqrt{h}}\sum_{\mathbf{G}\in\mathcal{R}_h^*}\left[V^{(2)}[\mathbf{G}] - V[\mathbf{G}]\left(e^{i\mathbf{G}_z\frac{d}{2}} + e^{-i\mathbf{G}_z\frac{d}{2}}\right)\right]\int_\Omega e^{i\mathbf{G}\cdot[\mathbf{x},z]}\mathrm{d}\mathbf{x} \\
&= \frac{1}{\sqrt{|\Omega|h}}\sum_{\mathbf{G}_z\in\mathcal{R}^*_{h,z}}\left[V^{(2)}[\mathbf{0},\mathbf{G}_z] - 2V[\mathbf{0},\mathbf{G}_z]\cos\big(\mathbf{G}_z\tfrac{d}{2}\big)\right]e^{i\mathbf{G}_z z}.
\end{aligned}$$

The second equality follows from the $\mathcal{R}_{\mathbf{x}}$-periodicity of V in the $\mathbf{x}$ direction. The approximation at the third line is reasonable as long as $h \gg 1$. We deduce from the last equality that $V_{\mathrm{int},d,,\mathbf{y}}$ can be approximated in Fourier space, for all $\mathbf{G}_z\in\mathcal{R}^*_{h,z}$, by

$$\boxed{V_{\mathrm{int},d,\mathbf{y}}[\mathbf{G}_z] = \frac{1}{\sqrt{|\Omega|}}\left[V^{(2)}[\mathbf{0},\mathbf{G}_z] - 2V[\mathbf{0},\mathbf{G}_z]\cos\big(\mathbf{G}_z\tfrac{d}{2}\big)\right].} \tag{5.3.23}$$

The total potential correction is then simply approached by a Riemann sum over a uniform sample of disregistries $\mathbf{y}$ (which is the optimal quadrature scheme due to the periodicity and smoothness of the potentials).

Bilayer scalar product

Let $f, g \in L^2_{\rm per}(\Omega \times [0,h); \mathbb{C})$. For all signs $\eta, \eta' \in \{-1,+1\}$ and vector $\mathbf{X} \in \mathbb{R}^2$

$$
\begin{aligned}
(\!(f,g)\!)_d^{\eta,\eta'}(\mathbf{X}) &:= \int_{\Omega\times\mathbb{R}} \overline{f(\mathbf{x}-\tfrac{\eta}{2}J\mathbf{X}, z-\eta\tfrac{d}{2})} g(\mathbf{x}-\tfrac{\eta'}{2}J\mathbf{X}, z-\eta'\tfrac{d}{2}) \mathrm{d}\mathbf{x}\mathrm{d}z \\
&\overset{\text{discretization}}{\simeq} \frac{1}{|\Omega|h} \sum_{\mathbf{G},\mathbf{G}'\in\mathcal{R}_h^*} \overline{f[\mathbf{G}]} g[\mathbf{G}'] e^{i\frac{\eta}{2}\mathbf{G}\cdot(J\mathbf{X},d)} e^{-i\frac{\eta'}{2}\mathbf{G}'\cdot(J\mathbf{X},d)} \underbrace{\int_{\Omega\times[0,h)} e^{i(\mathbf{G}'-\mathbf{G})\cdot(\mathbf{x},z)} \mathrm{d}\mathbf{x}\mathrm{d}z}_{=\delta_{\mathbf{G}\mathbf{G}'}|\Omega|h} \\
&= \sum_{\mathbf{G}\in\mathcal{R}_h^*} \overline{f[\mathbf{G}]} g[\mathbf{G}] e^{i\frac{1}{2}(\eta-\eta')\mathbf{G}\cdot(J\mathbf{X},d)} \\
&\overset{\mathrm{tr}\{J\}=-J}{=} \sum_{\mathbf{G}_\mathbf{x}\in\mathcal{R}_\mathbf{x}^*} \left(\sum_{\mathbf{G}_z\in\mathcal{R}_{z,h}^*} \overline{f[\mathbf{G}_\mathbf{x},\mathbf{G}_z]} g[\mathbf{G}_\mathbf{x},\mathbf{G}_z] e^{i\frac{1}{2}(\eta-\eta')\mathbf{G}_z d} \right) e^{i\frac{1}{2}(\eta'-\eta)J\mathbf{G}_\mathbf{x}\cdot\mathbf{X}} \\
&= \frac{1}{\sqrt{|J\Omega|}} \sum_{\mathbf{G}_\mathbf{x}\in\mathcal{R}_\mathbf{x}^*} \left(\sqrt{|J\Omega|} \sum_{\mathbf{G}_z\in\mathcal{R}_{z,h}^*} \overline{f[\mathbf{G}_\mathbf{x},\mathbf{G}_z]} g[\mathbf{G}_\mathbf{x},\mathbf{G}_z] e^{i\frac{1}{2}(\eta-\eta')\mathbf{G}_z d} \right) e^{i\frac{1}{2}(\eta'-\eta)J\mathbf{G}_\mathbf{x}\cdot\mathbf{X}}.
\end{aligned}
$$

Hence if

$$
C_d^{\eta,\eta'}(\mathbf{G}_\mathbf{x}) := \sqrt{|J\Omega|} \sum_{\mathbf{G}_z\in\mathcal{R}_{z,h}^*} \overline{f[\mathbf{G}_\mathbf{x},\mathbf{G}_z]} g[\mathbf{G}_\mathbf{x},\mathbf{G}_z] e^{i\frac{1}{2}(\eta-\eta')\mathbf{G}_z d},
$$

then

$$
(\!(f,g)\!)_d^{\eta,\eta'}(\mathbf{X}) = \frac{1}{\sqrt{|J\Omega|}} \sum_{\mathbf{G}_\mathbf{x}\in\mathcal{R}_x^*} \left(\delta_{\eta=\eta'} \sqrt{|J\Omega|} \langle f,g\rangle_{L^2_{\rm per}} + \delta_{\eta\neq\eta'} C_d^{\eta,\eta'}(\mathbf{G}_\mathbf{x}) \right) e^{i\frac{1}{2}(\eta'-\eta)J\mathbf{G}_x\cdot\mathbf{X}}.
$$

We deduce that in Fourier space for all $\mathbf{G} \in \mathcal{R}_{\rm M}^*$ or equivalently $J\mathbf{G}_\mathbf{x} \in \mathcal{R}_\mathbf{x}^*$

$$
\widehat{[(\!(f,g)\!)_d^{\eta,\eta'}]}(J\mathbf{G}_\mathbf{x}) = \begin{cases} \delta_{\mathbf{G}_\mathbf{x}=0}\sqrt{|J\Omega|}\langle f,g\rangle_{L^2_{\rm per}} & \text{for } \eta=\eta', \\ C_d^{\eta,\eta'}(\mathbf{G}_\mathbf{x}) & \text{for } \eta=-1,\ \eta'=1,, \\ C_d^{\eta,\eta'}(-\mathbf{G}_\mathbf{x}) & \text{for } \eta'=-1,\ \eta=1 \end{cases}
$$

which can be written in the simpler form

$$
\boxed{\widehat{[(\!(f,g)\!)_d^{\eta,\eta'}]}(J\mathbf{G}_\mathbf{x}) = \delta_{\eta=\eta'}\delta_{\mathbf{G}_x=0}\sqrt{|J\Omega|}\langle f,g\rangle_{L^2_{\rm per}} + \delta_{\eta\neq\eta'} C_d^{\eta,\eta'}(\eta'\mathbf{G}_\mathbf{x})} \tag{5.3.24}
$$

Potential correction terms

Most part of the effective local potential boils down to compute a bilayer scalar product, as described in equation (5.3.5.1). The only remaining term is the second term in the right-hand side of

$$
[\widetilde{\mathbb{W}}_d^\eta(\mathbf{X})]_{jj'} := (\!(u_j\overline{u_{j'}}, V)\!)_d^{\eta(-\eta)}(\mathbf{X}) + (W_{{\rm int},d}^\eta)_{jj'}, \tag{5.3.25}
$$

involving the potential correction $V_{{\rm int},d}$, for all $j,j' \in \{1,2\}$, sign $\eta \in \{-,+\}$ and $\mathbf{X} \in \mathbb{R}^2$. One has:

$$
\begin{aligned}
(W^{\eta}_{\mathrm{int},d})_{jj'} &= \int_{\Omega\times\mathbb{R}} \overline{u_j} u_{j'}(\mathbf{x}, z-\eta\tfrac{d}{2}) V_{\mathrm{int},d}(z)\mathbf{x}\mathrm{d}z \\
&\overset{\text{discretization}}{\simeq} \frac{1}{\sqrt{|\Omega|}\times h} \sum_{\mathbf{G}\in\mathcal{R}_h^*} \sum_{\mathbf{G}'_z\in\mathcal{R}^*_{z,h}} \overline{u_j} u_{j'}[\mathbf{G}] V_{\mathrm{int},d}[\mathbf{G}'_z] e^{-i\mathbf{G}_z\eta\frac{d}{2}} \underbrace{\int_{\Omega\times[0,h)} e^{i\mathbf{G}_{\mathbf{x}}\cdot\mathbf{X}} e^{i(\mathbf{G}_z+\mathbf{G}'_z)\cdot z}\mathbf{x}\mathrm{d}z}_{\delta_{\mathbf{G}_{\mathbf{x}}=0}|\Omega|\times\delta_{(\mathbf{G}_z=-\mathbf{G}'_z)}h} \\
&= \sqrt{|\Omega|} \sum_{\mathbf{G}_z\in\mathcal{R}^*_{z,h}} \overline{u_j} u_{j'}[\mathbf{0},\mathbf{G}_z] V_{\mathrm{int},d}[-\mathbf{G}_z] e^{-i\mathbf{G}_z\eta\frac{d}{2}} \\
&= \sqrt{|\Omega|} \sum_{\mathbf{G}_z\in\mathcal{R}^*_{z,h}} \overline{u_j} u_{j'}[\mathbf{0},\mathbf{G}_z] \overline{V_{\mathrm{int},d}[\mathbf{G}_z]} e^{-i\mathbf{G}_z\eta\frac{d}{2}}.
\end{aligned}
$$

Hence we have for all $J\mathbf{G}_{\mathbf{x}} \in \mathcal{R}^*_{\mathbf{x}}$

$$
\boxed{(\widehat{W}^{\eta}_{\mathrm{int},d})_{jj'}(J\mathbf{G}_{\mathbf{x}}) = \delta_{\mathbf{G}_{\mathbf{x}}=0} \sum_{\mathbf{G}_z\in\mathcal{R}^*_{z,h}} |\Omega|\, \overline{u_j} u_{j'}[\mathbf{0},\mathbf{G}_z] \overline{V_{\mathrm{int},d}[\mathbf{G}_z]} e^{-i\mathbf{G}_z\eta\frac{d}{2}}.} \tag{5.3.26}
$$

In practice the table of Fourier coefficients of $\overline{u_j}u_{j'}$ is obtained by multiplying both functions in real space and going back to frequencies space with an inverse fast Fourier transform.

Full CGG potential term

From the definitions of the CGG potential terms, using the equations (5.3.23), (5.3.24), (5.3.26) as well as algorithm 4, we can compute the moiré Fourier coefficients $\widetilde{\mathbb{W}}^+_d[\mathbf{G}]$, $\widetilde{\mathbb{W}}^-_d[\mathbf{G}]$, $\widetilde{\mathbb{V}}_d[\mathbf{G}]$ and $\widetilde{\mathbb{V}}^*_d[\mathbf{G}]$, for all $\mathbf{G}\in\mathcal{R}^*_{\mathrm{M}}$. The Fourier matrix of $\mathbf{T}^{(-1)}_{d,\theta}$ is then obtained for all $\mathbf{G},\mathbf{G}'\in\mathcal{R}^*_{\mathrm{M}}$ as

$$
\boxed{[\mathbf{T}^{(-1)}_{d,\theta}]_{\mathbf{G},\mathbf{G}'} = \frac{1}{\varepsilon_\theta\sqrt{|\Omega_{\mathrm{M}}}|} \begin{pmatrix} \widetilde{\mathbb{W}}^+_d[\mathbf{G}-\mathbf{G}'] & \widetilde{\mathbb{V}}_d[\mathbf{G}-\mathbf{G}'] \\ \widetilde{\mathbb{V}}^*_d[\mathbf{G}-\mathbf{G}'] & \widetilde{\mathbb{W}}^-_d[\mathbf{G}-\mathbf{G}'] \end{pmatrix}} \tag{5.3.27}
$$

5.3.5.2 Kinetic/Magnetic–like CGG terms $\mathbf{T}^{(0)}_{d,\theta}$

Matrix-free implementation

As seen in (5.3.9), the diagonal part of $\mathbf{T}^{(0)}_{d,\theta}$ is the same as the kinetic term of the BM Hamiltonian. It thus discretizes exactly as in (5.3.17). For the off-diagonal part, let $n\in\{1,2\}$. We abbreviate

$$
M_{\mathbf{K}_n}(\mathbf{x}) := J\left(-i\nabla - \mathbf{q}_1\right)\widetilde{\Sigma}_d(\mathbf{x})\cdot\left(-i\nabla-\mathbf{K}_n\right).
$$

If $\mathbf{x}_j$ denotes the j-th component of a given vector $\mathbf{x}$, one has

$$
\begin{aligned}
M_{\mathbf{K}_n}(\mathbf{x}) &= J\left[(-i\partial_1 - \mathbf{q}_{1,1})\widetilde{\Sigma}_d(\mathbf{x}), (-i\partial_2-\mathbf{q}_{1,2})\widetilde{\Sigma}_d(\mathbf{x})\right]\cdot\begin{pmatrix} -i\partial_1-\mathbf{K}_{n,1} \\ -i\partial_2-\mathbf{K}_{n,2}\end{pmatrix} \\
&= J\Xi(\mathbf{x})\cdot\begin{pmatrix} -i\partial_1-\mathbf{K}_{n,1} \\ -i\partial_2-\mathbf{K}_{n,2}\end{pmatrix},
\end{aligned} \tag{5.3.28}
$$

where we introduced $\Xi(\mathbf{x}) := \left[(-i\partial_1-\mathbf{q}_{1,1})\widetilde{\Sigma}_d(\mathbf{x}), (-i\partial_2-\mathbf{q}_{1,2})\widetilde{\Sigma}_d(\mathbf{x})\right] \in (\mathbb{C}^{2\times 2})^2$.

Let us first derive the matrix free implementation of the off-diagonal term. From (5.3.29), we obtain for all $\mathbf{k}$ in the moiré Brillouin zone and all vector $\begin{pmatrix} \alpha_{1,\mathbf{k}}(\mathbf{x}) \\ \alpha_{2,\mathbf{k}}(\mathbf{x}) \end{pmatrix}$

$$\boxed{M_{\mathbf{K}_n}(\mathbf{x}) \begin{pmatrix} \alpha_{1,\mathbf{k}}(\mathbf{x}) \\ \alpha_{2,\mathbf{k}}(\mathbf{x}) \end{pmatrix} = \sum_{j=1,2} (J\Xi)_j(\mathbf{x}) \times \begin{pmatrix} (-i\partial_j - \mathbf{K}_{n,j})\alpha_{1,\mathbf{k}}(\mathbf{x}) \\ (-i\partial_j - \mathbf{K}_{n,j})\alpha_{2,\mathbf{k}}(\mathbf{x}) \end{pmatrix}.} \tag{5.3.29}$$

The two matrix-vector products on the right-hand side of (5.3.29) can be computed in real space. However, since both Ξ and α_j are moiré-periodic functions, they are readily computed in Fourier space for all $\mathbf{G} \in \mathcal{R}^*_{\mathrm{M},h}$ with

$$(J\Xi)_j[\mathbf{G}] = J[(\mathbf{G}_j - \mathbf{q}_{1,j})\widetilde{\Sigma}_d[\mathbf{G}], (\mathbf{G}_j - \mathbf{q}_{1,j})\widetilde{\Sigma}_d[\mathbf{G}]] = [(\mathbf{G}_2 - \mathbf{q}_{1,2})\widetilde{\Sigma}_d[\mathbf{G}], -(\mathbf{G}_1 - \mathbf{q}_{1,1})\widetilde{\Sigma}_d[\mathbf{G}]] \tag{5.3.30}$$

and

$$\mathcal{F}\left((-i\partial_j - \mathbf{K}_{n,j})\alpha_{j,\mathbf{k}}\right)[\mathbf{G}] = (\mathbf{G} + \mathbf{k}_j - \mathbf{K}_{n,j})\alpha_{j,\mathbf{k}}[\mathbf{G}]. \tag{5.3.31}$$

Note that in (5.3.30), the matrix J acts on the components of the vector Ξ. We then recovered the vectors in real space by an inverse fast Fourier transform.

Full matrix implementation

Let us now derive the full matrix term $M_{\mathbf{K}_n}$ in the moiré Fourier basis. Again since Ξ is a (matrix-valued) moiré-periodic function, we introduce the Fourier decomposition for all $\mathbf{x} \in \mathbb{R}^2$ and $j = 1, 2$

$$\Xi_j(\mathbf{x}) = \frac{1}{\sqrt{|\Omega_{\mathrm{M}}|}} \sum_{\mathbf{G} \in \mathcal{R}^*_{\mathrm{M},L}} \Xi_j[\mathbf{G}] e^{i\mathbf{G}\cdot\mathbf{x}}.$$

A straightforward computation yields for all $\mathbf{G}, \mathbf{G}' \in \mathcal{R}^*_{\mathrm{M}}$

$$\langle e_{\mathbf{G}}, M_{\mathbf{K}_n}(\mathbf{x}) e_{\mathbf{G}'} \rangle_{L^2(\mathbb{R}^2,\mathbb{C})} = \frac{1}{\sqrt{|\Omega_{\mathrm{M}}|}} \sum_{j=1,2} (J\Xi)_j[\mathbf{G} - \mathbf{G}'](\mathbf{G}' + \mathbf{k}_j - \mathbf{K}_{n,j}).$$

In order to compute the sub-diagonal term $\langle e_{\mathbf{G}}, M_{\mathbf{K}_n}(\mathbf{x})^* e_{\mathbf{G}'} \rangle_{L^2(\mathbb{R}^2,\mathbb{C})}$, we need to compute with care the ajoint of $M_{\mathbf{K}_n}(\mathbf{x})^*$. For all $n \in \{1, 2\}$, we can show that

$$((J\Xi) \cdot (-i\nabla - K_n))^* = (J\Xi^*) \cdot (-i\nabla - K_n) = (J\Xi)^* \cdot (-i\nabla - K_n). \tag{5.3.32}$$

Therefore

$$\langle e_{\mathbf{G}}, M_{\mathbf{K}_n}(\mathbf{x})^* e_{\mathbf{G}'} \rangle_{L^2(\mathbb{R}^2,\mathbb{C})} = \frac{1}{\sqrt{|\Omega_{\mathrm{M}}|}} \sum_{j=1,2} (J\Xi^*)_j[\mathbf{G} - \mathbf{G}'] \times (\mathbf{G}' + \mathbf{k}_j - \mathbf{K}_{n,j}).$$

The full matrix of the magnetic term is given for all $\mathbf{G}, \mathbf{G}' \in \mathcal{R}^*_{\mathrm{M}}$ by

$$\boxed{\begin{pmatrix} 0 & \langle e_{\mathbf{G}}, M_{\mathbf{K}_n}(\mathbf{x}) e_{\mathbf{G}'} \rangle_{L^2(\mathbb{R}^2,\mathbb{C})} \\ \langle e_{\mathbf{G}}, M_{\mathbf{K}_n}(\mathbf{x})^* e_{\mathbf{G}'} \rangle_{L^2(\mathbb{R}^2,\mathbb{C})} & 0 \end{pmatrix} = \\ \frac{1}{\sqrt{|\Omega_{\mathrm{M}}|}} \sum_{j=1,2} \begin{pmatrix} 0 & (J\Xi)_j[\mathbf{G} - \mathbf{G}'](\mathbf{G}' + \mathbf{k}_j - \mathbf{K}_{2,j}) \\ (J\Xi^*)_j[\mathbf{G} - \mathbf{G}'] \times (\mathbf{G}' + \mathbf{k}_j - \mathbf{K}_{1,j}) & 0 \end{pmatrix}.}$$

5.3.5.3 Term with second derivatives $\mathbf{T}^{(1)}_{d,\theta}$

Diagonal term

The computations for the diagonal term are straightforward, using the same Fourier decomposition as the last sections. The action of the Laplacian term writes for all $\mathbf{k} \in \Omega^*$ and $n \in \{1, 2\}$.

$$-\Delta_{\mathbf{K}_n} \begin{pmatrix} \alpha_{1,\mathbf{k}}(\mathbf{x}) \\ \alpha_{2,\mathbf{k}}(\mathbf{x}) \end{pmatrix} = \frac{1}{\sqrt{|\Omega_{\mathrm{M}}|}} \sum_{\mathbf{G} \in \mathcal{R}^*_{\mathrm{M},L}} |\mathbf{G}_{\mathbf{k}} - \mathbf{K}_n|^2 \begin{pmatrix} \alpha_{1,\mathbf{k}}[\mathbf{G}] e^{i\mathbf{G}\cdot\mathbf{x}} \\ \alpha_{2,\mathbf{k}}[\mathbf{G}] e^{i\mathbf{G}\cdot\mathbf{x}} \end{pmatrix}.$$

Alternatively, the full Fourier matrix of $-\Delta_{\mathbf{K}_n}$ is block diagonal and writes for all $\mathbf{G}, \mathbf{G} \in \mathcal{R}^*_{\mathrm{M},L}$

$$[-\Delta_{\mathbf{K}_n}]_{\mathbf{G},\mathbf{G}} = \begin{pmatrix} |\mathbf{G}_{\mathbf{k}} - \mathbf{K}_n|^2 & 0 \\ 0 & |\mathbf{G}_{\mathbf{k}} - \mathbf{K}_n|^2 \end{pmatrix}. \tag{5.3.34}$$

Off-diagonal term

Let us consider the off-diagonal first order term

$$\mathbf{T}^{(2)}(\mathbf{x}) := \begin{pmatrix} 0 & \mathbf{T}^{(2)}_{\mathbf{K}_1,\mathbf{K}_2}(\mathbf{x}) \\ \left[\mathbf{T}^{(2)}_{\mathbf{K}_1,\mathbf{K}_2}\right]^*(\mathbf{x}) & 0 \end{pmatrix},$$

where

$$\mathbf{T}^{(2)}_{\mathbf{K}_1,\mathbf{K}_2}(\mathbf{x}) := (-i\nabla_{\mathbf{K}_1}) \cdot \left[\widetilde{\Sigma}_d(\mathbf{x})(-i\nabla_{\mathbf{K}_2})\bullet\right].$$

Then we have for all $\mathbf{G}, \mathbf{G}' \in \mathcal{R}^*_{\mathrm{M}}$

$$\begin{aligned} \left\langle e_{\mathbf{G}}, T^{(2)}_{\mathbf{K}_1,\mathbf{K}_2} e_{\mathbf{G}'} \right\rangle_{L^2(\mathbb{R}^2,\mathbb{C})} &= (\mathbf{G}+\mathbf{k}-\mathbf{K}_1)\cdot(\mathbf{G}'+\mathbf{k}-\mathbf{K}_2) + \left\langle e_{\mathbf{G}}, \widetilde{\Sigma}_d e_{\mathbf{G}'} \right\rangle_{L^2(\mathbb{R}^2,\mathbb{C})} \\ &= \frac{1}{\sqrt{|\Omega_{\mathrm{M}}|}}(\mathbf{G}+\mathbf{k}-\mathbf{K}_1)\cdot(\mathbf{G}'+\mathbf{k}-\mathbf{K}_2) \sum_{\mathbf{G}''\in\mathcal{R}^*_{\mathrm{M},L}} \widetilde{\Sigma}_d[\mathbf{G}''] \underbrace{\langle e_{\mathbf{G}}, e_{\mathbf{G}'+\mathbf{G}''}\rangle}_{\delta_{(\mathbf{G}-\mathbf{G}'),\mathbf{G}''}} \\ &= \frac{1}{\sqrt{|\Omega_{\mathrm{M}}|}}\left[(\mathbf{G}+\mathbf{k}-\mathbf{K}_1)\cdot(\mathbf{G}'+\mathbf{k}-\mathbf{K}_2)\right]\widetilde{\Sigma}_d[\mathbf{G}-\mathbf{G}']. \end{aligned}$$

Similarly,

$$\left\langle e_{\mathbf{G}}, \left[T^{(2)}_{\mathbf{K}_1,\mathbf{K}_2}\right]^* e_{\mathbf{G}'} \right\rangle_{L^2(\mathbb{R}^2,\mathbb{C})} = \frac{1}{\sqrt{|\Omega_{\mathrm{M}}|}}\left[(\mathbf{G}+\mathbf{k}-\mathbf{K}_2)\cdot(\mathbf{G}'+\mathbf{k}-\mathbf{K}_1)\right][\widetilde{\Sigma}_d]^*[\mathbf{G}-\mathbf{G}'].$$

As a result the full matrix of the off-diagonal first order term for all $\mathbf{G}, \mathbf{G}' \in \mathcal{R}^*_{\mathrm{M}}$

$$\left[\mathbf{T}^{(2)}\right]_{\mathbf{G},\mathbf{G}'} = \frac{1}{\sqrt{|\Omega_{\mathrm{M}}|}} \begin{pmatrix} 0 & (\mathbf{G}+\mathbf{k}-\mathbf{K}_1)\cdot(\mathbf{G}'+\mathbf{k}-\mathbf{K}_2)\widetilde{\Sigma}_d[\mathbf{G}-\mathbf{G}'] \\ (\mathbf{G}+\mathbf{k}-\mathbf{K}_2)\cdot(\mathbf{G}'+\mathbf{k}-\mathbf{K}_1)[\widetilde{\Sigma}_d]^*[\mathbf{G}-\mathbf{G}'] & 0 \end{pmatrix}$$

5.3.6 The TwistedBilayerGraphene.jl package

Let us now give a brief presentation of our Julia package `TwistedBilayerGraphene.jl`, in which we implemented the BM and CGG models for TBG. This package was designed as an overlay to the `DFTK` code [HLC21] from which it borrows the structure, the flexibility and ergonomics. We refer to Section 4 for a brief presentation of `DFTK`. As any Julia package, our code and `DFTK` can be easily installed on many platforms, and do not require any configuration or definition of specific environments.

The workflow of our package is simple:

1. The user first creates a `GrapheneSystem` structure, compatible with `DFTK` conventions, which encapsulates the geometry of monolayer graphene, the chosen twist-angle θ for TBG, and informations on the discretization basis set. For example, let us setup a quick computation for TBG at magic angle $\theta = 1.1°$, all other parameters having default value:

```
geometry = GeometryParameters(; θ =1.1)
convergence = ConvergenceParameters()
monolayer = MonolayerGraphene(; geometry, convergence)
```

The `ConvergenceParameters` structure controls the size of the plane-wave discretization basis, and when needed (*e.g.* for the computation of the CGG potential (5.3.27)), the size of the Monkhorst-Pack grid and the SCF tolerance for DFT computations on monolayer graphene.

2. The user then selects a model for TBG: currently either BM or CGG. The model structure contains the list of terms to include in the TBG Hamiltonian, compatible with `DFTK` routines for the resolution of the Bloch fibers' eigenvalue problems. In our example, the BM model is initialized with

```
tbg_model = BM(monolayer)
```

or alternatively in its chiral version [Bec+20]

```
tbg_model = BM(monolayer; chiral=true)
```

It is also possible to construct a custom model by selecting each terms independently, as in the `BM-like` model (see [CGG23]), obtained by selecting the diagonal kinetic term and off-diagonal potential term of CGG, and by neglecting the CGG overlap

```
import TwistedBilayerGraphene: CCGVlocOperators, miGradOperator
terms = [CGGVlocOperator, # off-diagonal CGG potential
         miGradOperator, # diagonal CGG kinetic term
]
tbg_model = TBG(monolayer, terms; name=''BM like'')
```

Note that the CGG overlap is not computed for custom models, unless explicitly asked.

3. When the band model is applied to the `GrapheneSystem`, our code launches with `DFTK` the *ab initio* computations needed to build the selected TBG Hamiltonian;

4. The band diagram is then simply computed in the moiré Brillouin zone with a call to the function `plot_tbg_bandstructure`, which uses the `DFTK` eigensolver routines. In our example, we write

```
BM_bandplot = plot_tbg_bandstructure(tbg_model)
```

which produces the first band diagram of Figure 5.3. The figure also contains band diagrams of chiral BM, CGG and BM-like for the same configuration. Other examples can be found in the documentation of our code at `https://tbg.hollved.org/stable/`.

5.3.7 Conclusions and perspectives

In this work, we have established the groundwork for a user-friendly playground for the simulation of 2D materials, as an overlay to `DFTK`. Our current implementation allows to generate state-of-the art BM and CGG band diagrams for twisted-bilayer graphene in a few simple calls. It also includes automated tests to ensure the code's resilience over time, enabling adaptation to future updates of `Julia` and `DFTK`. Additionally, we have provided a public documentation and an API to facilitate usage for new users and streamline the integration of new functionalities.

The perspective are numerous. At present day, our code constructs the full matrices of the BM and CGG Hamiltonians within the selected plane-wave discretization basis set. One potential direction is to explore algorithms for computing eigenvalues within the bulk spectrum of these operators, enabling a matrix-free implementation of the BM and CGG models. This approach could significantly reduce memory requirements and computational time for the calculations.

Other further developments will focus on incorporating additional features into the package. Recent updates in DFTK have introduced phonon diagrams, paving the way for integrating phonon and electron-phonon interaction models for TBG into our package. Additionally, the structure of our code, mainly based on general purpose routines of `DFTK`, allow the future integration of methods for the simulation of other 2D materials.

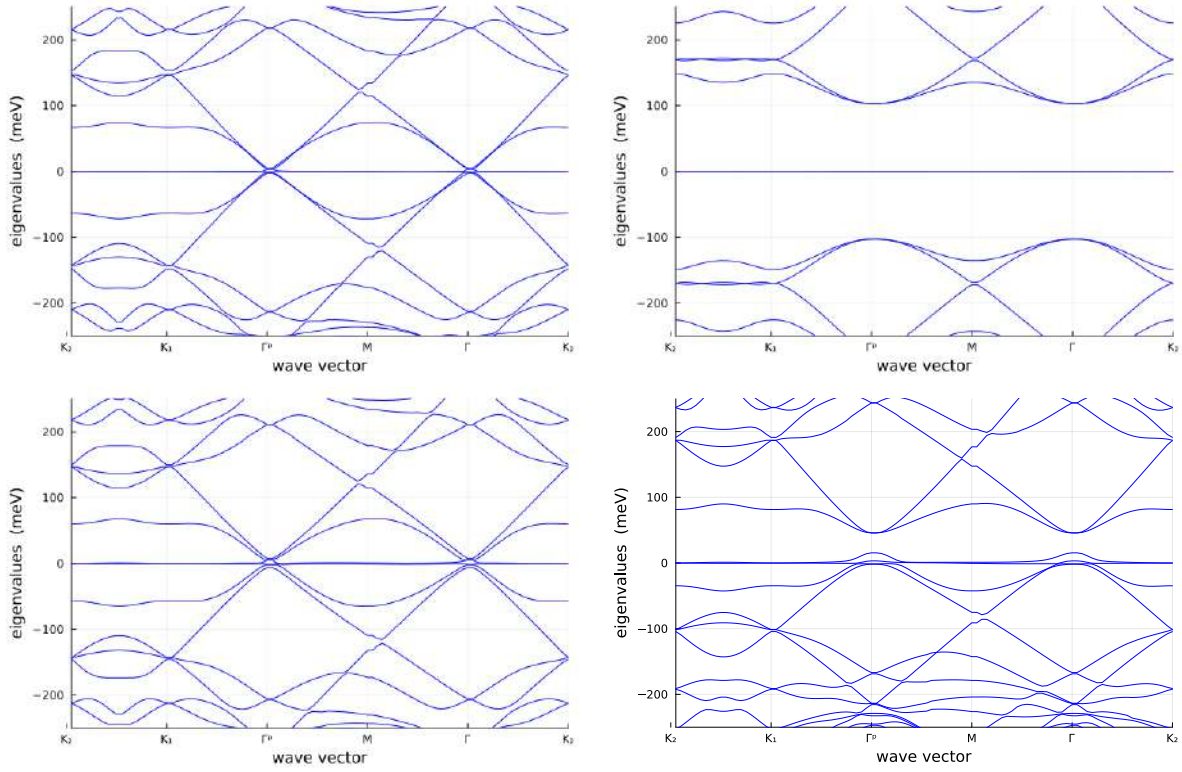


Figure 5.3 – (Up-left) BM, (up-right) chiral BM, (down-left) BM-like and (down-right) CGG band diagrams of TBG, as introduced in [CGG23]. In the **k**-path, the Γ point is the origin of the moiré lattice. The points $\mathbf{K}_1$ and $\mathbf{K}_2$ are moiré quasi-momenta defined in Table 5.1.

5.4 First steps toward large tight-binding simulation of multi-layer graphene with compressed Wannier functions

As discussed earlier, the physics of interest in TBG is partly related to the interaction of electronic states located in the **K** and **K′** valleys of graphene, where the valence bands of graphene intersect conically. A natural way to take these interactions into account is to use a tight-binding approximation parametrized by *Maximally localized Wannier functions* (MLWFs) for these two specific valence bands.

Using a standard Marzari-Vanderbilt (MV) wannierization procedure [MV97] (see Section 4), one can easily compute these two MLWFs w_1 and w_2, which have a shape resembling a p_z atomic orbital in non-interacting-electron atoms (Figure 5.4). In fact, one only needs a single Wannier function, as w_2 is obtained from w_1 by translation and mirror symmetry. In the following we simply denote $w_z = w_1$. Unfortunately, the MV numerical procedure produces fully numerical Wannier functions, obtained as tabulated values on a real or Fourier grid, which makes their practical use in a large tight-binding computation difficult.

In [Bak+18], the authors proposed a systematic way to expand a Wannier function on a basis of symmetry adapted gaussian-type orbitals (SAGTOs). Notably, they apply their method to the case of w_z. For simplicity, they use a restricted set of SAGTOs and mention that more elaborate strategies should be adopted in future works. In the second part of this chapter, we apply the method of [Bak+18] to w_z with a larger set of SAGTOs, in order to use the compressed w_z in large tight-binding computations on TBG.

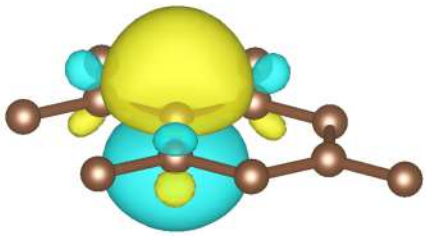


Figure 5.4 – Representation of w_z, a maximally localized Wannier function corresponding to a p_z-like valence band of monolayer graphene. The plot is a level set where the colors represent the sign of the function. It also shows in brown a small sample of monolayer graphene. The data has been generated with `DFTK` [HLC21] and visualized with `VESTA` [MI08].

After briefly describing the compression method in the general case, we derive the expression of a basis of SAGTOs $(\phi_i^{\mathrm{SA}})_{1\leqslant i\leqslant n}$ for w_z and compute in Fourier space the orthogonal projection of w_z on $(\phi_i^{\mathrm{SA}})_{1\leqslant i\leqslant n}$ for the $H^s(\mathbb{R}^3;\mathbb{C})$ canonical inner products, $s\in\mathbb{N}$. In Section 5.4.2, we present our preliminary results and discuss some perspectives.

5.4.1 Compression of w_z on symmetry adapted GTO basis

5.4.1.1 Overview of the general compression method

Let us start by a brief overview of the compression method introduced for a general crystalline system in [Bak+18]. For a given $s\in\mathbb{N}$, the authors consider a Wannier function $w\in H^s(\mathbb{R}^3;\mathbb{C})$ with symmetry point group G, and a set $\mathcal{B}$ of localized symmetry-adapted functions in the sense that

$$\forall\phi^{\mathrm{SA}}\in\mathcal{B},\quad \forall g\in G,\qquad g\cdot\phi^{\mathrm{SA}}:=\phi^{\mathrm{SA}}\circ g^{-1}=\phi^{\mathrm{SA}}. \tag{5.4.1}$$

The compression method reads as the following two-step greedy procedure: given current iterate and residual

$$w_{n-1}=\sum_{i=1}^{n-1}[C_{n-1}]_i\phi_i^{\mathrm{SA}}\quad\text{and}\quad r_{n-1}=\|w-w_{n-1}\|_{H^s}^2 \tag{5.4.2}$$

where $C_{n-1}\in\mathbb{R}^{n-1}$, $n\in\mathbb{N}^*$,

1. choose the next function $\phi_n^{\mathrm{SA}}\in\mathcal{B}$ that best approximates the residual r_n

$$\boxed{\phi_n^{\mathrm{SA}}\ \in\ \operatorname{argmin}\left\{\left\|r_n-\phi^{SA}\right\|_{H^s}^2\,|\,\phi^{\mathrm{SA}}\in\mathcal{B}\right\};} \tag{5.4.3}$$

2. compute w_n as the H^s-orthogonal projection on the basis $(\phi_i^{\mathrm{SA}})_{1\leqslant i\leqslant n}$ by solving the least square problem

$$\boxed{w_n:=\sum_{i=1}^{n}[C_n]_i\phi_i^{\mathrm{SA}},\quad\text{where } C_n\in\operatorname{argmin}\left\{\left\|w-\sum_{i=1}^{n}[C]_i\phi_i^{\mathrm{SA}}\right\|_{H^s}^2\,\middle|\,C\in\mathbb{R}^n\right\}.} \tag{5.4.4}$$

A solution to (5.4.4) is quickly obtained by first expanding the squared-norm in (5.4.4) as

$$\|w\|_{H^s}^2-2\sum_{i=1}^{n}c_i\langle w|\phi_i^{\mathrm{SA}}\rangle_{H^s}+\sum_{i,j=1}^{n}c_ic_j\langle\phi_i^{\mathrm{SA}}|\phi_j^{\mathrm{SA}}\rangle_{H^s}=:\|w\|_{H^s}^2-2C^TX+C^TSC \tag{5.4.5}$$

where we introduced the notation

$$X := [\langle w|\phi_i^{\mathrm{SA}}\rangle_{H^s}]_{1\leqslant i\leqslant n} \in \mathbb{C}^n \quad \text{and} \quad S := [\langle \phi_i^{\mathrm{SA}}|\phi_j^{\mathrm{SA}}\rangle_{H^s}]_{1\leqslant i,j\leqslant n} \in \mathbb{C}^{n\times n}. \tag{5.4.6}$$

A straightforward computation of the gradient of this expression with respect to C provides the set of optimal coefficients $C_n^T = S^{-1}X$. We conclude that

$$\boxed{w_n \;=\; \sum_{i=1}^{n} [X^T S^{-1}]_i \phi_i^{\mathrm{SA}}.} \tag{5.4.7}$$

We emphasize that the original presentation of the compression method is more general. We refer to [Bak+18] for further details.

5.4.1.2 Symmetries of w_z

Let us now identify the symmetry point group of w_z. For the sake of simplicity, we suppose that w_z is centered at the origin of $\mathbb{R}^3$. Using the notations of Section 5.2.1, we observe that

- w_z is odd in the out-of-plane direction

$$\boxed{\mathscr{S}(w_z) \;=\; -w_z;} \tag{5.4.8}$$

- in the monolayer graphene plane, w_z is invariant by rotations of angle $n\frac{2\pi}{3}$ ($n \in \mathbb{Z}$) around the z-axis, and by reflections with respect to the vertical planes containing the covalent bonds with the carbon atom at the origin. Let D_3 be the symmetry group of the equilateral triangle, generated by the symmetry $s_1(x,y) = (-x,y)$ and the rotation $\mathcal{R}_{\frac{2\pi}{3}}$ around the z-axis. Then w_z is D_3-invariant in the sense that

$$\boxed{\forall u \in \mathrm{D}_3 \quad u \cdot w_z(\mathbf{x},z) \;=\; w_z(u^{-1}(\mathbf{x}),z) \;=\; w_z(\mathbf{x},z).} \tag{5.4.9}$$

This implies that w_z belongs to the A_2'' irreducible representation of the $\mathrm{D}_{3\mathrm{h}}$ group [1].

5.4.1.3 Construction of a symmetry-adapted basis

As in [Bak+18], we begin by defining a reference set of Cartesian gaussian-polynomial functions

$$\mathcal{B}^{(0)} = \left\{ \phi : \mathbf{r} \in \mathbb{R}^3 \mapsto p(\mathbf{r}-\mathbf{R})e^{-\zeta|\mathbf{r}-\mathbf{R}|^2} \mid p \in \mathbb{R}^3[\mathbf{X}],\, \mathbf{R} \in \mathbb{R}^3,\, \zeta \in \mathbb{R}_+^* \right\} \tag{5.4.10}$$

where $\mathbb{R}^3[\mathbf{X}]$ denotes the polynomial functions of $\mathbb{R}^3$. We now wish to identify the A_2''-symmetric functions of $\mathcal{B}^{(0)}$, *i.e.* the GTOs $\phi \in \mathcal{B}^{(0)}$ that verify

$$(1)\ \mathscr{S}(\phi) = -\phi \quad \text{and} \quad (2)\ u\cdot\phi = \phi \text{ for all } u \in \mathrm{D}_3. \tag{5.4.11}$$

It is easily seen that the two conditions are met for all centered GTOs of the form

$$\phi(\mathbf{x},z) = \sum_{j\in\mathbb{N}} \sum_{p\in\mathcal{I}(\mathrm{D}_3)} \lambda_{jp} p(\mathbf{x}) z^{2j+1} e^{-\zeta|r|^2}, \quad \lambda_{jp} \in \mathbb{R}, \tag{5.4.12}$$

where $\mathcal{I}(\mathrm{D}_3)$ denotes all polynomials of $\mathbb{R}^2[\mathbf{X}]$ that are A_2''-symmetric. To identify $\mathcal{I}(\mathrm{D}_3)$, we remark that by linearity of the action of D_3 on 2D polynomials

$$\mathcal{I}(\mathrm{D}_3) = \bigoplus_{n\in\mathbb{N}} \mathcal{I}_n(\mathrm{D}_3) \tag{5.4.13}$$

[1] See http://symmetry.jacobs-university.de/cgi-bin/group.cgi?group=603&option=4

where $\mathcal{I}_n(\mathrm{D}_3)$ are the homogeneous polynomials of degree n in $\mathcal{I}(\mathrm{D}_3)$, and we proceed by increasing order n. For $n \leqslant 2$ a quick computation shows that

$$\mathcal{I}_0(\mathrm{D}_3) = \mathrm{Span}\{1\}, \quad \mathcal{I}_1(\mathrm{D}_3) = \{0\} \quad \text{and} \quad \mathcal{I}_2(\mathrm{D}_3) = \mathrm{Span}\{(x_1^2 + x_2^2)\}. \tag{5.4.14}$$

For $n > 2$, we use polar coordinates and decompose $p_n \in \mathcal{I}_n(\mathrm{D}_3)$ as

$$p_n(x_1 = r\cos(\theta), x_2 = r\sin(\theta)) = q\left(\sqrt{x_1^2 + x_2^2}\right) \sum_{|m| \leqslant n} c_m e^{im\theta} = q\left(\sqrt{x_1^2 + x_2^2}\right) \sum_{0 \leqslant m \leqslant \frac{n}{3}} d_m \cos(3m\theta) \tag{5.4.15}$$

where $q \in \mathbb{R}[X]$, and where c_m and d_m are complex numbers. The last equality comes from the fact that p_n verifies (5.4.11). Finally the equality

$$\cos(m\theta) = \mathrm{Re}\left(\sum_{k=0}^{m} i^k \sin^k(\theta)\cos^{m-k}(\theta)\right) \tag{5.4.16}$$

and the identification $x_1 = r\cos(\theta)$, $x_2 = r\sin(\theta)$ allows one to identify the polynomials in $\mathcal{I}_n(\mathrm{D}_3)$. The A_2''-symmetric homogeneous polynomials are shown in Table 5.2 up to degree $n_{\mathbf{x}} = 9$. Using (5.4.12),

degree $n_{\mathbf{x}}$	D_3-invariant homogeneous polynomial
0	1
2	${x_1}^2 + {x_2}^2$
3	${x_1}^3 - 3x_1{x_2}^2$
4	${x_1}^4 + 2{x_1}^2{x_2}^2 + {x_2}^4$
6	${x_1}^6 - 15{x_1}^4{x_2}^2 + 15{x_1}^2{x_2}^4 - {x_2}^6$
9	${x_1}^9 - 36{x_1}^7{x_2}^2 + 126{x_1}^5{x_2}^4 - 84{x_1}^3{x_2}^6 + 9x_1{x_2}^8$

Table 5.2 – Two dimensional homogeneous D_3-symmetric polynomials of up to order 9.

(5.4.13) and Table 5.2, and for given maximum orders $n_{\mathbf{x}}$ and n_z, we construct the set $\mathcal{B}_{n_{\mathbf{x}},n_z} \subset \mathcal{B}^{(0)}$ of A_2''-symmetric SAGTOs

$$B_{n_{\mathbf{x}},n_z} = \left\{ \phi(\mathbf{x}, z) = \sum_{j=0}^{n_z} p_{n_{\mathbf{x}}}(\mathbf{x}) z^{2j+1} e^{-\zeta|\mathbf{r}|^2}, \quad p_{n_{\mathbf{x}}} \in \bigoplus_{n=0}^{n_{\mathbf{x}}} \mathcal{I}_n(\mathrm{D}_3), \quad \zeta \in \mathbb{R}_+^* , \right\}. \tag{5.4.17}$$

To ease convergence, it proves advantageous to construct some SAGTOs by selecting a given $\phi_{\mathrm{CC}} \in \mathcal{B}_{2,n_z}$ with center anywhere on a carbon-carbon bond, and to set

$$\phi = \phi_{\mathrm{CC}} + \mathcal{R}_{\frac{2\pi}{3}}\phi_{\mathrm{CC}} + \mathcal{R}_{\frac{4\pi}{3}}\phi_{\mathrm{CC}} \tag{5.4.18}$$

which is indeed A_2''-symmetric. In practice, we proceed in an alternate fashion, by selecting ϕ as in (5.4.17) or as in (5.4.18).

5.4.1.4 Computation of w_z

As recalled in Section 4, w_z is defined as the inverse Bloch transform of a set of Bloch waves $\{u_{\mathbf{k}}\}$, corresponding to the p_z-like band of monolayer graphene:

$$w_z(\mathbf{x}, z) = \fint_{\Omega^*} u_{\mathbf{k}}(\mathbf{x}, z) e^{i\mathbf{k}\cdot\mathbf{x}} \mathrm{d}\mathbf{k}. \tag{5.4.19}$$

It therefore remains to generate the Bloch waves $\{u_{\mathbf{k}}\}$ using a wannierization procedure. To do so we follow the Marzari-Vanderbilt (MV) [MV97] wannierization method, as described in the introductory Section 4.

1. First we obtain an initial set of Bloch waves and energy bands $(u^{(0)}_{n,\mathbf{k}}, \varepsilon^{(0)}_{n,\mathbf{k}})_{1 \leqslant n \leqslant N, \mathbf{k} \in \Omega^*}$ for monolayer graphene, by a Kohn-Sham PW-DFT calculation.

2. Then we find an optimal gauge $\{U(\mathbf{k})\}$ that minimizes the MV functional Ω^{MV}.

3. Finally we set $u_{n,\mathbf{k}} = \sum_{m=1}^{N} u^{(0)}_{m,\mathbf{k}} U_{m,n}(\mathbf{k})$, identify the n-th band corresponding to a p_z-valence band of graphene and set $u_{\mathbf{k}} = u_{n,\mathbf{k}}$.

The MV functional Ω^{MV} is minimized by a direct minimization procedure, using a finite-difference approximation on a $\mathbf{k}$-point sampling of Ω^*. As detailed in [Mar+12; W90], this procedure only requires the initial data of a family of small $N \times N$ matrices

$$M^{(0)}_{mn}(\mathbf{k},\mathbf{b}) = \langle u^{(0)}_{m,\mathbf{k}} | u^{(0)}_{n,\mathbf{k}+\mathbf{b}} \rangle_{L^2_{\mathrm{per}}(\Omega)} \quad \text{and} \quad A^{(0)}_{mn}(\mathbf{k}) = \langle u_{m\mathbf{k}} e^{i\mathbf{k}\cdot r} | g_n \rangle_{L^2}. \tag{5.4.20}$$

The vectors $\mathbf{b}$ connect a given $\mathbf{k}$-point from the finite-difference grid with its neighbors, and are selected depending on the chosen finite-difference scheme and symmetries. The matrices $A^{(0)}_{mn}$ allow to compute the projection of the Bloch states $\psi_{n\mathbf{k}}(r) = u_{n\mathbf{k}}(r)e^{i\mathbf{k}\cdot r}$ on initial trial localized orbitals $(g_n) \in (L^2(\mathbb{R}^3;\mathbb{C}))^N$. The matrix elements of (5.4.20) are easily computed in Fourier space after introducing a discretization basis set, as in the following section. For some vectors $\mathbf{b}$, the point $\mathbf{k}+\mathbf{b}$ is out of the Brillouin zone. In that case we use the fact that

$$u_{n,(\mathbf{k}+\mathbf{b})}(\mathbf{x},z) = e^{-i\mathbf{b}\cdot\mathbf{x}} u_{n,\mathbf{k}}(\mathbf{x},z). \tag{5.4.21}$$

5.4.1.5 Discretization in a plane-wave basis

In order to evaluate the inner products in (5.4.6) and (5.4.20), we proceed as in the previous section and introduce a truncation of the out-of-plane direction with boundary conditions. For a given height $h \gg 1$, let

$$\mathcal{R}_h = \mathcal{R}_{\mathbf{x}} + h\mathbb{Z} \tag{5.4.22}$$

with unit cell Ω_h and suppose that the Bloch waves are $\mathcal{R}_h$-periodic. We then approximate the continuous Brillouin zone Ω^*_h using the three-dimensional Monkhorts-Pack grid

$$\Omega^*_h \simeq \Omega^*_{h,L} := \left\{ \mathbf{k} = \frac{j_1}{n_1}\mathbf{a}^*_1 + \frac{j_2}{n_2}\mathbf{a}^*_2, \quad j_1, j_2 \in \{0,\dots,L-1\} \right\} \oplus \{0\} \tag{5.4.23}$$

composed of a regular 2D sampling of size L^2 ($L \in \mathbb{N}^*$) in the in-plane direction and of a single point at the origin in the out-of-plane direction. For all $\mathbf{k} \in \Omega^*_{h,L}$, let us decompose $u_{\mathbf{k}}$ in Fourier series

$$u_{\mathbf{k}}(\mathbf{r}) = \frac{1}{L} \sum_{\mathbf{G} \in \mathcal{R}^*_h} u[\mathbf{G}] e^{i\mathbf{G}\cdot\mathbf{r}}. \tag{5.4.24}$$

In that framework, (5.4.19) discretizes as

$$w_z(\mathbf{x},z) = \fint_{\Omega^*} u_{\mathbf{k}}(\mathbf{x},z) e^{i\mathbf{k}\cdot\mathbf{x}} \mathrm{d}\mathbf{k} \simeq \frac{1}{L^3} \sum_{\mathbf{k} \in \Omega^*_{h,L}} \sum_{\mathbf{G} \in \mathcal{R}^*_h} u_{\mathbf{k}}[\mathbf{G}] e^{i(\mathbf{k}+\mathbf{G})\cdot r}. \tag{5.4.25}$$

An immediate consequence of discretization, as appearing in (5.4.25), is that w_z becomes periodic of the lattice $\mathcal{R}_{h,L} = L\mathcal{R}_{\mathbf{x}} + h\mathbb{Z}$. In other words, the symmetry group of the discrete w_z is the space group

$$\mathrm{D}_{3\mathrm{h}} \ltimes \mathcal{R}_{h,L}. \tag{5.4.26}$$

We therefore have to adapt the construction of SAGTOs of the previous section to the discrete setting by considering

$$\mathcal{B}_{n_{\mathbf{x}},n_z}(h,L) = \left\{ \phi^{\mathrm{SA}} = \sum_{\mathbf{R} \in \mathcal{R}_{h,L}} \phi(\cdot - \mathbf{R}), \quad \phi \in \mathcal{B}_{n_{\mathbf{x}},n_z} \right\}. \tag{5.4.27}$$

To avoid any confusion, we call $\mathcal{R}_{h,L}$ the *supercell* lattice and adopt the following conventions:

- $\mathbf{q}$ denotes a point of the reciprocal supercell lattice $\mathcal{R}^*_{h,L}$ (the plane-wave discretization basis for w_z);

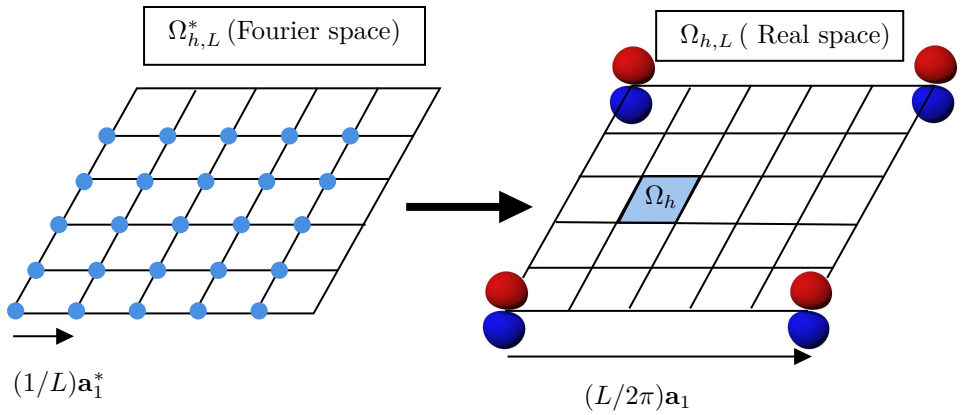


Figure 5.5 – (Left) Monkhorst-Pack discretization of the reciprocal unit cell $\Omega^*_{h,L}$ as a $L \times L \times 1$ grid. (Right) The corresponding Wannier function, pictured as a standard p_z atomic orbital, is periodic of a supercell made of L^2 monolayer graphene unit cells Ω_h. We omitted the periodicity in the out-of-plane direction for readability.

- $\mathbf{k}$ denotes a point of the sampled reciprocal supercell $\Omega^*_{h,L}$;
- $\mathbf{G}$ denotes a point of the truncated (with standard cell) lattice $\mathcal{R}^*_h$ (corresponding to the plane-wave discretization of the monolayer graphene Bloch waves $u_{n,\mathbf{k}}$).

Note that for all $\mathbf{k} \in \Omega^*_{h,L}$ and $\mathbf{G} \in \mathcal{R}_h$, $\mathbf{q} := \mathbf{k} + \mathbf{G} \in \mathcal{R}^*_{h,L}$.

For all $\phi^{\mathrm{SA}} \in \mathcal{B}_{n_\mathbf{x},n_z}(h,L)$, we write

$$\forall \mathbf{r} \in \mathbb{R}^3 \quad w_z(\mathbf{r}) \simeq \frac{1}{L} \sum_{\mathbf{q} \in \mathcal{R}^*_{L,h}} w_z[\mathbf{q}] e^{i\mathbf{q}\cdot\mathbf{r}}, \qquad \phi^{\mathrm{SA}}(\mathbf{r}) \simeq \frac{1}{L} \sum_{\mathbf{q} \in \mathcal{R}^*_{L,h}} \phi^{\mathrm{SA}}[\mathbf{q}] e^{i\mathbf{q}\cdot\mathbf{r}}. \tag{5.4.28}$$

The last step is to compute explicitly the Fourier coefficients $w_z[\mathbf{q}]$ and $\phi^{\mathrm{SA}}[\mathbf{q}]$. For all basis function $\phi^{\mathrm{SA}} \in \mathcal{B}_{n_\mathbf{x},n_z}(h,L)$ and quasi-momentum $\mathbf{q} \in \mathcal{R}^*_{L,h}$, we first remark that

$$\phi^{\mathrm{SA}}[\mathbf{q}] = \fint_{\Omega_{L,h}} \phi^{\mathrm{SA}}(\mathbf{r}) e^{-i\mathbf{q}\cdot\mathbf{r}} \mathrm{d}\mathbf{r} = \sum_{\mathbf{R} \in \mathcal{R}_{L,h}} \fint_{\Omega_{L,h}} \phi(\mathbf{r}-\mathbf{R}) e^{-i\mathbf{q}\cdot\mathbf{r}} \mathrm{d}\mathbf{r} = \mathcal{F}(\phi^{\mathrm{SA}})(\mathbf{q}). \tag{5.4.29}$$

Let us write $\phi^{\mathrm{SA}}(\mathbf{r}) = \sum_{j=1}^{n_z} p_{n_\mathbf{x}}(\mathbf{r}) z^{2j+1} g_\zeta(\mathbf{r})$ with $g_z(\mathbf{r}) := e^{-\zeta|\mathbf{r}|^2}$ and $\lambda_1, \ldots, \lambda_{n_\mathbf{x}} \in \mathbb{R}$ be such that

$$p_{n_\mathbf{x}} = \sum_{n=1}^{n_\mathbf{x}} \lambda_n p_n, \quad p_n \in \mathcal{I}_n(\mathrm{D}_3), \quad \forall n \in \{1, \ldots, n_\mathbf{x}\}. \tag{5.4.30}$$

The Fourier transform $\mathcal{F}(\phi^{\mathrm{SA}})$ can be computed using Fourier duality as

$$\mathcal{F}(\phi^{\mathrm{SA}}) = \left(\sum_{j=1}^{n_z} \sum_{n=1}^{n_\mathbf{x}} \lambda_n (-i)^{n+2j+1} p(\partial_\mathbf{x}) \partial_z^{2j+1} \right) \mathcal{F}(g_\zeta). \tag{5.4.31}$$

Since $\mathcal{F}(g_\zeta)$ is known, (5.4.31) can be computed analytically, although deriving (5.4.31) by hand is very cumbersome for high degrees $n_\mathbf{x}$ and n_z. In practice, we obtain $\mathcal{F}(\phi^{\mathrm{SA}})$ by applying forward automatic

differentiation [RLP16] iteratively, using the chain rule, which brings no additional numerical errors. When it comes to w_z, we simply identify in (5.4.25) with our discretization conventions

$$\boxed{w_z[\mathbf{k}+\mathbf{G}] = \frac{u_{\mathbf{k}}[\mathbf{G}]}{L^2}, \quad \forall (\mathbf{k}+\mathbf{G}) \in \mathcal{R}^*_{L,h}.} \tag{5.4.32}$$

Finally if $w_n = \sum_{i=1}^{n} \phi_i^{\mathrm{SA}}$ is the current iterate, then equations (5.4.31) and (5.4.32) yield for all $1 \leqslant i, j \leqslant n$

$$\boxed{\begin{aligned} X_i &= \frac{1}{L\sqrt{|\Omega_h|}} \sum_{\mathbf{q}=(\mathbf{k}+\mathbf{G})\in\mathcal{R}^*_{L,h}} (1+|\mathbf{q}|^2)^s \overline{u_{\mathbf{k}}[\mathbf{G}]} \mathcal{F}(\phi_i^{\mathrm{SA}})(\mathbf{q}) \\ S_{ij} &= \sum_{\mathbf{q}\in\mathcal{R}^*_{L,h}} (1+|\mathbf{q}|^2)^s \overline{\mathcal{F}(\phi_i^{\mathrm{SA}})(\mathbf{q})} \mathcal{F}(\phi_j^{\mathrm{SA}})(\mathbf{q}). \end{aligned}} \tag{5.4.33}$$

5.4.2 Numerical results

We start this numerical section by giving some implementations details. More information can be found in our code freely available at `https://github.com/LaurentVidal95/Wannier2GTO`.

5.4.2.1 Implementation details

Monolayer self-consistent field computation. For given parameters h and L, we use `DFTK` to produce a set of $N = 15$ Bloch waves and energy bands $(u^{(0)}_{n,\mathbf{k}}, \varepsilon^{(0)}_{n,\mathbf{k}})$, using a KS-DFT method with Perdew-Burke-Ernzerhof (PBE) functional [PBE96] and Hartwigsen-Goedecker-Teter-Hutter separable dual-space Gaussian pseudopotentials [HGH98]. The parameter h is set to match the value of [Bak+18, Table 2]. The $\mathbf{k}$-fiber eigenvalue problems are solved using $\mathbf{k}$-dependent Fourier discretization basis sets (as presented in Section 4) defined by a cut-off parameter $\mathrm{E_c}$. The cut-off energy $\mathrm{E_c}$ and sampling precision L are set by a convergence analysis with respect to w_z. The value $\mathrm{E_c} = 50$ Ha and $L = 8$ provide stable results within a range of 1% in the H^1-norm of w_z.

Wannierization. To obtain the MLWFs of the valence band of graphene, we originally interfaced `DFTK` with `Wannier90` [Piz+20], a commonly used software that implements the MV wannierization procedure. The interface works by generating two input files for `Wannier90` and by parsing and integrating the output data in `DFTK`: the first input file contains the geometry and convergence parameters needed to generate a list of $\mathbf{b}$ vectors; the second input files contains values of the matrix elements (5.4.20) (computed in `DFTK`) and produces the optimal gauge $\{U(\mathbf{k})\}$. The whole procedure has been integrated in `DFTK`.

Parts of this work were later used to integrate the wannierization package `Wannier.jl` [Qia], fully written in `Julia` language, in `DFTK`.

5.4.2.2 Preliminary results

We implemented the compression routine as described above in a `Julia` proof-of-concept code. By setting $n_{\mathbf{x}} = 5$ and $n_z = 5$, we managed to construct a basis $\Phi^{\mathrm{SA}} \in \mathcal{B}_{5,5}(h, L)^{10}$ of 10 basis functions, depending on 155 parameters, such that the H^1-projection $\Pi_{\Phi^{\mathrm{SA}}}(w_z)$ of w_z on Φ^{SA} verifies

$$\|w_z - \Pi_{\Phi^{\mathrm{SA}}}(w_z)\|_{H^1} \simeq 0.12\,. \tag{5.4.34}$$

In the original paper [Bak+18], the same precision is obtained for at least twice the number of parameters. Note that this choice of parameter $n_{\mathbf{x}}$ and n_z produced the best performance.

Unfortunately, adding more basis functions to Φ^{SA} resulted in the conditioning of the overlap matrix S to blow-up. This main limitation of our proof-of-concept code, which hindered the compression procedure, can be remedied for example by minimizing the condition number of S in the compression procedure, or by applying other standard methods to cure ill-conditioned basis sets. However, this aspect remains unexplored and is left for future investigation.

5.4.2.3 Perspectives

As we lacked the time to complete this study, the potential path for further explorations are numerous. Regarding the compression procedure, our initial results suggest that using a larger set of SAGTOs than in [Bak+18] allows to achieve the same precision with a smaller basis set. However, our naive implementation resulted in ill-conditioned SAGTOs, preventing further investigation into the use of larger and potentially more accurate basis sets. Our first task should therefore be devoted to address the bad conditioning inherent to the construction of SAGTOs within our implementation.

Our objective in obtaining a compressed version of w_z is to speed-up the computation of tight-binding elements for large-scale calculations in TBG. Presently, our code is interfaced with the GaIn [Duc] `C++` library developed by Ivan Duchemin, which handles analytical integrals involving GTOs and Hamiltonian terms. In a future work, we should evaluate the time saved by using analytical integrals compared to standard quadrature methods, and quantify the impact of compression on the accuracy in the calculation of tight-binding matrix elements.

Bibliography

[Abi] *ABINIT Software Suite User Guide*. https://docs.abinit.org/variables/rlx/#ecutsm.

[AM76] N. Ashcroft and N. Mermin. *Solid State Physics*. Holt, Rinehart and Winston, New York, 1976.

[Amm+78] J.H. Ammeter, H.B. Bürgi, J.C. Thibeault, and R. Hoffmann. "Counterintuitive orbital mixing in semiempirical and ab initio molecular orbital calculations". In: *Journal of the American Chemical Society* 100.12 (1978), pp. 3686–3692.

[AMS08] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

[And65] Donald G Anderson. "Iterative procedures for nonlinear integral equations". In: *Journal of the ACM (JACM)* 12.4 (1965), pp. 547–560.

[Ang+02] Celestino Angeli, Stefano Evangelisti, Renzo Cimiraglia, and Daniel Maynau. "A novel perturbation-based complete active space–self-consistent-field algorithm: Application to the direct calculation of localized orbitals". In: *J. Chem. Phys.* 117.23 (2002), pp. 10525–10533.

[ARS+05] A. A. Ramírez-Solís, R. Poteau, A. Vela, and J. P. Daudey. "Comparative studies of the spectroscopy of $CuCl_2$: DFT versus standard ab initio approaches". In: *J. Chem. Phys.* 122.16 (2005), p. 164306.

[AT87] Jan Almlöf and Peter R. Taylor. "General Contraction of Gaussian Basis Sets. I. Atomic Natural Orbitals for First- and Second-row Atoms". In: *The Journal of Chemical Physics* 86.7 (1987), pp. 4070–4077. DOI: 10.1063/1.451917.

[Ata+06] Mihail Atanasov et al. "DFT models for copper(II) bispidine complexes: Structures, stabilities, isomerism, spin distribution, and spectroscopy". In: *J. Comput. Chem.* 27.12 (2006), pp. 1263–1277. ISSN: 0192-8651. DOI: 10.1002/jcc.20412.

[BA15] Nicolas Boumal and P-A Absil. "Low-rank matrix completion via preconditioned optimization on the Grassmann manifold". In: *Linear Algebra and its Applications* 475 (2015), pp. 200–239.

[Bac81] G.B. Bacskay. "A quadratically convergent Hartree—Fock (QC-SCF) method. Application to closed shell systems". In: *Chemical Physics* 61.3 (1981), pp. 385–404. DOI: 10.1016/0301-0104(81)85156-7.

[Bac82] George B. Bacskay. "A quadratically convergent Hartree-Fock (QC-SCF) method. Application to open shell orbital optimization and coupled perturbed Hartree-Fock calculations". In: *Chem. Phys.* 65 (1982), pp. 383–396. DOI: 10.1016/0301-0104(82)85211-7.

[Bak+18] Athmane Bakhta, Eric Cancès, Paul Cazeaux, Shiang Fang, and Efthimios Kaxiras. "Compression of Wannier functions into Gaussian-type orbitals". In: *Computer Physics Communications* 230 (2018), pp. 27–37.

[BCS14] Markus Bachmayr, Huajie Chen, and Reinhold Schneider. "Error estimates for Hermite and even-tempered Gaussian approximations in quantum chemistry". In: *Numer. Math.* 128.1 (Sept. 2014), pp. 137–165.

[BDJ02] Jean-Louis Basdevant, Jean Dalibard, and Manuel Joffre. *Mécanique quantique*. Editions Ecole Polytechnique, 2002.

[BE50] S F Boys and Alfred Charles Egerton. “Electronic wave functions - I. A general method of calculation for the stationary states of any molecular system”. In: *Proc. R. Soc. Lond. A Math. Phys. Sci.* 200.1063 (Feb. 1950), pp. 542–554.

[Bec+20] Simon Becker, Mark Embree, Jens Wittsten, and Maciej Zworski. “Spectral Characterization of Magic Angles in Twisted Bilayer Graphene”. In: (2020). DOI: 10.48550/ARXIV.2010.05279. (Visited on 05/10/2023).

[Ber+95] M. Bernasconi, G. L. Chiarotti, P. Focher, S. Scandolo, E. Tosatti, and M. Parrinello. “First-principle-constant pressure molecular dynamics”. In: *Journal of Physics and Chemistry of Solids* 56.3-4 (1995), pp. 501–505.

[Bez+17] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. “Julia: A fresh approach to numerical computing”. In: *SIAM Review* 59.1 (2017), pp. 65–98. DOI: 10.1137/141000671. URL: https://epubs.siam.org/doi/10.1137/141000671.

[BJA94] P. E. Blöchl, O. Jepsen, and O. K. Andersen. “Improved tetrahedron method for Brillouin-zone integrations”. In: *Physical Review B* 49.23 (1994), p. 16223.

[Bla+97] J.-P. Blaudeau, M.P. McGrath, L.A. Curtiss, and L. Radom. “Extension of Gaussian-2 (G2) theory to molecules containing third-row atoms K and Ca”. In: *The Journal of Chemical Physics* 107.13 (1997), pp. 5016–5021. DOI: 10.1063/1.474865.

[Bli19] SM Blinder. “Eigenvalues for a Pure Quartic Oscillator”. In: *arXiv preprint arXiv:1903.07471* (2019).

[Blö94] Peter E Blöchl. “Projector augmented-wave method”. In: *Physical review B* 50.24 (1994), p. 17953.

[BM07] Rodney J Bartlett and Monika Musiał. “Coupled-cluster theory in quantum chemistry”. In: *Reviews of Modern Physics* 79.1 (2007), p. 291.

[BM11] Rafi Bistritzer and Allan H MacDonald. “Moiré bands in twisted double-layer graphene”. In: *Proceedings of the National Academy of Sciences* 108.30 (2011), pp. 12233–12237.

[Boo+13] George H. Booth, Andreas Grüneis, Georg Kresse, and Ali Alavi. “Towards an exact description of electronic wavefunctions in real solids”. In: *Nature* 493.7432 (2013), pp. 365–370. ISSN: 1476-4687. DOI: 10.1038/nature11770.

[Bou23] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023. DOI: 10.1017/9781009166164. URL: https://www.nicolasboumal.net/book.

[BPH80] J S Binkley, J A Pople, and W J Hehre. “Self-consistent molecular orbital methods. 21. Small split-valence basis sets for first-row elements”. In: *Journal of the American* (1980).

[Bro+07] Christian Brouder, Gianluca Panati, Matteo Calandra, Christophe Mourougane, and Nicola Marzari. “Exponential localization of Wannier functions in insulators”. In: *Physical review letters* 98.4 (2007), p. 046402.

[Can00] E. Cancès. “SCF algorithms for HF electronic calculations”. In: *Mathematical models and methods for ab initio quantum chemistry*. Springer, 2000, pp. 17–43.

[Can01] E. Cancès. “Self-consistent field algorithms for Kohn–Sham models with fractional occupation numbers”. In: *The Journal of Chemical Physics* 114.24 (2001), pp. 10616–10622. DOI: 10.1063/1.1373430. eprint: https://doi.org/10.1063/1.1373430. URL: https://doi.org/10.1063/1.1373430.

[Can+03] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday. *Computational quantum chemistry: a primer*. Vol. X. Handbook of Numerical Analysis. North-Holland, Amsterdam, 2003, pp. 3–270.

[Can+07] Claudio Canuto, M Yousuff Hussaini, Alfio Quarteroni, and Thomas A Zang. *Spectral methods: fundamentals in single domains*. Springer Science & Business Media, 2007.

[Can+20] E. Cancès, V. Ehrlacher, D. Gontier, A. Levitt, and D. Lombardi. “Numerical quadrature in the Brillouin zone for periodic Schrödinger operators”. In: *Numerische Mathematik* 144.3 (2020), pp. 479–526. ISSN: 0029-599X. DOI: 10.1007/s00211-019-01096-w. URL: https://doi.org/10.1007/s00211-019-01096-w.

[Can+21a] E. Cancès, C. Fermanian-Kammerer, A. Levitt, and S. Siraj-Dine. "Coherent electronic transport in periodic crystals". In: *Annales Henri Poincaré. A Journal of Theoretical and Mathematical Physics* 22.8 (2021), pp. 2643–2690. ISSN: 1424-0637. DOI: 10.1007/s00023-021-01026-3. URL: https://doi.org/10.1007/s00023-021-01026-3.

[Can+21b] Eric Cancès, Geneviève Dusson, Gaspard Kemlin, and Antoine Levitt. *Practical Error Bounds for Properties in Plane-Wave Electronic Structure Calculations*. https://hal.inria.fr/hal-03408321, 2021.

[Cao+18] Yuan Cao et al. "Unconventional superconductivity in magic-angle graphene superlattices". In: *Nature* 556.7699 (2018), pp. 43–50.

[CB00] E. Cancès and C. Le Bris. "Can we outperform the DIIS approach for electronic structure calculations?" In: *International Journal of Quantum Chemistry* 79.2 (2000), pp. 82–90. DOI: 10.1002/1097-461x(2000)79:2<82::aid-qua3>3.0.co;2-i.

[CCM10] E. Cancès, R. Chakir, and Y. Maday. "Numerical analysis of nonlinear eigenvalue problems". In: *Journal of Scientific Computing* 45.1 (2010), pp. 90–117.

[CF23] Eric Cancès and Gero Friesecke. *Density Functional Theory: Modeling, Mathematical Analysis, Computational Methods, and Applications*. Springer Nature, 2023.

[CGG23] Éric Cancès, Louis Garrigue, and David Gontier. "Simple derivation of moiré-scale continuous models for twisted bilayer graphene". In: *Physical Review B* 107.15 (2023), p. 155403.

[CGS16] E. Cancés, D. Gontier, and G. Stoltz. "A mathematical analysis of the GW0 method for computing electronic excited energies of molecules". In: *Reviews in Mathematical Physics* 28.04 (2016), p. 1650008. DOI: 10.1142/S0129055X16500082.

[Chu+21] Maxime Chupin, Mi-Song Dupuy, Guillaume Legendre, and Eric Séré. "Convergence analysis of adaptive DIIS algorithms with application to electronic ground state calculations". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 55.6 (2021), pp. 2785–2825.

[CKL21] Eric Cancès, Gaspard Kemlin, and Antoine Levitt. "Convergence analysis of direct minimization and self-consistent iterations". In: *SIAM Journal on Matrix Analysis and Applications* 42.1 (2021), pp. 243–274.

[CLB00a] E. Cancès and C. Le Bris. "Can we outperform the DIIS approach for electronic structure calculations?" In: *International Journal of Quantum Chemistry* 79.2 (2000), pp. 82–90. DOI: 10.1002/1097-461X(2000)79:2<82::AID-QUA3>3.0.CO;2-I. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1097-461X%282000%2979%3A2%3C82%3A%3AAID-QUA3%3E3.0.CO%3B2-I. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-461X%282000%2979%3A2%3C82%3A%3AAID-QUA3%3E3.0.CO%3B2-I.

[CLB00b] E. Cancès and C. Le Bris. "On the convergence of SCF algorithms for the Hartree-Fock equations". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 34.4 (2000), pp. 749–774. DOI: 10.1051/m2an:2000102.

[CLBM06] Eric Cancès, Claude Le Bris, and Yvon Maday. *Méthodes mathématiques en chimie quantique. Une introduction*. Vol. 53. Mathématiques & Applications. Springer Berlin Heidelberg, 2006.

[Con78] J. B. Conway. *Functions of one complex variable*. Second. Vol. 11. Graduate Texts in Mathematics. Springer-Verlag, New York-Berlin, 1978, pp. xiii+317. ISBN: 0-387-90328-3.

[CSG97] G. Chaban, M.W. Schmidt, and M.S. Gordon. "Approximate second order method for orbital optimization of SCF and MCSCF wavefunctions". In: *Theoretical Chemistry Accounts* 97.1 (1997), pp. 88–95.

[Das+22] Sambit Das, Phani Motamarri, Vishal Subramanian, David M Rogers, and Vikram Gavini. "DFT-FE 1.0: A massively parallel hybrid CPU-GPU density functional theory code using finite-element discretization". In: *Computer Physics Communications* 280 (2022), p. 108473.

[DCM20] Loredana Edith Daga, Bartolomeo Civalleri, and Lorenzo Maschio. "Gaussian basis sets for crystalline solids: All-purpose basis set libraries vs system-specific optimizations". In: *Journal of chemical theory and computation* 16.4 (2020), pp. 2192–2201.

[Dft] *DFTK.jl: The density-functional toolkit*. https://docs.dftk.org/stable/.

[DG12] R. Dreizler and E. K. Gross. *Density functional theory: an approach to the quantum many-body problem*. Springer Science & Business Media, 2012.

[DL18] Anil Damle and Lin Lin. "Disentanglement via entanglement: a unified method for Wannier localization". In: *Multiscale Modeling & Simulation* 16.3 (2018), pp. 1392–1410.

[DLL19] Anil Damle, Antoine Levitt, and Lin Lin. "Variational formulation for Wannier functions with entangled band structure". In: *Multiscale Modeling & Simulation* 17.1 (2019), pp. 167–191.

[DLY15] Anil Damle, Lin Lin, and Lexing Ying. "Compressed representation of Kohn–Sham orbitals via selected columns of the density matrix". In: *Journal of chemical theory and computation* 11.4 (2015), pp. 1463–1469.

[DSH18] Hans De Sterck and Alexander JM Howse. "Nonlinearly preconditioned L-BFGS as an acceleration mechanism for alternating least squares with application to tensor decomposition". In: *Numerical Linear Algebra with Applications* 25.6 (2018), e2202.

[Duc] Ivan Duchemin. *GaIn - a simple Gaussian Integral library*. https://gitlab.maisondelasimulation.fr/beDeft/GaIn.

[Dun89] T.H. Dunning. "Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen". In: *The Journal of Chemical Physics* 90.2 (1989), pp. 1007–1023. DOI: 10.1063/1.456153.

[Dup18] Mi-Song Dupuy. "Analysis of the projector augmented-wave method for electronic structure calculations in periodic settings". PhD thesis. Université Sorbonne Paris Cité, 2018.

[DVHG02] B.D. Dunietz, T. Van Voorhis, and M. Head-Gordon. "Geometric direct minimization of Hartree-Fock calculations involving open shell wavefunctions with spin restricted orbitals". In: *Journal of Theoretical and Computational Chemistry* 01.02 (Oct. 2002), pp. 255–261. DOI: 10.1142/s0219633602000233.

[EAS98] A. Edelman, T.A. Arias, and S.T. Smith. "The geometry of algorithms with orthogonality constraints". In: *SIAM journal on Matrix Analysis and Applications* 20.2 (1998), pp. 303–353.

[Eva14] Francesco A Evangelista. "Adaptive multiconfigurational wave functions". In: *The Journal of Chemical Physics* 140.12 (2014).

[Fra+82] M.M. Francl et al. "Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements". In: *The Journal of Chemical Physics* 77.7 (1982), pp. 3654–3665. DOI: 10.1063/1.444267.

[FW12] Charles Fefferman and Michael Weinstein. "Honeycomb lattice potentials and Dirac points". In: *Journal of the American Mathematical Society* 25.4 (2012), pp. 1169–1220.

[Gar+19] Y. Garniron et al. "Quantum package 2.0: An open-source determinant-driven suite of programs". In: *Journal of chemical theory and computation* 15.6 (2019), pp. 3591–3609.

[GL16] D. Gontier and S. Lahbabi. "Convergence rates of supercell calculations in the reduced Hartree-Fock model". In: *ESAIM. Mathematical Modelling and Numerical Analysis* 50.5 (2016), pp. 1403–1424. ISSN: 2822-7840. DOI: 10.1051/m2an/2015084. URL: https://doi.org/10.1051/m2an/2015084.

[GP13] G. Grosso and G. P. Parravicini. *Solid state physics*. Academic press, 2013.

[GS74] M.F. Guest and V.R. Saunders. "On methods for converging open-shell Hartree-Fock wavefunctions". In: *Molecular Physics* 28.3 (1974), pp. 819–828.

[Har57] D. R. Hartree. *The calculation of atomic structures*. Wiley, London, 1957.

[Har80] W. A. Harrison. *Solid state theory*. Courier Corporation, 1980.

[HDP72] Warren J Hehre, Robert Ditchfield, and John A Pople. "Selfconsistent molecular orbital methods. XII. Further extensions of Gaussiantype basis sets for use in molecular orbital studies of organic molecules". In: *The Journal of Chemical Physics* 56.5 (1972), pp. 2257–2261.

[Hen01] J. Henk. "Integration over two-dimensional Brillouin zones by adaptive mesh refinement". In: *Physical Review B* 64.3 (2001).

[HGA15] Wen Huang, Kyle A Gallivan, and P-A Absil. "A Broyden class of quasi-Newton methods for Riemannian optimization". In: *SIAM Journal on Optimization* 25.3 (2015), pp. 1660–1685.

[HGH98] C. Hartwigsen, S. Gœdecker, and J. Hutter. "Relativistic separable dual-space Gaussian pseudopotentials from H to Rn". In: *Physical Review B* 58.7 (1998).

[Hin+17] Yoyo Hinuma, Giovanni Pizzi, Yu Kumagai, Fumiyasu Oba, and Isao Tanaka. "Band structure diagram paths based on crystallography". In: *Computational Materials Science* 128 (2017), pp. 140–184.

[HJO14] Trygve Helgaker, Poul Jorgensen, and Jeppe Olsen. *Molecular Electronic-Structure Theory*. en. John Wiley & Sons, Aug. 2014.

[HK64] Pierre Hohenberg and Walter Kohn. "Inhomogeneous electron gas". In: *Physical review* 136.3B (1964), B864.

[HL06] Ulrich Hetmaniuk and Rich Lehoucq. "Basis selection in LOBPCG". In: *Journal of Computational Physics* 218.1 (2006), pp. 324–332.

[HL20] Michael F Herbst and Antoine Levitt. "Black-box inhomogeneous preconditioning for self-consistent field iterations in density functional theory". In: *Journal of Physics: Condensed Matter* 33.8 (2020), p. 085503.

[HLC21] M. F. Herbst, A. Levitt, and E. Cancès. "DFTK: A Julian approach for simulating electrons in solids". In: *Proc. JuliaCon Conf.* 3 (2021), p. 69. DOI: 10.21105/jcon.00069.

[HMW23] Muhammad Hassan, Yvon Maday, and Yipeng Wang. "Analysis of the single reference coupled cluster method for electronic structure calculations: the full-coupled cluster equations". In: *Numerische Mathematik* 155.1-2 (2023), pp. 121–173.

[Hof63] R. Hoffmann. "An extended Hückel theory. I. Hydrocarbons". In: *The Journal of Chemical Physics* 39.6 (1963), pp. 1397–1412.

[HP74] P.C. Hariharan and J.A. Pople. "Accuracy of AHn equilibrium geometries by single determinant molecular orbital theory". In: *Molecular Physics* 27.1 (1974), pp. 209–214. DOI: 10.1080/00268977400100171.

[HP85] Tracy P. Hamilton and Peter Pulay. "Direct inversion in the iterative subspace (DIIS) optimization of open-shell, excited-state, and small multiconfiguration SCF wave functions". In: *J. Chem. Phys.* 84 (1985), pp. 5728–5734. DOI: 10.1063/1.449880.

[HP86] T.P. Hamilton and P. Pulay. "Direct inversion in the iterative subspace (DIIS) optimization of open-shell, excited-state, and small multiconfiguration SCF wave functions". In: *The Journal of Chemical Physics* 84.10 (1986), pp. 5728–5734. DOI: 10.1063/1.449880.

[HSC79] DR Hamann, M Schlüter, and C Chiang. "Norm-conserving pseudopotentials". In: *Physical review letters* 43.20 (1979), p. 1494.

[HSP69] W J Hehre, R F Stewart, and J A Pople. "SelfConsistent MolecularOrbital Methods. I. Use of Gaussian Expansions of SlaterType Atomic Orbitals". In: *J. Chem. Phys.* 51.6 (Sept. 1969), pp. 2657–2664.

[HY10] Xiangqian Hu and Weitao Yang. "Accelerating self-consistent field convergence with the augmented Roothaan-Hall energy function". In: *The Journal of Chemical Physics* 132.5 (Feb. 2010), p. 054109. ISSN: 0021-9606. DOI: 10.1063/1.3304922. eprint: https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/1.3304922/16151464/054109_1_online.pdf. URL: https://doi.org/10.1063/1.3304922.

[HZ05] William W Hager and Hongchao Zhang. "A new conjugate gradient method with guaranteed descent and an efficient line search". In: *SIAM Journal on optimization* 16.1 (2005), pp. 170–192.

[HZ06] William W Hager and Hongchao Zhang. "Algorithm 851: CG_DESCENT, a conjugate gradient method with guaranteed descent". In: *ACM Transactions on Mathematical Software (TOMS)* 32.1 (2006), pp. 113–137.

[ICM96a] F. De Vuyst I. Charpentier and Y. Maday. "A component mode synthesis method of infinite order of accuracy using subdomain overlapping: numerical analysis and experiments". In: *Publication du laboratoire dAnalyse Numerique* 96002 (1996), pp. 55–65.

[ICM96b] F. De Vuyst I. Charpentier and Y. Maday. “Méthode de synthèse modale avec une décomposition de domaine par recouvrement”. In: *Comptes rendus de l’Académie des sciences. Série 1, Mathématique* 322.9 (1996), pp. 881–888.

[IR08] I. Ipsen and R. Rehman. “Perturbation bounds for determinants and characteristic polynomials”. In: *SIAM Journal on Matrix Analysis and Applications* 30.2 (2008), pp. 762–776. ISSN: 0895-4798. DOI: 10.1137/070704770.

[JA86] Hans Jørgen Aa Jensen and Hans Agren. “A Direct, restricted-step, second-order MC SCF program for large scale ab initio calculations”. In: *Chem. Phys.* 104 (1986), pp. 229–250.

[Jan+16] J. L. Janssen, Y. Gillet, S. Poncé, A. Martin, M. Torrent, and X. Gonze. “Precise effective masses from density functional perturbation theory”. In: *Physical Review B* 93.20 (2016).

[Jen01] Frank Jensen. “Polarization consistent basis sets: Principles”. In: *The Journal of Chemical Physics* 115.20 (2001), pp. 9113–9125.

[Jen13] Frank Jensen. “Atomic orbital basis sets”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 3.3 (2013), pp. 273–295.

[JJ84] Hans Jørgen Aa Jensen and Poul Jørgensen. “A direct approach to secondorder MCSCF calculations using a norm extended optimization scheme”. In: *J. Chem. Phys.* 80.3 (1984), pp. 1204–1214. ISSN: 0021-9606. DOI: 10.1063/1.446797.

[Kat57] Tosio Kato. “On the eigenfunctions of many-particle systems in quantum mechanics”. In: *Communications on Pure and Applied Mathematics* 10.2 (1957), pp. 151–177.

[Kax03] E. Kaxiras. *Atomic and Electronic Structure of Solids*. Cambridge University Press, 2003. DOI: 10.1017/CBO9780511755545.

[KDH92] Rick A. Kendall, Thom H. Dunning, and Robert J. Harrison. “Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions”. In: *J. Chem. Phys.* 96 (1992), pp. 6796–6806. DOI: 10.1063/1.462569.

[Ker81] GP Kerker. “Efficient iteration scheme for self-consistent pseudopotential calculations”. In: *Physical Review B* 23.6 (1981), p. 3082.

[KKN07] Simone Kossmann, Barbara Kirchner, and Frank Neese. “Performance of modern density functional theory for the prediction of hyperfine structure: meta-GGA and double hybrid functionals”. In: *Mol. Phys.* 105.15-16 (2007), pp. 2049–2071.

[KMM96] C. Kittel, P. McEuen, and P. McEuen. *Introduction to solid state physics*. Vol. 8. Wiley New York, 1996.

[Kol+19] Christian Kollmar, Kantharuban Sivalingam, Benjamin Helmich-Paris, Celestino Angeli, and Frank Neese. “A perturbation-based super-CI approach for the orbital optimization of a CASSCF wave function”. In: *J. Comp. Chem.* 40.14 (2019), pp. 1463–1470.

[KR09] A. Kurganov and J. Rauch. “The order of accuracy of quadrature formulae for periodic functions”. In: *Advances in phase space analysis of partial differential equations*. Vol. 78. Progr. Nonlinear Differential Equations Appl. Birkhäuser Boston, MA, 2009, pp. 155–159. DOI: 10.1007/978-0-8176-4861-9_9. URL: https://doi.org/10.1007/978-0-8176-4861-9_9.

[KS65] W. Kohn and L. J. Sham. “Self-consistent equations including exchange and correlation effects”. In: *Phys. Rev.* 140 (4A 1965), A1133–A1138. DOI: 10.1103/PhysRev.140.A1133.

[KSC02] K.N. Kudin, G.E. Scuseria, and E. Cancès. “A black-box self-consistent field convergence algorithm: One step closer”. In: *The Journal of Chemical Physics* 116.19 (2002), p. 8255. DOI: 10.1063/1.1470195.

[Kut94] W Kutzelnigg. “Theory of the expansion of wave functions in a Gaussian basis”. In: *Int. J. Quantum Chem.* (1994).

[Lax07] P. D. Lax. *Linear algebra and its applications*. second. Pure and Applied Mathematics (Hoboken). Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2007, pp. xvi+376. ISBN: 978-0-471-75156-4.

[LBL05] Claude Le Bris and Pierre-Louis Lions. “From atoms to crystals: a mathematical journey”. In: *Bulletin of the American Mathematical Society* 42.3 (2005), pp. 291–363.

[Lee13] J.M. Lee. In: *Introduction to smooth manifolds*. Springer, 2013.

[Leh19a] Susi Lehtola. "A review on nonrelativistic, fully numerical electronic structure calculations on atoms and diatomic molecules." In: *International Journal of Quantum Chemistry* (2019).

[Leh19b] Susi Lehtola. "Communication: Curing basis set overcompleteness with pivoted Cholesky decompositions". In: *arXiv preprint arXiv:1911.10372* (2019).

[Leh19c] Susi Lehtola. "Fully numerical Hartree-Fock and density functional calculations. I. Atoms". In: *International Journal of Quantum Chemistry* (2019).

[Leh24] Susi Lehtola. "Importance profiles. Visualization of atomic basis set requirements". In: *Electronic Structure* (2024).

[Lev12] A. Levitt. "Convergence of gradient-based algorithms for the Hartree-Fock equations". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 46.6 (2012), pp. 1321–1336. DOI: 10.1051/m2an/2012008.

[Lev20] Antoine Levitt. "Mathematical and numerical analysis of models of condensed-matter physics". PhD thesis. Université Paris-Est, 2020.

[Lev79] Mel Levy. "Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem". In: *Proceedings of the National Academy of Sciences* 76.12 (1979), pp. 6062–6065.

[Lew04] Mathieu Lewin. "Solutions of the multiconfiguration equations in quantum chemistry". In: *Archive for rational mechanics and analysis* 171 (2004), pp. 83–114.

[Lew22] Mathieu Lewin. *Théorie spectrale et mécanique quantique*. Vol. 87. Springer, 2022.

[Li+23] Kangbo Li, Hsin-Yu Ko, Robert A DiStasio Jr, and Anil Damle. "An unambiguous and robust formulation for Wannier localization". In: *arXiv preprint arXiv:2305.09929* (2023).

[Lie02] Elliott H Lieb. "Density functionals for Coulomb systems". In: *Inequalities: Selecta of Elliott H. Lieb* (2002), pp. 269–303.

[Lin+24] Peize Lin, Xinguo Ren, Xiaohui Liu, and Lixin He. "Ab initio electronic structure calculations based on numerical atomic orbitals: Basic fomalisms and recent progresses". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 14.1 (2024), e1687.

[Liu+14] Xin Liu, Xiao Wang, Zaiwen Wen, and Yaxiang Yuan. "On the convergence of the self-consistent field iteration in Kohn–Sham density functional theory". In: *SIAM Journal on Matrix Analysis and Applications* 35.2 (2014), pp. 546–558.

[LL19] Lin Lin and Jianfeng Lu. *A mathematical introduction to electronic structure theory*. SIAM, 2019.

[LLS22] Mathieu Lewin, Elliott H Lieb, and Robert Seiringer. "Universal functionals in density functional theory". In: *Density Functional Theory: Modeling, Mathematical Analysis, Computational Methods, and Applications*. Springer, 2022, pp. 115–182.

[LMA18] G. Li Manni and A. Alavi. "Understanding the mechanism stabilizing intermediate spin states in Fe (II)-porphyrin". In: *The Journal of Physical Chemistry A* 122.22 (2018), pp. 4935–4947.

[Löw70] Per-Olov Löwdin. "On the nonorthogonality problem". In: *Advances in quantum chemistry*. Vol. 5. Elsevier, 1970, pp. 185–199.

[LS10] Mathieu Lewin and Éric Séré. "Spectral pollution and how to avoid it". In: *Proceedings of the London mathematical society* 100.3 (2010), pp. 864–900.

[LT72] G. Lehmann and M. Taut. "On the numerical calculation of the density of states and related properties". In: *Physica Status Solidi (b)* 54.2 (1972), pp. 469–477.

[LV1] Eric Cancès, Geneviève Dusson, Gaspard Kemlin, and Laurent Vidal. "On basis set optimisation in quantum chemistry". In: *ESAIM: Proceedings and Surveys* 73 (2023), pp. 107–129.

[LV2] Eric Cancès, Muhammad Hassan, and Laurent Vidal. "Modified-operator method for the calculation of band diagrams of crystalline materials". In: *Mathematics of Computation* (2023).

[LVp1] Laurent Vidal, Tommaso Nottoli, Filippo Lipparini, and Eric Cancès. "Geometric optimization of Restricted-Open and Complete Active Space Self-Consistent Field wavefunctions". *Submitted*.

[LVp2] Robert Benda, Eric Cancès, Emmanuel Giner, and Laurent Vidal. "Self-consistent field algorithms in Restricted Open-Shell Hatree-Fock". *Submitted*.

[Mar+12] Nicola Marzari, Arash A Mostofi, Jonathan R Yates, Ivo Souza, and David Vanderbilt. "Maximally localized Wannier functions: Theory and applications". In: *Reviews of Modern Physics* 84.4 (2012), p. 1419.

[Mar20] R. M. Martin. *Electronic structure: basic theory and practical methods*. Cambridge university press, 2020.

[Mat+20] Devin A Matthews et al. "Coupled-cluster techniques for computational chemistry: The CFOUR program package". In: *The Journal of Chemical Physics* 152.21 (2020).

[McW50] R. McWeeny. "Gaussian Approximations to Wave Functions". In: *Nature* 166.4209 (1950), pp. 21–22. DOI: 10.1038/166021a0.

[MI08] Koichi Momma and Fujio Izumi. "VESTA: a three-dimensional visualization system for electronic and structural analysis". In: *Journal of Applied crystallography* 41.3 (2008), pp. 653–658.

[MKW16] Filipe Menezes, Daniel Kats, and Hans-Joachim Werner. "Local complete active space second-order perturbation theory using pair natural orbitals (PNO-CASPT2)". In: *J. Chem. Phys.* 145 (2016), p. 124115. ISSN: 00219606. DOI: 10.1063/1.4963019.

[Mod] *Modified operator approach - numerics*. https://github.com/LaurentVidal95/ModifiedOp.

[Mor+18] W. S. Morgan, J. J. Jorgensen, B. C. Hess, and G. L. W. Hart. "Efficiency of generalized regular k-point grids". In: *Computational Materials Science* 153 (2018), pp. 424–430.

[MP76] H. J. Monkhorst and J. D. Pack. "Special points for Brillouin-zone integrations". In: *Physical Review. B. Condensed Matter. Third Series* 13.12 (1976), pp. 5188–5192. ISSN: 0163-1829.

[MP89] M. Methfessel and A. T. Paxton. "High-precision sampling for Brillouin-zone integration in metals". In: *Physical Review B* 40.6 (1989).

[MR18] P Mogensen and A Riseth. "Optim: A mathematical optimization package for Julia". In: *Journal of Open Source Software* 3.24 (2018).

[MRR90] Per Åke Malmqvist, Alistair Rendell, and Björn O Roos. "The restricted active space self-consistent-field method, implemented with a split graph unitary group approach". In: *J. Phys. Chem.* 94.14 (1990), pp. 5477–5482.

[MSS21] Quentin Mérigot, Filippo Santambrogio, and Clément Sarrazin. "Non-asymptotic convergence bounds for Wasserstein approximation using point clouds". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 12810–12821.

[MT94] Jorge J Moré and David J Thuente. "Line search algorithms with guaranteed sufficient decrease". In: *ACM Transactions on Mathematical Software (TOMS)* 20.3 (1994), pp. 286–307.

[MV97] Nicola Marzari and David Vanderbilt. "Maximally localized generalized Wannier functions for composite energy bands". In: *Physical review B* 56.20 (1997), p. 12847.

[MW20] Qianli Ma and Hans-Joachim Werner. "Scalable Electron Correlation Methods. 7. Local Open-Shell Coupled-Cluster Methods Using Pair Natural Orbitals: PNO-RCCSD and PNO-UCCSD". In: *J. Chem. Theory Comput.* 16.5 (2020), pp. 3135–3151. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.0c00192.

[Nee00] F. Neese. "Approximate second-order SCF convergence for spin unrestricted wavefunctions". In: *Chemical Physics Letters* 325.1-3 (2000), pp. 93–98.

[Net+09] AH Castro Neto, Francisco Guinea, Nuno MR Peres, Kostya S Novoselov, and Andre K Geim. "The electronic properties of graphene". In: *Reviews of modern physics* 81.1 (2009), p. 109.

[NGL21a] T. Nottoli, J. Gauss, and F. Lipparini. "Second-Order CASSCF Algorithm with the Cholesky Decomposition of the Two-Electron Integrals". In: *Journal of chemical theory and computation* 17.11 (2021), pp. 6819–6831.

[NGL21b] Tommaso Nottoli, Jürgen Gauss, and Filippo Lipparini. "A Black-Box, General Purpose Quadratic Self-Consistent Field Code with and without Cholesky Decomposition of the Two-Electron Integrals". In: *Molecular Physics* 119.21-22 (Nov. 2021), e1974590. ISSN: 0026-8976, 1362-3028. DOI: 10.1080/00268976.2021.1974590. arXiv: 2106.04836 [physics]. (Visited on 02/08/2023).

[NW99] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[Ols21] Jeppe Olsen. "An Introduction and Overview of Basis Sets for Molecular and Solid-State Calculations". In: *Basis Sets in Computational Chemistry*. Ed. by Eva Perlt. Cham: Springer International Publishing, 2021, pp. 1–16.

[Pag15] Gilles Pagès. "Introduction to vector quantization and its applications for numerics". en. In: *ESAIM: Proceedings and Surveys* 48 (Jan. 2015), pp. 29–79.

[Pan07] Gianluca Panati. "Triviality of bloch and bloch–dirac bundles". In: *Annales Henri Poincaré*. Vol. 8. Springer. 2007, pp. 995–1011.

[PBE96] J. P. Perdew, K. Burke, and M. Ernzerhof. "Generalized gradient approximation made simple". In: *Physical Review Letters* 77.18 (1996), p. 3865.

[PD14] B.N. Plakhutin and E.R. Davidson. "Canonical form of the Hartree-Fock orbitals in open-shell systems". In: *The Journal of Chemical Physics* 140.1 (Jan. 2014), p. 014102. DOI: 10.1063/1.4849615.

[Per21] Eva Perlt. *Basis Sets in Computational Chemistry*. Springer, 2021.

[Pes+23] Federica Pes, Etienne Polack, Patrizia Mazzeo, Genevieve Dusson, Benjamin Stamm, and Filippo Lipparini. "A Quasi Time-Reversible Scheme Based on Density Matrix Extrapolation on the Grassmann Manifold for BornOppenheimer Molecular Dynamics". In: *The Journal of Physical Chemistry Letters* 14.43 (2023). PMID: 37879072, pp. 9720–9726. DOI: 10.1021/acs.jpclett.3c02098. eprint: https://doi.org/10.1021/acs.jpclett.3c02098. URL: https://doi.org/10.1021/acs.jpclett.3c02098.

[PH88] Peter Pulay and Tracy P. Hamilton. "UHF natural orbitals for defining and starting MC-SCF calculations". In: *J. Chem. Phys.* 88 (1988), pp. 4926–4933. DOI: 10.1063/1.454704.

[Pha17] Dinh Huong Pham. "Galerkin method using optimized wavelet-Gaussian mixed bases for electronic structure calculations in quantum chemistry". en. PhD thesis. Université Grenoble Alpes, June 2017.

[Piz+20] Giovanni Pizzi et al. "Wannier90 as a community code: new features and applications". In: *Journal of Physics: Condensed Matter* 32.16 (2020), p. 165902.

[Po+18] Hoi Chun Po, Liujun Zou, Ashvin Vishwanath, and T Senthil. "Origin of Mott insulating behavior and superconductivity in twisted bilayer graphene". In: *Physical Review X* 8.3 (2018), p. 031089.

[PP13] Gianluca Panati and Adriano Pisante. "Bloch bundles, Marzari-Vanderbilt functional and maximally localized Wannier functions". In: *Communications in Mathematical Physics* 322 (2013), pp. 835–875.

[PP99] C. J. Pickard and M. C. Payne. "Extrapolative approaches to Brillouin-zone integration". In: *Physical Review B* 59.7 (1999).

[Pul80] P. Pulay. "Convergence acceleration of iterative sequences. the case of SCF iteration". In: *Chemical Physics Letters* 73.2 (1980), pp. 393–398. DOI: 10.1016/0009-2614(80)80396-4.

[Pul82] P. Pulay. "Improved SCF convergence acceleration". In: *Journal of Computational Chemistry* 3.4 (1982), pp. 556–560. DOI: 10.1002/jcc.540030413.

[Qbo] *QBOX: First Principles Molecular Dynamics Documentation*. http://qboxcode.org/doc/html/usage/variables.html#ecuts-var.

[Qia] J. Qiao. *Wannier.jl: A playground for experimentation with Wannier functions (WFs)*. https://github.com/qiaojunfeng/Wannier.jl.

[RLP16] Jarrett Revels, Miles Lubin, and Theodore Papamarkou. "Forward-mode automatic differentiation in Julia". In: *arXiv preprint arXiv:1607.07892* (2016).

[Roo51] C.C.J. Roothaan. “New Developments in Molecular Orbital Theory”. In: *Reviews of Modern Physics* 23.2 (1951), pp. 69–89. DOI: 10.1103/revmodphys.23.69.

[Roo60] C.C.J. Roothaan. “Self-consistent field theory for open shells of electronic systems”. In: *Reviews of Modern Physics* 32.2 (1960), p. 179.

[Roo80] Björn O Roos. “The complete active space SCF method in a Fock-matrix-based super-CI formulation”. In: *Int. J. Quant. Chem.* 18.S14 (1980), pp. 175–189.

[Roo87] Björn O. Roos. “The Complete Active Space Self-Consistent Field Method and its Applications in Electronic Structure Calculations”. In: *Adv. Chem. Phys.* 69 (1987), pp. 399–445. DOI: 10.1002/9780470142943.ch7.

[RS11] Thorsten Rohwedder and Reinhold Schneider. “An analysis for the DIIS acceleration method used in quantum chemistry calculations”. In: *Journal of mathematical chemistry* 49 (2011), pp. 1889–1914.

[RS72] Michael Reed and Barry Simon. *Methods of modern mathematical physics. I. Functional analysis*. Academic Press, New York-London, 1972, pp. xvii+325.

[RS78] M. Reed and B. Simon. *Methods of modern mathematical physics. IV. Analysis of operators*. Academic Press [Harcourt Brace Jovanovich publishers], New York-London, 1978, pp. xv+396. ISBN: 0-12-585004-2.

[Rui+98] Eliseo Ruiz, Joan Cano, Santiago Alvarez, and Pere Alemany. “Magnetic Coupling in End-On Azido-Bridged Transition Metal Complexes: A Density Functional Study”. In: *J. Am. Chem. Soc.* 120.43 (1998), pp. 11122–11129. ISSN: 0002-7863. DOI: 10.1021/ja981661n.

[Sce+16] A. Scemama, T. Applencourt, Y. Garniron, E. Giner, G. David, and M. Caffarel. “Quantum package v1. 0”. In: *Zenodo. http://dx. doi. org/10.5281/zenodo* 200970 (2016).

[Sch+93] M.W. Schmidt et al. “General atomic and molecular electronic structure system”. In: *Journal of Computational Chemistry* 14.11 (1993), pp. 1347–1363. DOI: 10.1002/jcc.540141112.

[Ser10] D. Serre. *Matrices*. Vol. 216. Graduate Texts in Mathematics. Theory and applications. Springer, New York, 2010, pp. xiv+289. ISBN: 978-1-4419-7682-6. DOI: 10.1007/978-1-4419-7683-3. URL: https://doi.org/10.1007/978-1-4419-7683-3.

[SH73] V.R. Saunders and I.H. Hillier. “A "Level-Shifting" method for converging closed shell Hartree-Fock wave functions”. In: *International Journal of Quantum Chemistry* 7.4 (1973), pp. 699–705. DOI: 10.1002/qua.560070407.

[Sha+15] Yihan Shao et al. “Advances in molecular quantum chemistry contained in the Q-Chem 4 program package”. In: *Molecular Physics* 113.2 (2015), pp. 184–215.

[Sha20] Robert A Shaw. “The completeness properties of Gaussiantype orbitals in quantum chemistry”. In: *Int. J. Quantum Chem.* 120.17 (Sept. 2020), p. 93.

[She87] Ron Shepard. “The Multiconfiguration Self-Consistent Filed Method”. In: *Adv. Chem. Phys.* 69 (1987), pp. 63–200.

[She+98] C David Sherrill, Anna I Krylov, Edward FC Byrd, and Martin Head-Gordon. “Energies and analytic gradients for a coupled-cluster doubles model using variational Brueckner orbitals: Application to symmetry breaking in O 4+”. In: *The Journal of chemical physics* 109.11 (1998), pp. 4171–4181.

[Sie+81] Per EM Siegbahn, Jan Almlöf, Anders Heiberg, and Björn O Roos. “The complete active space SCF (CASSCF) method in a Newton–Raphson formulation with application to the HNO molecule”. In: *J. Chem. Phys.* 74.4 (1981), pp. 2384–2396.

[Sim81] Barry Simon. “Spectrum and continuum eigenfunctions of Schrödinger operators”. In: *Journal of Functional Analysis* 42.3 (1981), pp. 347–355.

[Sla30] J C Slater. “Atomic shielding constants”. In: *Physical Review* (1930).

[SMS02] Robert K. Szilagyi, Markus Metz, and Edward I. Solomon. “Spectroscopic Calibration of Modern Density Functional Methods Using [CuCl4]2-”. In: *J. Phys. Chem. A* 106.12 (2002), pp. 2994–3007. ISSN: 1089-5639. DOI: 10.1021/jp014121c.

[SPAS95] Daniel Sanchez-Portal, Emilio Artacho, and Jose M Soler. "Projection of plane-wave calculations into atomic orbitals". In: *Solid State Communications* 95.10 (1995), pp. 685–690.

[SPAS96] Daniel Sánchez-Portal, Emilio Artacho, and José M Soler. "Analysis of atomic orbital basis sets from the projection of plane-wave results". In: *Journal of Physics: Condensed Matter* 8.21 (1996), p. 3859.

[SSI87] Gustavo E Scuseria and Henry F Schaefer III. "The optimization of molecular orbitals for coupled cluster wavefunctions". In: *Chemical physics letters* 142.5 (1987), pp. 354–358.

[Ste69] Robert F Stewart. "Small Gaussian Expansions of Atomic Orbitals". In: *J. Chem. Phys.* 50.6 (Mar. 1969), pp. 2485–2495.

[Sun+20] Qiming Sun et al. "Recent developments in the PySCF program package". In: *The Journal of chemical physics* 153.2 (2020).

[SY17] Stephan Scholz and Harry Yserentant. "On the approximation of electronic wavefunctions by anisotropic Gauss and Gauss–Hermite functions". In: *Numer. Math.* 136.3 (July 2017), pp. 841–874.

[Thø+04] L. Thøgersen, J. Olsen, D. Yeager, P. Jørgensen, P. Sałek, and T. Helgaker. "The trust-region self-consistent field method: Towards a black-box optimization in Hartree–Fock and Kohn–Sham theories". In: *The Journal of Chemical Physics* 121.1 (2004), p. 16. DOI: 10.1063/1.1755673.

[Tou22] Julien Toulouse. "Review of approximations for the exchange-correlation energy in density-functional theory". In: *Density Functional Theory: Modeling, Mathematical Analysis, Computational Methods, and Applications*. Springer, 2022, pp. 1–90.

[TP20] Zsuzsanna Tóth and Peter Pulay. "Comparison of Methods for Active Orbital Selection in Multiconfigurational Calculations". In: *J. Chem. Theory Comput.* 16.12 (2020), pp. 7328–7341. DOI: 10.1021/acs.jctc.0c00123. URL: https://pubs.acs.org/doi/10.1021/acs.jctc.0c00123.

[TPA89] Michael P Teter, Michael C Payne, and Douglas C Allan. "Solution of Schrödingers equation for large systems". In: *Physical Review B* 40.18 (1989), p. 12255.

[Tri+16] Georgios A Tritsaris et al. "Perturbation theory for weakly coupled two-dimensional layers". In: *Journal of Materials Research* 31.7 (2016), pp. 959–966.

[TS10] T. Tsuchimochi and G.E. Scuseria. "Communication: ROHF theory made simple". In: *The Journal of Chemical Physics* 133.141102 (2010).

[Tur+12] J.M. Turney et al. "Psi4: an open-source ab initio electronic structure program". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2.4 (2012), pp. 556–565.

[TV+01] G t Te Velde et al. "Chemistry with ADF". In: *Journal of Computational Chemistry* 22.9 (2001), pp. 931–967.

[TW14] L. N. Trefethen and J. A. C. Weideman. "The exponentially convergent trapezoidal rule". In: *SIAM Review* 56.3 (2014), pp. 385–458. ISSN: 0036-1445. DOI: 10.1137/130932132. URL: https://doi.org/10.1137/130932132.

[VHG02] T. Van Voorhis and M. Head-Gordon. "A geometric approach to direct minimization". In: *Molecular Physics* 100.11 (June 2002), pp. 1713–1721. DOI: 10.1080/00268970110103642.

[VT20] Pragya Verma and Donald G. Truhlar. "Status and Challenges of Density Functional Theory". In: *Trends Chem.* 2.4 (2020), pp. 302–318. ISSN: 2589-7209. DOI: 10.1016/j.trechm.2020.02.005.

[W90] *Wannier90: user guide*. https://wannier.org/support/.

[WA05] Florian Weigend and Reinhart Ahlrichs. "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy". In: *Physical Chemistry Chemical Physics* 7.18 (2005), pp. 3297–3305.

[Wer87] Hans-Joachim Werner. "Matrix-Formulated Direct Multiconfiguration Self-Consistent Field and Multiconfiguration Reference Configuration-Interaction Methods". In: *Adv. Chem. Phys.* 69 (1987), pp. 1–62. DOI: 10.1002/9780470142943.ch1.

[WH52] M. Wolfsberg and L. Helmholz. “The spectra and electronic structure of the tetrahedral ions MnO4-, CrO4–, and ClO4-”. In: *The Journal of Chemical Physics* 20.5 (1952), pp. 837–843.

[Whi72] H. Whitney. *Complex analytic varieties*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1972, pp. xii+399.

[YWL22] K. Ye, K. Wong, and L. Lim. “Optimization on flag manifolds”. In: *Mathematical Programming* 194.1 (2022), pp. 621–660.

[Zhi60] Grigorii Moiseevich Zhislin. “A study of the spectrum of the Schrodinger operator for a system of several particles”. In: *Trudy Moskovskogo Matematicheskogo Obshchestva* 9 (1960), pp. 81–120.

[ZS24] Xiaojing Zhu and Chungen Shen. “Practical gradient and conjugate gradient methods on flag manifolds”. In: *Computational Optimization and Applications* (2024), pp. 1–34.