

# Analyse numérique pour la théorie de la fonctionnelle de densité

Gaspard Kemlin

▶ **To cite this version:**

Gaspard Kemlin. Analyse numérique pour la théorie de la fonctionnelle de densité. Analyse numérique [math.NA]. École des Ponts ParisTech, 2022. Français. NNT : 2022ENPC0042 . tel-03941417

**HAL Id: tel-03941417**
**https://pastel.hal.science/tel-03941417**

Submitted on 16 Jan 2023

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

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

École des Ponts ParisTech

Inria

THÈSE DE DOCTORAT

de l'École des Ponts ParisTech

# Numerical analysis for Kohn–Sham density functional theory

École doctorale MSTIC

Discipline: Mathématiques Appliquées

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

Thèse soutenue le 15 décembre 2022, par

**Gaspard Kemlin**

**Composition du jury**

| | |
|---|---|
| Benjamin Stamm<br>Professeur, Université de Stuttgart | *Président du jury* |
| Daniel Peterseim<br>Professeur, Université d'Augsbourg | *Rapporteur* |
| Simona Rota Nodari<br>Professeure, Université Côte d'Azur | *Rapportrice* |
| Geneviève Dusson<br>Chargée de recherche, CNRS & Université Bourgogne Franche-Comté | *Examinatrice* |
| François Gygi<br>Professeur, Université de Californie Davis | *Examinateur* |
| Ingrid Lacroix-Violet<br>Professeure, Université de Lorraine & Polytech Nancy | *Examinatrice* |
| Eric Cancès<br>Professeur, École des Ponts & Inria | *Directeur de thèse* |
| Antoine Levitt<br>Chargé de recherche, École des Ponts & Inria | *Directeur de thèse* |

EMC² Extreme-scale Mathematically-based Computational Chemistry

erc European Research Council
Established by the European Commission

*This thesis has been written in LaTeX, using the Bredele class by Framatophe, available at https://framagit.org/Framatophe/Bredele.*

# Résumé

La simulation moléculaire et le calcul de structures électroniques sont des outils fondamentaux utilisés en chimie, physique de la matière condensée, biologie moléculaire, science des matériaux, nanosciences... La théorie de la fonctionnelle de densité (DFT) est aujourd'hui une des méthodes les plus utilisées, car elle offre un bon compromis entre efficacité et précision. Il s'agit d'un problème formidable qui nécessite toute une hiérarchie de choix entraînant un certain nombre d'approximations et d'erreurs associées : choix du modèle, choix de la base de discrétisation, choix des solveurs, erreur de troncature, erreur numérique... Cette thèse traite certains de ces problèmes, du point de vue de l'analyse numérique, et porte une attention particulière à la simulation de cristaux et autres systèmes périodiques avec DFTK, un récent logiciel de DFT en `Julia`.

Les premiers chapitres de ce manuscrit concernent l'analyse asymptotique d'algorithmes utilisés en calcul de structures électroniques et l'estimation d'erreurs. Dans le premier chapitre, nous analysons et comparons la structure algébrique de deux classes d'algorithmes : les algorithmes de minimisation directe et les algorithmes de champ auto-cohérent. Ce cadre commun permet de dériver des taux de convergence asymptotiques pour ces algorithmes et nous analysons leur dépendance en fonction du trou spectral et d'autres propriétés du problème. Le second chapitre tire profit de la structure algébrique étudiée dans le premier chapitre pour proposer des estimateurs d'erreur pour les équations de Kohn–Sham : le caractère non linéaire de ces équations rend difficile l'obtention de tels estimateurs et la stratégie proposée dans cette thèse consiste à linéariser les équations de Kohn–Sham pour obtenir une relation entre l'erreur et le résidu, que l'on peut ensuite inverser de façon efficace, sous des approximations raisonnables. En particulier, cette méthode est utilisée pour obtenir des estimateurs sur des quantités d'intérêt comme les forces interatomiques, dont la dérivation faisait défaut jusqu'à présent. Un autre chapitre est consacré à la conception et l'implémentation de méthodes de calculs pour les propriétés de réponse des matériaux, dans l'objectif de les rendre plus stables et rapides. Nous y décrivons un cadre commun, dans lequel entre la plupart des méthodes existantes dans la littérature et justifions son intérêt par une analyse de stabilité. Nous proposons également une nouvelle méthode de résolution de l'équation de Sternheimer, pierre centrale du calcul de réponse, qui réduit de façon significative le temps de calcul.

Le reste de ce manuscrit est composé de travaux menés en parallèle de la première partie. Un chapitre traite de la régularité des solutions d'équations de Schrödinger périodiques. Nous y étendons des résultats antérieurs au cas de potentiels analytiques et prouvons (dans le cas linéaire) la convergence exponentielle avec la taille de la base de discrétisation. Enfin, dans le dernier chapitre, fruit de travaux menés pendant l'école d'été du CEMRACS 2021, nous proposons des critères généraux pour construire des bases atomiques localisées optimales en chimie quantique.

## Summary

Molecular simulation and electronic structure calculation are fundamental tools used in chemistry, solid-state physics, molecular biology, materials science, nanosciences... Density functional theory (DFT) is one of the most widely used methods nowadays, as it offers a good compromise between efficiency and accuracy. It is a formidable problem that requires a whole hierarchy of choices, which lead to a number of approximations and associated errors: choice of model, choice of discretization basis, choice of solvers, truncation error, numerical error... This thesis deals with some of these problems, from a numerical analysis point of view, and pays particular attention to the simulation of crystals and other periodic systems with DFTK, a recent DFT software in `Julia`.

The first chapters of this manuscript concern the asymptotic analysis of algorithms used in electronic structure calculation and error estimation. In the first chapter, we analyse and compare the algebraic structure of two classes of algorithms: direct minimization algorithms and self-consistent field algorithms. This common framework allows us to derive asymptotic convergence rates for these algorithms and we analyse their dependence on the spectral gap and other properties of the problem. The second chapter takes advantage of the algebraic structure studied in the first chapter to propose error estimators for the Kohn–Sham equations: the nonlinear nature of these equations makes it difficult to obtain such estimators and the strategy proposed in this thesis consists in linearizing the Kohn–Sham equations to obtain a relation between the error and the residual, which can then be efficiently inverted, under reasonable approximations. In particular, this method is used to obtain estimators for quantities of interest such as interatomic forces, the derivation of which has been lacking until now. Another chapter is devoted to the design and implementation of methods for calculating the response properties of materials, with the aim of making them more stable and fast. We describe a common framework, in which most of the existing methods in the literature fit, and justify its interest by a stability analysis. We also propose a new method for solving the Sternheimer equation (the cornerstone of response calculations) which significantly reduces the computational time.

The rest of this manuscript is composed of works carried out in parallel to the first part. One chapter deals with the regularity of solutions to periodic Schrödinger equations. We extend previous results to the case of analytic potentials and prove (in the linear case) the exponential convergence with the size of the discretization basis. Finally, in the last chapter, resulting from works carried out during the CEMRACS 2021 summer school, we propose general criteria for constructing optimal localized atomic bases in quantum chemistry.

# Remerciements

À quelques jours de la soutenance de cette thèse et après avoir corrigé la 1 856$^{\text{ème}}$ coquille, il est temps pour moi de remercier toutes les personnes qui m'ont permis de (sur)vivre pendant ces trois dernières années, entre thèse, bugs et COVID.

Tout d'abord, j'aimerais remercier mes directeurs de thèse, Eric et Antoine, de m'avoir fait confiance il y a presque quatre ans, au moment de se lancer dans l'aventure de la thèse. Quand je regarde en arrière, je revois le jeune étudiant en M2 qui hésite entre différents sujets tous aussi intéressants et, aujourd'hui, je réalise la chance que j'ai eu de travailler sous votre encadrement pendant ces années : vous avez su m'enseigner la profondeur des liens entre mathématiques, analyse numérique et calcul de structures électroniques, tout en prenant le temps de répondre à mes questions (même les plus naïves), de m'aider dans la recherche de bugs ou de me suggérer des idées quand je me retrouvais dans des impasses. Vous m'avez également appris à ne pas renoncer à comprendre le moindre détail, tant pour des convergences aux comportements un peu surprenant que pour des concepts mathématiques. Merci pour votre disponibilité sans failles, même dans des périodes chargées, et votre accompagnement pendant les confinements successifs : si j'ai décidé de tenter ma chance en postdoc à la suite de cette thèse malgré les contraintes spatio-temporelles que cela implique, c'est aussi grâce à vous et vos conseils quand il a fallu réfléchir à la suite.

Je voudrais ensuite remercier grandement Daniel Peterseim et Simona Rota Nodari d'avoir accepté de rapporter ma thèse, j'en suis très honoré. Je suis également très reconnaissant à Ingrid Lacroix-Violet et François Gygi pour avoir accepté de faire partie du jury (même depuis la lointaine Californie). François, merci aussi pour les discussions à l'IPAM.

Cette thèse n'aurait jamais pris la forme qu'elle a aujourd'hui sans les collaborations dont elle a fait l'objet. Je pense en particulier à Michael, à qui j'aimerais adresser mes plus sincères remerciements. Tout d'abord, pour avoir donné naissance, avec Antoine, à DFTK : je ne sais pas si nous aurions pu aller aussi loin dans certains chapitres sans un logiciel de cette qualité, à l'interface entre mathématiques, calcul scientifique et physique de la matière condensée. Merci aussi pour les conseils en code et toutes les explications de concepts de chimie au début de la thèse, quand je découvrais ces domaines nouveaux pour moi. Plus généralement, merci à toi et Carine pour les visites de musées, les welshs, l'accueil à Aachen, la découverte de Bohnanza… En parlant de DFTK, merci à Niklas et Markus pour leurs contributions à la différentiation automatique dans DFTK ainsi que ce chouette DFTK workshop à Aachen. Je voudrais aussi remercier Geneviève pour nos précieuses collaborations et pour tes nombreux conseils et encouragements sur la suite : c'est toujours un plaisir de discuter avec toi. Merci également à Ben pour ces deux mois passés à Aachen, la disponibilité dont tu as fait preuve pour qu'on puisse travailler ensemble et tes nombreux conseils, tant sur le plan scientifique que sur la vie de jeune chercheur en général. J'ai hâte d'arriver à Stuttgart pour commencer notre collaboration. Ben et Geneviève, je suis aussi très heureux que vous fassiez partie de mon jury. Enfin, merci Laurent pour les résultats prometteurs de cette belle collaboration du CEMRACS 2021, et je voudrais terminer en remerciant Susi Lehtola pour l'attention qu'il a porté à ces travaux.

Je souhaite maintenant remercier tou·te·s les permanent·e·s du CERMICS et des Ponts qui m'ont accompagné, d'abord pendant mes études, puis pendant la thèse. Merci G d'avoir préparé avec moi mon semestre à l'ETH puis de m'avoir fait confiance pour prendre en charge une petite classe pendant deux ans. Merci également à Frédéric et Frédéric pour m'avoir respectivement confié des TDs d'optimisation et d'éléments finis, même de façon ponctuelle. L'enseignement est une dimension de la thèse et de la recherche que j'apprécie particulièrement, je vous suis donc très reconnaissant de m'avoir donné l'occasion de la découvrir. Je voudrais aussi remercier Virginie pour les bières au CEMRACS et les cours sous les arbres du CIRM, Tony pour les cours de master et les conseils en postdoc, Julien pour le coup de main un soir de galère à vélo devant les Ponts, Vincent pour le séminaire jeux (qui n'a malheureusement pas survécu aux confinements), Jean-François pour la mise en place de la mobilité Erasmus en IMI (et accessoirement pour m'avoir appris que $L^\infty$ était isomorphe au dual de $L^1$ *à condition que la mesure soit*

*σ-finie*), Jean-Philippe pour le soutien informatique et Urbain pour tes nombreux conseils sur l'entrée dans le monde académique. Je remercie tout particulièrement Isabelle et Stéphanie, le super binôme de secrétaires du CERMICS, grâce à qui tous nos problèmes trouvent toujours une solution. Merci de toujours nous défendre auprès de l'administration, de nous accompagner et nous conseiller dans les tâches administratives. Enfin, j'aimerais remercier Sandrine pour tout ce qu'elle fait pour les élèves d'IMI : de notre première rencontre dans ton bureau parce que j'avais pris un cours incompatible avec un parcours IMI à nos nombreux repas au Descartes, c'est toujours un plaisir de venir papoter avec toi. Merci aussi à Mohammed pour m'avoir accompagné tout au long de mon parcours aux Ponts.

Je souhaiterais également remercier mes encadrants de stages entre le M1 et le M2 pour avoir guidé mes premiers pas dans la recherche. Merci donc à Frank Hülsemann et Jérôme Bonelle pour avoir accompagné, à EDF, ma découverte de l'algèbre linéaire numérique et d'un code de calcul industriel. Merci aussi à Antoine Rousseau, de l'Inria, pour ses nombreux enseignements en analyse numérique et son accompagnement dans ce moment charnière de la vie d'un étudiant.

La vie de thésard, c'est aussi ~~devoir supporter~~ les autres doctorant·e·s du labo. La bonne ambiance qui règne au CERMICS est, je pense, un gros point fort pour avancer même dans les moments les plus difficiles (bien qu'un étage soit clairement mieux que l'autre). Merci d'abord aux ancien·ne·s pour m'avoir chaleureusement accueilli dans leur maison. Je ne me risquerai pas à l'exercice de citer tous vos noms, donc si tu le cherches et que je suis arrivé en thèse au milieu de la tienne, c'est ici que je pense à toi. Merci à Rémi, Inass et Michel pour les apéros confinement et le partage de recettes de pizzas. Merci à Cyrille et Rutger pour vos délicieux gâteaux. Merci à la team pause (dont je tairai le nom des membres pour ne pas balancer) de toujours venir nous rappeler qu'il est bon de s'aérer l'esprit de temps en temps. Merci Epiphane, Edoardo et Eloise d'avoir défendu avec bravoure les murs de la cité de Provins. Merci Laurent et Alfred pour vos sketchs improvisés. Merci Solal, en qui j'ai enfin trouvé quelqu'un avec qui parler rugby, dommage que l'interface entre nos thèses soit si courte. Merci aussi aux postdocs de la team quantique, Etienne et Louis. Pour toutes les discussions autour d'un café ou d'un repas à la cantine et pour tout le reste, merci à Hervé, Léo, Noé, Andrea, Louis, Jean, Louis-Pierre, Emanuele, Shiva, Roberta, Zoé, Maël, Fabian, Clément, Hélène, Kacem, Albéric, Nathan, Camila, Carlos, Stefano, Thomas, Vitor, Mohamad, Simon, Coco, Renato, Morgane, Nerea et Julien. Milles mercis à nos représentants : Guillaume pour avoir eu ce difficile rôle en pleine pandémie et ton site plein de ressources, puis Régis pour avoir pris la succession et organisé avec brio la première édition des JSJC. Enfin, j'aimerais réitérer mes remerciements à toute la team quantique qui n'a pas cessé de grandir depuis 3 ans : Laurent, Eloise, Alfred, Etienne, Louis et plus récemment Solal et Andrea, quel bonheur d'avoir grandi à vos côtés.

J'ai cité les camarades de thèse du labo, mais je n'oublie pas Ioanna-Maria, Hassan, Siwar, Matthias, Rémi, Matthieu, Agustin, Jesus, Alexiane et Ramon que j'ai toujours plaisir à croiser, au LJLL ou ailleurs. Merci aussi à toute la team CEMRACS 2021 pour nos longues soirées au milieu des calanques et des sangliers. Francesco, depuis notre rencontre en stage à EDF, on suit des chemins similaires dans des pays différents, mais merci pour toutes nos discussions sur le monde, la politique ou les différences de fromages entre la France et l'Italie. Enfin, merci à Danh pour tout, des vacances dans les Pyrénées aux discussions sur les maths et la vie par téléphone pendant les confinements.

Merci également à Mi-Song pour sa contribution à l'étude anatomique des poissons de la Méditerranée, la découverte des *atomic chess* ou (de temps en temps) les maths[1]. J'aimerais en profiter pour remercier de façon plus générale la petite communauté française qui orbite autour du groupe EMC2, à l'interface des maths, du calcul scientifique et du calcul de structures électroniques. Que cela soit pendant les cours des mini-school du GDR NBODY, en conférence ou autour d'un café, j'ai aussi beaucoup appris pendant ces quelques années à vos côtés.

Enfin, d'un point de vue scientifico-professionnel, si je devais remercier une figure des mathématiques, je choisirais sans hésiter Issai Schur pour son complément. J'en profite pour remercier au passage toutes les personnes qui contribuent aux logiciels qui rythment ma vie depuis tant d'années, avec une pensée particulière pour les contributeur·rice·s de `Julia`, LaTeX et Neovim.

Aussi saugrenu que cela puisse paraître, il existe (heureusement) une vie à côté de la thèse et j'aimerais à présent remercier toutes les belles personnes qui ont, indirectement ou pas, participé au succès de celle-ci. Merci d'abord à mes colocs, Jonas, Camille, Fred et Clément. La vie en coloc n'est pas une chose facile, mais je trouve que, malgré nos désaccords, on s'en est plutôt bien tiré. Merci pour tous ces

[1] Tu savais, toi, que c'est le carré de la Gaussienne qui est normalisé ?

moments partagés, ces repas, ces soirées, ces jeux et j'en passe, grâce auxquels je garderai un excellent souvenir de ces quelques années de vie commune. Merci à Nath pour la modération du chat. Merci Krumpy et Milouch pour les soirées jeux, j'espère qu'on pourra aller au festival d'Essen une fois installé à Stuttgart ! Merci aussi Roxane et Justin pour les jeux vidéos, les concerts ou les discussions sur vos dernières lectures. Merci Cédric pour tes concerts et m'avoir ouvert au monde du rap. Merci Alice pour les pogos. Merci Solène et Quentin pour cette belle année de prépa en trinôme. Un de mes principaux regrets ces dernières années est qu'on ne se soit pas vu plus souvent. Je remercie également tous les potes des Ponts : c'est un plaisir de continuer à vous croiser, en concert ou dans un bar à jeux. Merci aux copain·e·s du lycée, on ne se voit pas aussi souvent que dans le temps, mais c'est toujours chouette de vous retrouver le temps d'une soirée ou d'un pique-nique. Merci à MLF pour m'avoir introduit au monde des murders et fait découvrir une face de moi-même que je ne connaissais pas. Merci aussi à tou·te·s les camarades rencontré·e·s ces dernières années qui luttent pour un monde meilleur.

Toulousain en exil en Île-de-France, j'ai eu la chance d'être régulièrement accueilli dans deux familles que je souhaiterais remercier infiniment pour leur aide, tant pendant des galères sans appartement que pour survivre à des canicules insupportables sous les toits parisiens. Caroline et Thomas, merci pour toutes ces fois où vous m'avez accueilli chez vous, le temps d'un repas, d'une fête ou d'un déménagement. Cette thèse vous doit beaucoup. Merci aussi à Nathalie et Dominique, c'est toujours une joie de venir dîner chez vous ou de partir en balade. Merci aussi à Hervé d'avoir été un super voisin pendant mon année parisienne.

Parmi les gens qui nous entourent, la famille occupe une place importante et j'ai l'immense chance d'être entouré de personnes qui ont pu me conseiller et m'accompagner pendant ces trois années. Merci Vincent et Cécile pour vos conseils au moment de se lancer dans la thèse. Merci Cyrille, Laura et le petit Marius pour tous ces moments partagés. Merci Lali pour tous nos repas à Jussieu. Merci Aline pour ces beaux week-ends à Bouffémont. Merci Gene et Chris pour vos conseils et votre accueil à Montpellier. Même s'iels ne sont plus là aujourd'hui, j'aimerais finir en ayant une pensée pour mes grands-parents, qui m'ont enseigné les bases de la vie, comme le Scrabble, le clafoutis aux pommes, l'apéro ou le décollé de Citroën Saxo.

Je ne peux évidemment pas conclure ces ~~lignes~~ pages sans remercier mes parents. Déjà, pour m'avoir donné naissance, condition nécessaire à toute personne qui veut se lancer dans la vie. Puis pour m'avoir guidé sur le chemin sinueux qu'elle représente, malgré des allergies pas simples à gérer. Et ce, en étant toujours là pour m'aider lors des choix difficiles (d'autant plus qu'il n'y en pas de mauvais) mais aussi pour profiter des moments plus joyeux ! Merci aussi pour ces 3 mois inattendus de vie commune entre mars et juin 2020. Quant à mes frères, on suit chacun notre route dans trois directions orthogonales : c'est ce qui fait notre force. Merci pour les parties de belote, les rires, les randos et ce que vous êtes.

Lucile, que dire que tu ne sais déjà ? Merci d'être là. Merci de me supporter, moi et mes blagues, et d'être toujours prête à m'accompagner, pour aller attraper le Ronflex au coin de la rue comme pour partir à l'assaut des ballons des Vosges. Merci pour ce beau chemin parcouru ensemble, entre la campagne berrichonne, les montagnes pyrénéennes, Amiens et Noisy. J'espère que nous continuerons à le tracer encore longtemps.

Merci aussi à toi qui a lu ces remerciements jusqu'au bout.

# Résumé détaillé

## Un peu d'histoire

Traditionnellement, les mathématiques appliquées ont joué un rôle fondamental dans les sciences de l'ingénieur, telles que la dynamique des fluides, la mécanique ou l'électromagnétisme. En effet, les méthodes numériques sous-jacentes ont été analysées en profondeur et reposent sur des bases théoriques rigoureuses. Par exemple, la méthode des éléments finis (FEM), inventée au milieu du 20e siècle par des ingénieurs, a ensuite été considérablement améliorée avec l'aide de mathématicien·ne·s appliqué·e·s et elle n'en serait pas au stade actuel sans l'intervention des mathématiques. Cependant, pour des raisons historiques, cela a rarement été le cas en chimie computationnelle, en physique des solides ou en science des matériaux.

Cela est quelque peu surprenant car les problèmes mathématiques dérivés de l'équation de Schrödinger, pierre angulaire de nombreux modèles dans ces domaines, sont souvent des problèmes aux valeurs propres qui doivent être discrétisés pour que des solutions numériques puissent être calculées. Bien que la discrétisation des problèmes aux valeurs propres (éventuellement non linéaires) soit un sujet bien établi en mathématiques appliquées, il existe peu d'interactions avec la chimie computationnelle, la physique des solides et la science des matériaux dans la littérature. Le tableau suivant montre le nombre d'occurrences sur les bases de données Google Scholar et MathSciNet, pour les mots-clés "Navier–Stokes" et "density functional theory" :

| Mot-clé: | Google Scholar | MathSciNet |
|---|---:|---:|
| "Navier–Stokes" | 955,000 | 11,744 |
| "density functional theory" | 1,900,000 | 227 |

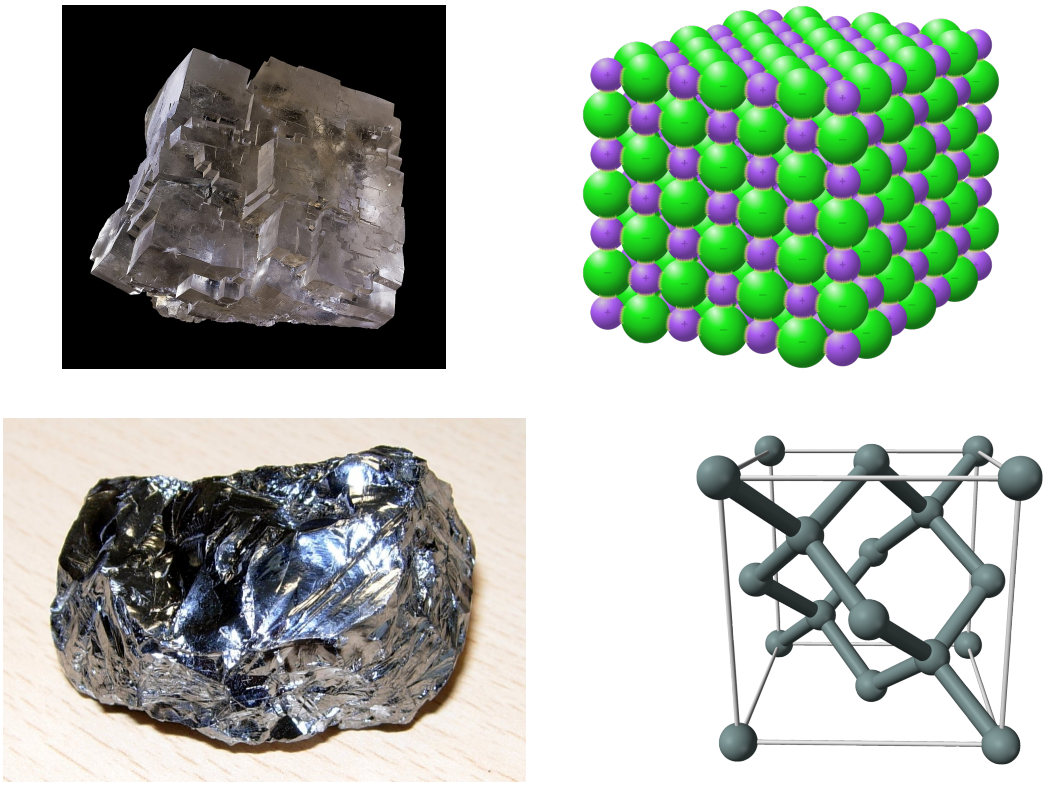
Table 1 – "Navier–Stokes" *vs* "density functional theory", au 1er Janvier 2022.

Si l'on considère ces données comme une mesure de la pertinence de ces deux domaines de recherche, il apparaît qu'il existe une déconnexion au niveau des tendances entre la communauté scientifique et son sous-ensemble de mathématicien·ne·s. Outre les chiffres impressionnants de Google Scholar, l'importance de la chimie computationnelle et du calcul de structures électroniques est confirmée par le fait que Kohn et Pople ont reçu le prix Nobel de chimie en 1998 pour leur contribution à la théorie de la fonctionnelle de densité ("density functional theory", DFT) et que les travaux fondateurs sur les méthodes multiéchelles (modèles de champ de force QM/MM) de Karplus, Levitt et Warshel ont été récompensés par le même prix en 2013. Aujourd'hui, la chimie computationnelle est pleinement considérée comme un troisième pilier de la chimie, aux côtés de la chimie expérimentale et théorique. Comme autre indicateur de l'importance de la DFT dans la science moderne, 12 des 100 articles les plus cités s'y réfèrent. En particulier, deux d'entre eux sont dans le top 10 et consistent en des "recettes techniques sur lesquelles sont construites les méthodes et logiciels de DFT les plus populaires" (parmi les articles du Web of Science de Thomson Reuter, de 1900 à 2014 [195]). Comme cela a été le cas dans le passé pour la FEM, le domaine du calcul de structures électroniques a bénéficié ces vingt dernières années du travail de mathématicien·ne·s du monde entier, qui ont analysé les méthodes existantes et développé des outils mathématiques pour améliorer les aspects numériques du calcul de structures électroniques et de la chimie quantique. Cette thèse vise à contribuer à ces améliorations.

# Description détaillée des chapitres

## Chapitre 1

Le premier chapitre a vocation à introduire les concepts nécessaires à la compréhension de ce manuscrit. Après une brève introduction historique et une présentation du contexte général dans les Sections 1.1 et 1.2, nous présentons dans la Section 1.3 le cadre mathématique nécessaire au calcul de l'état fondamental de systèmes moléculaires généraux, avec un accent particulier sur la DFT de Kohn–Sham. Il est volontairement bref par souci de clarté et lae lecteur·rice intéressé·e est invité·e à consulter les références fournies pour plus de détails sur les différents sujets, en particulier les ouvrages suivants : [29, 42, 133]. Dans la Section 1.4, nous nous concentrons sur le cadre de la DFT en ondes planes, qui utilise une discrétisation en modes de Fourier des objets que nous étudions. Nous introduisons également dans cette section des approximations utiles pour la DFT en ondes planes : l'approximation dite des *pseudopotentiels*, qui a motivé les résultats du Chapitre 5, et l'échantillonnage de la zone de Brillouin pour les opérateurs périodiques de type Schrödinger. Ces concepts sont utiles pour comprendre les systèmes que nous étudions dans cette thèse : la plupart des simulations portent sur des systèmes cristallins, qui ont une structure périodique intrinsèque qui se prête bien à la discrétisation en ondes planes. Ces simulations sont réalisées avec le paquet `Julia` DFTK, que nous présentons dans la Section 1.4.4.



Source: *Wikipedia Commons*

FIGURE 1 – Exemples de matériaux étudiés dans cette thèse et ayant une structure périodique : (haut) un morceau de cristal de chlorure de sodium, communément appelé sel, et sa structure cristalline – les atomes de sodium sont en violet et les atomes de chlore en vert –, (bas) un morceau de cristal de silicium purifié, un système simple pour tester les méthodes numériques, et sa structure cristalline.

Dans la Section 1.5, nous décrivons la résolution des équations découlant des sections précédentes, en mettant l'accent sur deux classes d'algorithmes : les algorithmes de minimisation directe et les algorithmes de champ auto-cohérent. Nous présentons ensuite les résultats du Chapitre 2, où ces deux classes sont analysées et comparées. Dans la Section 1.6, nous passons en revue la littérature existante sur les estimations d'erreur pour les simulations numériques et le calcul de structures électroniques. Ensuite, nous discutons les résultats du Chapitre 3, où des estimateurs d'erreurs sont développés dans le cadre de la DFT en ondes planes. Enfin, nous considérons dans la Section 1.7 le cadre de la DFPT ("density

functional perturbation theory"), qui a pour objectif de calculer les dérivées de la densité électronique de l'état fondamental par rapport à des perturbations externes, et nous présentons les contributions du Chapitre 4.

Enfin, nous soulignons que le Chapitre 6 est le résultat d'un projet mené à l'école d'été du CEMRACS 2021[2]. Bien qu'il soit lié à la chimie quantique, ce chapitre traite de l'optimisation de bases, qui n'est pas du ressort de la DFT en ondes planes. Il n'est donc pas mentionné dans le chapitre introductif puisqu'il est relativement auto-contenu.

## Chapitre 2

Dans le Chapitre 2, nous étudions des problèmes de minimisation exprimés de façon générale sous la forme suivante :

$$\min_{P \in \mathcal{M}_{N_{\mathrm{el}}}} E(P),$$

où

$$\mathcal{M}_{N_{\mathrm{el}}} = \left\{ P \in \mathcal{H},\ P = P^T,\ \mathrm{Tr}(P) = N_{\mathrm{el}},\ P^2 = P \right\}.$$

$\mathcal{H} = \mathbb{R}^{N_{\mathrm{b}} \times N_{\mathrm{b}}}$ est muni du produit scalaire de Frobenius $\langle A, B \rangle_{\mathrm{F}} = \mathrm{Tr}(A^T B)$, l'extension au cas des matrices complexes étant immédiate. De plus, $\mathcal{M}_{N_{\mathrm{el}}}$ est une variété Riemannienne, dont les éléments sont appelés en chimie quantique des *matrices densités*, difféomorphe à la variété de Grassmann $\mathrm{Grass}(N_{\mathrm{el}}, N_{\mathrm{b}})$ : nous définissons alors $\mathcal{T}_P \mathcal{M}_{N_{\mathrm{el}}}$, le plan tangent à $\mathcal{M}_{N_{\mathrm{el}}}$ au point $P$. Ce cadre convient parfaitement au calcul de l'état fondamental d'un système moléculaire ou cristallin pour des modèles de type Kohn–Sham DFT, tel que décrit dans le Chapitre 1 : ici, $N_{\mathrm{el}}$ correspond aux nombres d'électrons du système étudié et $N_{\mathrm{b}}$ à la taille de la base de discrétisation choisie.

Nous supposons par ailleurs que la fonctionnelle $E$ est suffisamment régulière et qu'il existe un minimum non dégénéré $P_*$ : il existe une constante $\eta > 0$ telle que

$$E(P) \geqslant E(P_*) + \eta \|P - P_*\|_{\mathrm{F}}^2 \quad \text{pour } P \text{ dans un voisinage de } P_*.$$

Nous dérivons alors des conditions d'optimalité du premier et second ordre :

- La condition du premier ordre s'écrit $\boldsymbol{\Pi}_{P_*} H(P_*) = 0$, où $H(P) = \nabla E(P)$ est l'Hamiltonien du système et $\boldsymbol{\Pi}_P$ la projection orthogonale sur $\mathcal{T}_P \mathcal{M}_{N_{\mathrm{el}}}$.
- La condition du second ordre est obtenue par linéarisation et s'écrit

  $$\forall\ X \in \mathcal{T}_{P_*} \mathcal{M}_{N_{\mathrm{el}}}, \quad \langle X, (\boldsymbol{\Omega}_* + \boldsymbol{K}_*) X \rangle_{\mathrm{F}} \geqslant \eta \|X\|_{\mathrm{F}}^2,$$

  où $\boldsymbol{K}_* = \boldsymbol{\Pi}_{P_*} \nabla^2 E(P_*) \boldsymbol{\Pi}_{P_*}$ est l'Hessienne de l'énergie projetée sur $\mathcal{T}_{P_*} \mathcal{M}_{N_{\mathrm{el}}}$ et

  $$\forall\ X \in \mathcal{T}_{P_*} \mathcal{M}_{N_{\mathrm{el}}}, \quad \boldsymbol{\Omega}_* X = -[P_*, [H(P_*), X]]$$

  représente l'influence de la courbure de $\mathcal{M}_{N_{\mathrm{el}}}$. Cet opérateur est étudié de façon plus approfondie dans le Chapitre 2 et cette condition traduit la non dégénérescence du minimiseur $P_*$. En effet, pour un problème d'optimisation sans contraintes et non dégénéré, la condition du second ordre se lit dans le caractère défini positif de l'Hessienne de la fonction objectif. Cette condition est ici modifiée par les contraintes.

À l'aide de ces deux conditions, nous étudions la convergence des plus simples représentants de deux classes distinctes d'algorithmes : les algorithmes de minimisation directe (représentés par une simple descente de gradient projetée) et les algorithmes de champ auto-cohérent (représentés par un algorithme SCF – "Self-consistent field" – amorti), où des problèmes aux valeurs propres sont successivement résolus jusqu'à convergence. La première classe d'algorithmes est basée sur une minimisation directe de l'énergie sur la variété $\mathcal{M}_{N_{\mathrm{el}}}$ : il s'agit d'une descente de gradient contrainte à rester sur la variété à l'aide d'un

[2] http://smai.emath.fr/cemracs/cemracs21/

opérateur de rétraction $R$. La seconde classe d'algorithmes est basée sur l'interprétation des équations d'Euler–Lagrange du problème de minimisation, qui s'écrivent sous la forme suivante :

$$\begin{cases} H(P_*)\phi_n = \varepsilon_n \phi_n, \ \varepsilon_1 \leqslant \cdots \leqslant \varepsilon_{N_{\mathrm{el}}} \\ \phi_n^* \phi_m = \delta_{nm}, \\ P_* = \displaystyle\sum_{n=1}^{N_{\mathrm{el}}} \phi_n \phi_n^* \in \mathcal{M}_{N_{\mathrm{el}}}, \end{cases}$$

où, à nouveau, $H(P) = \nabla E(P)$. Il s'agit d'un problème aux valeurs propres non linéaire de par la nature auto-cohérente de ces équations : pour construire l'opérateur $H(P_*)$ que l'on souhaite diagonaliser, il faut déjà connaître ses vecteurs propres afin de pouvoir construire la matrice densité $P_*$. L'algorithme SCF amorti est alors obtenu en introduisant un paramètre de *damping* dans un algorithme de point fixe standard. Ces deux classes d'algorithmes sont brièvement présentées ci-dessous.

ALGORITHME – Descente de gradient projetée

**Data:** $P^0 \in \mathcal{M}_{N_{\mathrm{el}}}$
**while** *convergence non atteinte* **do**
  $P^{k+1} = R\big(P^k - \beta \mathbf{\Pi}_{P^k}\big(\nabla E(P^k)\big)\big)$ ;
**end**

ALGORITHME – SCF amorti

**Data:** $P^0 \in \mathcal{M}_{N_{\mathrm{el}}}$
**while** *convergence non atteinte* **do**
  résoudre $\begin{cases} H(P^k)\phi_n^k = \varepsilon_n^k \phi_n^k, \ \varepsilon_1^k \leqslant \cdots \leqslant \varepsilon_{N_{\mathrm{el}}}^k < \varepsilon_{N_{\mathrm{el}}+1}^k \\ (\phi_n^k)^* \phi_m^k = \delta_{nm}, \end{cases}$ ;
  $\widetilde{P}^k = \displaystyle\sum_{n=1}^{N_{\mathrm{el}}} \phi_n^k (\phi_n^k)^*$;
  $P^{k+1} = R\Big(P^k + \beta \mathbf{\Pi}_{P^k}\Big(\widetilde{P}^k - P^k\Big)\Big)$;
**end**

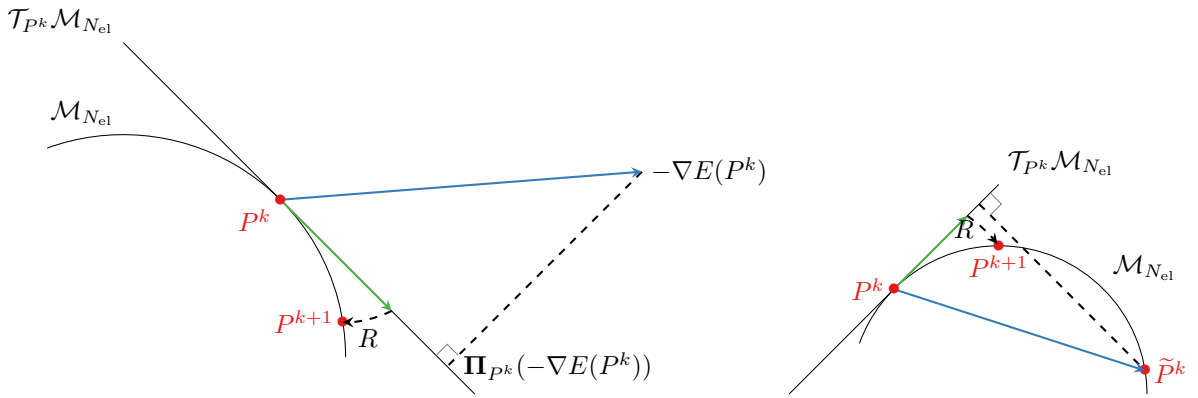



FIGURE 2 – Représentation graphique de la descente de gradient projetée (gauche) et du SCF amorti (droite).

Nous montrons en particulier dans ce chapitre deux théorèmes décrivant la convergence de ces deux méthodes :

**Théorème 1.** *Sous les bonnes hypothèses, si $P^0 \in \mathcal{M}_{N_{\mathrm{el}}}$ est suffisamment proche de $P_*$, la descente de gradient projetée converge linéairement vers $P_*$ pour $\beta > 0$ assez petit, avec comme taux de convergence asymptotique $r(1 - \beta J_{\mathrm{grad}})$ où $J_{\mathrm{grad}} = \mathbf{\Omega}_* + \boldsymbol{K}_*$.*

**Théorème 2.** *Sous les bonnes hypothèses, pour $\beta > 0$ assez petit et $P^0 \in \mathcal{M}_{N_{\mathrm{el}}}$ suffisamment proche de $P_*$, le SCF amorti converge vers $P_*$, avec comme taux de convergence asymptotique $r(1 - \beta J_{\mathrm{SCF}})$ où $J_{\mathrm{SCF}} = 1 + \mathbf{\Omega}_*^{-1} \boldsymbol{K}_*$.*

La convergence de ces algorithmes dépend donc respectivement du rayon spectral $r$ de l'opérateur $1-\beta J$ où $J = J_{\text{grad}} = \boldsymbol{\Omega}_* + \boldsymbol{K}_*$ pour la descente de gradient et $J = J_{\text{SCF}} = 1 + \boldsymbol{\Omega}_*^{-1}\boldsymbol{K}_*$ pour le SCF amorti. On notera en particulier que l'inversibilité de $\boldsymbol{\Omega}_*$ nécessite l'existence d'un gap strictement positif entre la plus haute valeur propre occupée et la plus basse valeur propre non occupée de l'Hamiltonien auto-cohérent $H(P_*)$ (cette hypothèse transparait d'ailleurs dans la description de l'algorithme afin de pouvoir définir de façon unique $\widetilde{P}^k$ étant donnée $P^k$). On remarque alors immédiatement que plus ce gap est petit, plus difficile sera la convergence des algorithmes de type SCF : il s'agit là d'un problème classique rencontré par les chimistes, que nous sommes en mesure de quantifier mathématiquement. Ces résultats permettent alors une meilleure compréhension du problème de minimisation initial et créent des liens entre deux méthodes *a priori* complètement différentes. En guise de conclusion, cette comparaison systématique aide à discuter de la pertinence d'une méthode ou l'autre en fonction de la situation.

## Chapitre 3

Dans ce chapitre, nous nous intéressons à l'estimation de l'erreur de discrétisation pour l'approximation numérique de problèmes de calcul de structures électroniques. Pour cela, nous utilisons une approche basée sur la linéarisation des équations de Kohn–Sham discrètes effectuée dans le Chapitre 2. Pour simplifier, cette approche peut être vue de la façon suivante : supposons que nous cherchions $x \in \mathbb{R}^n$ tel que $f(x) = 0$, pour une fonction non linéaire $f : \mathbb{R}^n \to \mathbb{R}^n$ (le résidu). Au voisinage d'une solution $x_*$, on peut écrire $f(x) \approx f'(x)(x - x_*)$ et alors, si $f'(x)$ est inversible, nous obtenons la relation erreur-résidu suivante :

$$x - x_* \approx f'(x)^{-1} f(x).$$

Lae lecteur·rice attentif·ve notera que c'est cette relation même qui est utilisée dans l'algorithme de Newton. Supposons maintenant que l'on veuille calculer une quantité d'intérêt (QoI) $A(x_*)$, où $A : \mathbb{R}^n \to \mathbb{R}$ est une fonction de classe $C^1$ (par exemple l'énergie, une composante des forces interatomiques, de la densité…), nous obtenons alors l'approximation suivante, où le membre de droite est calculable sans connaissance de la solution exacte $x_*$,

$$A(x) - A(x_*) \approx \nabla A(x) \cdot \big(f'(x)^{-1} f(x)\big).$$

Cela donne alors une première estimation (naïve) de l'erreur :

$$|A(x) - A(x_*)| \leqslant |\nabla A(x)| \big\| f'(x)^{-1} \big\|_{\text{op}} |f(x)|,$$

où $|\cdot|$ est une norme quelconque de $\mathbb{R}^n$, et $\|\cdot\|_{\text{op}}$ la norme d'opérateur associée sur $\mathbb{R}^{n\times n}$ (notons que $\nabla A(x) \in \mathbb{R}^n$ et $f'(x) \in \mathbb{R}^{n\times n}$). En étendant cette approche aux problèmes qui nous intéressent, nous rencontrons plusieurs difficultés qui ont mené à plusieurs résultats, que nous résumons brièvement ici.

Premièrement, dans le cas du calcul de structures électroniques, la présence de contraintes et de dégénérescences donne naissance à des problèmes dont la structure n'est pas facilement transcrite sous la forme présentée ci-dessus. Pour remédier à ce premier obstacle, nous utilisons le cadre géométrique mis en place dans le Chapitre 2 afin d'identifier le bon analogue de la Jacobienne $f'(x)$. En effet, nous avons prouvé que l'opérateur $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ est la Jacobienne du résidu $R : P \mapsto \boldsymbol{\Pi}_P H(P)$, qui s'annule en $P = P_*$. Notre approche est donc basée sur l'approximation

$$P - P_* \approx (\boldsymbol{\Omega}_* + \boldsymbol{K}_*)^{-1} R(P),$$

où $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ joue le rôle de $f'$ dans le cadre général, $P_*$ est une solution de référence (idéalement exacte, mais en pratique obtenue avec une discrétisation très fine) et $P$ est une solution approchée, obtenue avec une discrétisation plus grossière. Il s'avère que cette approximation est très bonne, même pour des discrétisations très grossières, mais elle n'est pas utilisable en pratique à cause du coût prohibitif de l'inversion de $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ dans l'espace de référence.

Ensuite, le choix d'une norme appropriée n'est pas évident dans ce contexte. En général, pour des problèmes impliquant des EDPs, il est naturel de se tourner vers des normes de Sobolev avec les bons exposants afin de faire de la Jacobienne un opérateur borné entre les espaces fonctionnels associés. Dans ce chapitre, nous explorons différents choix de normes et leur influence sur les estimées d'erreur. Cependant, dans notre cas, les bornes naïves

$$\big|\nabla A(x) \cdot \big(f'(x)^{-1} f(x)\big)\big| \leqslant |\nabla A(x)| \big\| f'(x)^{-1} \big\|_{\text{op}} |f(x)|,$$

où $A$ représente les forces interatomiques, sont largement sous-optimales (de plus de cinq ordres de magnitude), même avec des normes de Sobolev appropriées. Nous montrons alors que cela est dû, dans le cas de la DFT en ondes planes, au fait que l'erreur de discrétisation est principalement supportée par les hautes fréquences alors que $\nabla A$ est surtout supporté par les basses fréquences pour les forces interatomiques.

Nous suivons alors naturellement une autre idée, qui consiste à remplacer l'erreur par le résidu, correctement préconditionné. Cela présente l'avantage d'être une quantité facilement accessible en pratique, au prix de remplacer les bornes d'erreurs par des approximations du type de celles décrites plus haut. Cela abouti à des estimations raisonnables de l'erreur sur les QoI qui nous intéressent, mais qui ne sont ni des bornes supérieures systématiques, ni asymptotiquement valides. Ce second point est de nouveau dû au fait que l'erreur et le résidu préconditionné diffèrent essentiellement sur les basses fréquences. Nous proposons alors une approche basée sur un complément de Schur entre les hautes et basses fréquences afin d'approcher l'inverse de $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$. Cela améliore de façon systématique l'estimation de l'erreur de discrétisation sur les basses fréquences, à partir de laquelle nous pouvons estimer l'erreur sur $A$ à un coût raisonnable : la Jacobienne $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ ne doit plus être inversée que sur les basses fréquences (au lieu de l'espace tout entier), les hautes fréquences étant approchées à l'aide d'un préconditionneur cinétique (diagonal en ondes planes). Les deux parties sont enfin couplées à l'aide d'un complément de Schur, ce qui donne une estimation de l'erreur pour un coût limité : l'inversion de la Jacobienne sur les basses fréquences a un coût du même ordre de grandeur que les algorithmes SCF utilisés pour obtenir l'approximation $P$ dans l'espace grossier.

## Chapitre 4

Dans les Chapitres 2 et 3, nous avons étudié des algorithmes de calcul d'états fondamentaux ainsi que les erreurs de discrétisation associées. Nous avons également pu nous intéresser aux forces interatomiques qui, grâce au théorème de Hellmann–Feynman, ne requièrent que la connaissance de l'état fondamental pour être calculées. En revanche, de nombreuses quantités d'intérêt, telles que la polarisabilité, la susceptibilité magnétique, les spectres de phonons…, requièrent le calcul de dérivées de l'état fondamental par rapport à certains paramètres. Plus récemment, l'utilisation du machine-learning en DFT nécessite également des dérivées par rapport aux paramètres des modèles. En pratique, ces dérivées sont calculées grâce à la théorie des perturbations, un cadre aussi connu sous le nom de DFPT ("density functional perturbation theory") en chimie.

Dans ce chapitre, nous introduisons donc dans un premier temps le cadre nécessaire à la DFPT, en particulier l'introduction d'une température numérique lorsque l'on souhaite travailler avec des systèmes métalliques. Nous rappelons ensuite les différents résultats existants, obtenus en général à l'aide de la théorie des perturbations au premier ordre, et nous présentons l'équation de Sternheimer, pierre angulaire du calcul de réponses en DFT. Nous montrons que différents choix de jauges sont possibles, sans altérer le résultat final, mais dont le choix peut être déterminant pour la stabilité et la robustesse des méthodes numériques sous-jacentes. Après une revue des techniques utilisées en pratique, nous proposons un cadre commun dans lequel celles-ci rentrent et nous analysons leur stabilité numérique. Enfin, nous proposons une nouvelle approche dans la résolution de l'équation de Sternheimer qui, à l'aide d'un complément de Schur, tire profit des états d'énergies non occupés, mais calculés au préalable, afin d'améliorer la robustesse des solveurs linéaires utilisés. En particulier, nous montrons comment nous parvenons à gagner plus de 40% de temps de calcul sur des systèmes connus pour être numériquement difficiles.

## Chapitre 5

Le Chapitre 5 est dédié à l'étude de la régularité des solutions d'équations de type Schrödinger, linéaires et non linéaires, avec des potentiels analytiques (*i.e.* dont l'extension analytique est une fonction entière). L'étude menée dans ce chapitre est motivée par l'approximation dites des *pseudopotentiels*, introduite dans le Chapitre 1.

Nous étudions donc dans ce chapitre la régularité de solutions d'équations de la forme $-\Delta u + Vu + g(u) = f$ ou $-\Delta u + Vu + g(u) = \lambda u$ où les données $V$, $g$ et $f$ sont analytiques et les conséquences sur l'analyse *a priori* de l'erreur de discrétisation en ondes planes. Le cas du problème aux valeurs propres

avec des potentiels ayant une régularité de Sobolev donnée a déjà été étudié dans [31]. En particulier, il a été prouvé, dans le cas tridimensionnel et pour une certaine classe de non linéarités, que si le potentiel $V$ appartient à un espace de Sobolev périodique d'exposant $s > 3/2$, alors les solutions $u$ sont dans l'espace de Sobolev périodique d'exposant $s + 2$. Des taux de convergence polynomiaux (et optimaux) sont également dérivés dans tous les espaces de Sobolev de coefficients $-s < r < s + 2$. Dans le cas de potentiels analytiques, on s'attend à une convergence exponentielle de l'erreur de discrétisation et les taux polynomiaux mentionnés précédemment, même s'ils sont valides, ne semblent pas optimaux. Le but de ce chapitre est donc de quantifier cette convergence.

Pour des raisons pédagogiques, nous travaillons avec des équations de Schrödinger unidimensionnelles, linéaires ou non linéaires, car (i) visualiser des extensions analytiques dans le plan complexe de fonctions initialement définies sur $\mathbb{R}^d$ est plus facile quand $d = 1$, et (ii) détecter des taux de convergence exponentiels est plus facile avec des simulations numériques unidimensionnelles. Nous introduisons dans ce chapitre les espaces fonctionnels $(\mathcal{H}_A)_{A>0}$ composés des fonctions $2\pi$-periodiques sur l'axe réel admettant une extension analytique sur la bande $\mathbb{R} + \mathrm{i}(-A, A)$, avec une norme $\|\cdot\|_A$, et nous prouvons, dans le cas linéaire, le théorème suivant.

**Théorème 3.** *Soient $B > 0$ et $V \in \mathcal{H}_B$ à valeurs réelles et telle que $V \geqslant 1$ sur $\mathbb{R}$. Alors, pour tout $0 < A < B$ et $f \in \mathcal{H}_A$, l'unique solution $u$ de $-\Delta u + Vu = f$ est dans $\mathcal{H}_A$. De plus, nous avons l'inégalité suivante*

$$\exists\, C > 0 \text{ indépendante de } f \text{ telle que } \|u\|_A \leqslant C\|f\|_A.$$

*Par conséquent, si $V$ et $f$ sont entières, alors $u$ l'est aussi.*

Un résultat similaire peut être prouvé pour le problème aux valeurs propres

$$\begin{cases} -\Delta u + Vu = \lambda u, \\ \|u\| = 1, \end{cases}$$

dans le sens où, si $V \in \mathcal{H}_B$, alors $u \in \mathcal{H}_A$ pour tout $0 < A < B$. Une conséquence directe de ces résultats est que l'erreur de discrétisation en ondes planes (*i.e.* l'erreur entre la solution exacte et une solution variationnelle ayant des modes de Fourier à support fini) converge plus vite que n'importe quelle exponentielle si les données du problème sont entières.

Cependant, dans le cas non linéaire, de tels résultats ne sont en général plus vrais : nous mettons en avant dans ce chapitre un contre-exemple basé sur une équation de Gross–Pitaevskii unidimensionnelle et pour laquelle nous montrons, en utilisant une combinaison d'outils théoriques et numériques, que les solutions de

$$-\varepsilon\Delta u_\varepsilon + u_\varepsilon + u_\varepsilon^3 = \mu \sin, \quad \varepsilon \geqslant 0,$$

ne sont pas entières, même si le terme source et la non linéarité le sont.

## Chapitre 6

Ce dernier chapitre est un peu différent des autres, car il ne traite pas directement de DFT en ondes planes. Il traite néanmoins toujours de chimie quantique. En effet, nous nous intéressons pour finir à la construction de bases atomiques optimales. Après un court passage en revue des différentes (et nombreuses) bases existantes, nous proposons une première méthode de construction de bases optimales pour des critères d'optimisation généraux. L'objectif sous-jacent à ce chapitre est la construction de bases atomiques, optimales pour un critère, qui peuvent être systématiquement raffinées afin d'améliorer la précision des résultats. Nous traitons et analysons cette approche pour deux critères en particulier, l'un basé sur la matrice densité et l'autre sur l'énergie du fondamental, pour un modèle jouet qui correspond à une version simplifiée et unidimensionnelle de la dissociation d'une molécule diatomique.

# List of contributions

## Accepted or published papers

[GK1] Eric Cancès, Gaspard Kemlin, and Antoine Levitt. Convergence analysis of direct minimization and self-consistent iterations. *SIAM Journal on Matrix Analysis and Applications*, 42(1):243–274, 2021.

[GK2] Eric Cancès, Geneviève Dusson, Gaspard Kemlin, and Antoine Levitt. Practical error bounds for properties in plane-wave electronic structure calculations. *SIAM Journal on Scientific Computing*, 44(5):B1312–B1340, 2022.

[GK3] Eric Cancès, Geneviève Dusson, Gaspard Kemlin, and Laurent Vidal. On basis set optimisation in quantum chemistry. *Accepted in ESAIM Proceedings*, 2022.

⇝ [GK1] coincides with Chapter 2.

⇝ [GK2] coincides with Chapter 3.

⇝ [GK3] coincides with Chapter 6.

## Preprints

[GKp1] Eric Cancès, Michael F. Herbst, Gaspard Kemlin, Antoine Levitt, and Benjamin Stamm. Numerical stability and efficiency of response property calculations in density functional theory. *Submitted*, 2022.

⇝ [GKp1] coincides with Chapter 4.

## In preparation

[GKip1] Eric Cancès, Gaspard Kemlin, and Antoine Levitt. A priori error analysis of linear and nonlinear periodic Schrödinger equations with analytic potentials. *In preparation*, 2022.

⇝ [GKip1] coincides with Chapter 5.

## Software

⇝ During this thesis, I made several contributions to the DFTK software (see Section 1.4.4):

- the Kohn–Sham equations, in their discrete form, are linearized in Chapter 2, from which I could implement a Newton solver;
- error estimators for quantities of interest are developed in Chapter 3 and I implemented them in DFTK[3][4] for interatomic forces;
- a framework to perform response calculations, a cornerstone to the implementation of AD in DFTK, is proposed in Chapter 4 and I implemented it in DFTK.

[3] https://juliamolsim.github.io/DFTK.jl/stable/examples/error_estimates_forces/
[4] https://github.com/gkemlin/paper-forces-estimator

# Contents


**1 Introduction** **1**

1.1 Historical overview . . . 2

1.2 General organization and context . . . 2

1.3 Mathematical framework of electronic structure theory . . . 3

1.3.1 The quantum many-body problem . . . 3

1.3.2 Approximation models . . . 5

1.4 Plane-wave density functional theory . . . 12

1.4.1 Plane-wave discretization . . . 13

1.4.2 Pseudopotential approximation – **Results from Chapter 5** . . . 13

1.4.3 Brillouin zone sampling . . . 15

1.4.4 DFTK: the Density Functional ToolKit . . . 17

1.5 Computing the ground-state . . . 19

1.5.1 General setting . . . 19

1.5.2 Direct minimization algorithms . . . 19

1.5.3 Self-consistent field algorithms . . . 20

1.5.4 Direct minimization or SCF? – **Results from Chapter 2** . . . 21

1.6 Estimating the error . . . 23

1.6.1 Sources of error . . . 24

1.6.2 The linear case . . . 24

1.6.3 The nonlinear case . . . 25

1.6.4 Practical error estimates for plane-wave KS-DFT – **Results from Chapter 3** . . . 25

1.7 Density functional perturbation theory . . . 27

1.7.1 The Kohn–Sham equations at finite temperature . . . 28

1.7.2 Density functional perturbation theory . . . 28

1.7.3 Calculations of response properties for metals – **Results from Chapter 4** . . . 29

**2 Convergence analysis of direct minimization and self-consistent iterations** **33**

2.1 Introduction . . . 34

2.2 Optimization on Grassmann manifolds . . . 36

2.2.1 First-order condition . . . 37

2.2.2 Second-order condition . . . 38
2.2.3 Fixed-point iterations on a manifold . . . 39
2.3 Algorithms and analysis of convergence . . . 39
2.3.1 Direct minimization . . . 39
2.3.2 Damped self-consistent field . . . 42
2.3.3 Comparison . . . 45
2.4 Numerical tests . . . 46
2.4.1 The retraction . . . 47
2.4.2 A toy model with tunable spectral gap . . . 47
2.4.3 Chaos in SCF iterations . . . 49
2.4.4 Local convergence for a 1D nonlinear Schrödinger equation . . . 50
2.4.5 Kohn–Sham density functional theory . . . 55
2.5 Conclusion . . . 57

**3 Practical error bounds for properties in plane-wave electronic structure calculations 61**
3.1 Introduction . . . 62
3.2 Mathematical framework . . . 64
3.2.1 General framework . . . 64
3.2.2 First-order geometry . . . 65
3.2.3 Second-order geometry . . . 65
3.2.4 Density matrix and orbitals . . . 66
3.2.5 Metrics on the tangent space . . . 68
3.2.6 Correspondence rules . . . 69
3.3 The periodic Kohn–Sham problem . . . 69
3.3.1 The continuous problem . . . 69
3.3.2 Discretization . . . 70
3.3.3 Forces . . . 71
3.3.4 Numerical setup . . . 71
3.4 A first error bound using linearization . . . 72
3.4.1 Linearization in the asymptotic regime . . . 72
3.4.2 A simple error bound based on operator norms . . . 73
3.4.3 Error bounds on QoIs and applications to interatomic forces . . . 75
3.5 Improved error bounds based on frequencies splitting . . . 77
3.5.1 Spectral decomposition of the error . . . 77
3.5.2 Improving the error estimation . . . 78
3.6 Numerical examples with more complex systems . . . 80

3.7 Conclusion . . . 82

**4 Numerical stability and efficiency of response property calculations in DFT 87**

4.1 Introduction . . . 88

4.2 Mathematical framework . . . 90

4.2.1 Periodic Kohn–Sham equations . . . 90

4.2.2 Density functional perturbation theory . . . 92

4.2.3 Plane-wave discretization and numerical resolution . . . 93

4.3 Computing the response . . . 93

4.3.1 Practical implementation . . . 93

4.3.2 Occupied-occupied contributions . . . 95

4.3.3 Computation of unoccupied-occupied contributions employing a Schur complement 96

4.4 Numerical tests . . . 98

4.4.1 Insulators and semiconductors . . . 98

4.4.2 Metals . . . 99

4.4.3 Comparison to shifted Sternheimer approaches . . . 101

4.5 Conclusion . . . 102

**5 A priori error analysis of periodic Schrödinger equations with analytic potentials 109**

5.1 Introduction . . . 110

5.2 Spaces of analytic functions . . . 111

5.3 The linear case . . . 112

5.3.1 The linear elliptic problem . . . 112

5.3.2 The linear eigenvalue problem . . . 114

5.3.3 Plane-wave approximation of the linear Schrödinger equation . . . 114

5.4 The nonlinear case: a counter-example . . . 115

5.5 Extension to the multidimensional case with application to Kohn–Sham models. . . . 121

**6 On basis set optimization in quantum chemistry 127**

6.1 Introduction . . . 128

6.2 Optimization criteria . . . 129

6.3 Application to 1D toy model . . . 131

6.3.1 Description of the model . . . 131

6.3.2 Variational approximation in AO basis sets . . . 132

6.3.3 Overcompleteness of Hermite Basis Sets . . . 134

6.3.4 Practical computation of the criterion $J_A$ and $J_E$ . . . 134

6.4 Numerical results . . . 136

6.4.1 Numerical setting and first results . . . . . 136
6.4.2 Influence of numerical parameters . . . . . 142

**Bibliography** **147**

Chapter 1

# Introduction

## Contents


**1.1 Historical overview . . . 2**
**1.2 General organization and context . . . 2**
**1.3 Mathematical framework of electronic structure theory . . . 3**
1.3.1 The quantum many-body problem . . . 3
1.3.2 Approximation models . . . 5
**1.4 Plane-wave density functional theory . . . 12**
1.4.1 Plane-wave discretization . . . 13
1.4.2 Pseudopotential approximation – **Results from Chapter 5** . . . 13
1.4.3 Brillouin zone sampling . . . 15
1.4.4 DFTK: the Density Functional ToolKit . . . 17
**1.5 Computing the ground-state . . . 19**
1.5.1 General setting . . . 19
1.5.2 Direct minimization algorithms . . . 19
1.5.3 Self-consistent field algorithms . . . 20
1.5.4 Direct minimization or SCF? – **Results from Chapter 2** . . . 21
**1.6 Estimating the error . . . 23**
1.6.1 Sources of error . . . 24
1.6.2 The linear case . . . 24
1.6.3 The nonlinear case . . . 25
1.6.4 Practical error estimates for plane-wave KS-DFT – **Results from Chapter 3** 25
**1.7 Density functional perturbation theory . . . 27**
1.7.1 The Kohn–Sham equations at finite temperature . . . 28
1.7.2 Density functional perturbation theory . . . 28
1.7.3 Calculations of response properties for metals – **Results from Chapter 4** . . . 29

## 1.1 Historical overview

Traditionally, applied mathematics has played a fundamental role in computational engineering sciences, such as computational fluid dynamics, mechanics or electromagnetism. The underlying numerical methods have been thoroughly analysed and stand on rigorous theoretical foundations. For instance, the Finite Elements Method (FEM) was invented in the middle of the $20^{\text{th}}$ century by engineers but then substantially improved jointly with applied mathematicians and it would not be at the present stage without the intervention of mathematics. However, because of historical reasons, it has rarely been the case in computational chemistry, solid-state physics or materials science.

This is somewhat surprising since the mathematical problems derived from the Schrödinger equation, which is the cornerstone for many models in these fields, are often eigenvalue problems that need to be discretized for numerical solutions to be computed. Although the discretization of (possibly nonlinear) eigenvalue problems is a well-established topic in applied mathematics, little interaction with computational chemistry, solid-state physics and materials science exists in the literature. The following table shows the number of hits on the databases Google Scholar and MathSciNet, for the keywords "Navier–Stokes" and "density functional theory":

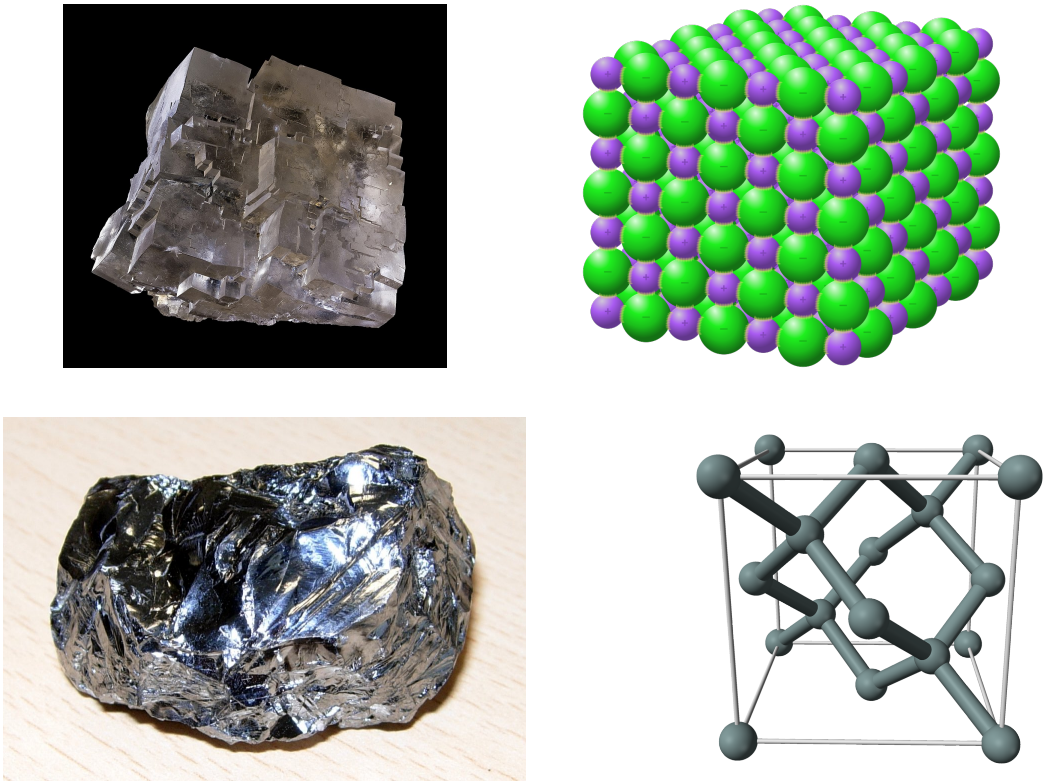
| Keyword: | Google Scholar | MathSciNet |
|---|---|---|
| "Navier–Stokes" | 955,000 | 11,744 |
| "density functional theory" | 1,900,000 | 227 |

TABLE 1.1 – "Navier–Stokes" *vs* "density functional theory", as of January 1st, 2022.

If we assume this data to be a measure of relevance of these two fields of research, it appears that there is a disconnect in trends between the scientific community and its subset of mathematicians. Besides the impressive numbers in Google Scholar, the relevance of computational chemistry and electronic structure is further substantiated by the fact that Kohn and Pople were awarded the Nobel Prize in chemistry in 1998 for their contribution to the density functional theory (DFT) and the seminal work on multiscale methods (QM/MM force field models) of Karplus, Levitt and Warshel has been recognized with the same award in 2013. Nowadays, computational chemistry is fully regarded as a third pillar in chemistry, besides experimental and theoretical chemistry. As an indicator of the importance of DFT in modern science, 12 among the most cited 100 articles relate to it. In particular, 2 of them are in the top 10 and consist in "technical recipes on which the most popular DFT methods and software packages are built" (among the papers in Thomson Reuter's Web of Science, from 1900 to 2014 [195]). As has been the case in the past for the FEM, the field of electronic structure calculation benefited in the last twenty years from the work of mathematicians all over the world, who analysed the existing methods and developed mathematical tools to improve the computational aspects of electronic structure and quantum chemistry. This thesis aims to contribute to these improvements.

## 1.2 General organization and context

This introductory chapter is organized as follows. In Section 1.3, we present the general mathematical framework required to compute the ground-state of general molecular systems, with a particular focus on Kohn–Sham DFT. It is intentionally brief for the sake of clarity and the interested reader is advised to consult the provided references for more details on the different topics, in particular the books [29, 42, 133]. In Section 1.4, we focus on the framework of plane-wave DFT, which uses a Fourier discretization of the objects we study. We also introduce in this section useful approximations for plane-wave DFT: the pseudopotential approximation, which motivated the results from Chapter 5, and the sampling of the Brillouin zone for periodic Schrödinger-like operators. These concepts are useful for understanding the systems we study in this thesis: most of the simulations deal with crystalline systems, that have an intrinsic periodic structure, well suited for plane-wave discretization. These simulations are realized with the DFTK `Julia` package, which we present in Section 1.4.4.



Source: *Wikipedia Commons*

FIGURE 1.1 – Examples of materials that are of interest in this thesis and have a periodic structure: (top) a piece of sodium chloride crystal, commonly known as salt, and its crystalline structure – sodium atoms are in purple and chlorine atoms are in green –, (bottom) a piece of purified silicon crystal, a simple system to test numerical methods, and its diamond cubic crystal structure.

In Section 1.5, we describe the resolution of the equations arising from the previous sections, with a particular focus on two classes of algorithms: direct minimization algorithms and self-consistent field algorithms. We then present the results from Chapter 2, where these two classes are analysed and compared. In Section 1.6, we review the existing literature on error estimates for numerical simulations and electronic structure theory. Then, we discuss the results from Chapter 3, where practical error estimates are derived for plane-wave Kohn–Sham DFT. Finally, we consider in Section 1.7 the framework of density functional perturbation theory, which aims at computing the derivatives of the ground-state density with respect to external perturbations, and present the contributions from Chapter 4.

Finally, we should emphasize that Chapter 6 is the result of a project conducted at the CEMRACS 2021 summer school[1]. While it is related to quantum chemistry, it deals with basis optimization, which is out of the scope of plane-wave DFT. We therefore chose not to mention it in this introduction as it is more or less self-contained.

## 1.3 Mathematical framework of electronic structure theory

### 1.3.1 The quantum many-body problem

All the models studied in this manuscript are based on the solution of a standard problem in quantum chemistry: we seek to determine the ground-state, *i.e.* the one with the lowest energy, of a given molecular system. The models we are going to present are suitable for describing isolated systems, such as molecules. We then present in Section 1.4 how to adapt them to periodic systems, such as crystals.

[1] http://smai.emath.fr/cemracs/cemracs21/

The models that we are going to use are *ab initio*, which means that they are derived directly from the Schrödinger equation and only contain fundamental constants of physics. We work in the scope of the Born–Oppenheimer approximation, of which a precise description is given in [42, Appendix A], so that the nuclei are considered as point-like classical particles. We also ignore for simplicity the spin degrees of freedom. Finally, as it is usual in quantum chemistry, we use atomic units, *i.e.*

$$m_{\mathrm{e}} = 1, \quad e = 1, \quad \hbar = 1, \quad \frac{1}{4\pi\varepsilon_0} = 1, \tag{1.3.1}$$

where $m_{\mathrm{e}}$ is the mass of an electron, $e$ the elementary charge, $\hbar$ the reduced Planck constant and $\varepsilon_0$ the dielectric permittivity of the vacuum. In this framework, a molecular system is composed of:

- $N_{\mathrm{nuc}}$ nuclei, considered as point charges, with position $\overline{\boldsymbol{r}}_k \in \mathbb{R}^3$ and charge $z_k$, for $1 \leqslant k \leqslant N_{\mathrm{nuc}}$. The positions of these particles are considered here as fixed.
- $N_{\mathrm{el}}$ electrons described as quantum particles with wave-function $\Psi$ that belongs to the Hilbert space

$$\bigotimes_{n=1}^{N_{\mathrm{el}}} \mathrm{L}^2(\mathbb{R}^3, \mathbb{C}) \simeq \mathrm{L}^2(\mathbb{R}^{3N_{\mathrm{el}}}, \mathbb{C}). \tag{1.3.2}$$

  Physically, $|\Psi(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{N_{\mathrm{el}}})|^2$ represent the probability density of finding the electrons in a given configuration $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{N_{\mathrm{el}}})$: it therefore integrates to 1 over $\mathbb{R}^{3N_{\mathrm{el}}}$. In addition, due to the fermionic nature of electrons, the Pauli principle states that the electronic wave-function is an antisymmetric function of the positions:

$$\Psi\big(\boldsymbol{r}_{p(1)}, \ldots, \boldsymbol{r}_{p(N_{\mathrm{el}})}\big) = \sigma(p)\Psi(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{N_{\mathrm{el}}}) \tag{1.3.3}$$

  for any permutation $p$ of the indices, $\sigma(p)$ being the parity of $p$. Note that the antisymmetry relation implies in particular that $\Psi(\ldots, \boldsymbol{r}_n, \ldots, \boldsymbol{r}_n, \ldots) = 0$, *i.e.* two electrons cannot occupy the same quantum state[2].

The nuclei positions being fixed, finding the ground-state energy of the system reduces to the resolution of the minimization problem

$$\inf\Big\{\langle \Psi, H_{\mathrm{e}}\Psi\rangle_{\mathrm{L}^2(\mathbb{R}^{3N_{\mathrm{el}}},\mathbb{C})},\ \Psi \in \mathcal{H}_{\mathrm{e}},\ \|\Psi\|_{\mathrm{L}^2} = 1\Big\}, \tag{1.3.4}$$

where

$$H_{\mathrm{e}} = \sum_{n=1}^{N_{\mathrm{el}}} -\frac{1}{2}\Delta_{\boldsymbol{r}_n} - \sum_{n=1}^{N_{\mathrm{el}}}\sum_{k=1}^{N_{\mathrm{nuc}}} \frac{z_k}{|\boldsymbol{r}_n - \overline{\boldsymbol{r}}_k|} + \sum_{1\leqslant n<m\leqslant N_{\mathrm{el}}} \frac{1}{|\boldsymbol{r}_n - \boldsymbol{r}_m|}, \tag{1.3.5}$$

and

$$\mathcal{H}_{\mathrm{e}} = \left\{\Psi \in \bigotimes_{n=1}^{N_{\mathrm{el}}} \mathrm{L}^2(\mathbb{R}^3, \mathbb{C}),\ \int_{\mathbb{R}^{3N_{\mathrm{el}}}} |\nabla\Psi|^2 < +\infty\right\} \tag{1.3.6}$$

is the form domain of the Hamiltonian $H_{\mathrm{e}}$, which is composed of three different terms:

- $-\frac{1}{2}\Delta_{\boldsymbol{r}_n}$ represents the kinetic energy of the electron $n$ and the condition $\displaystyle\int_{\mathbb{R}^{3N_{\mathrm{el}}}} |\nabla\Psi|^2 < +\infty$ ensures finiteness of the kinetic energy;
- the second sum represents the Coulomb interaction between electrons of charge $-1$ and nuclei of charge $z_k$;
- the last term corresponds to the Coulomb interaction between electrons.

The total energy of the system is then recovered by adding the (constant) energy of the interactions between nuclei.

[2] On the contrary, a bosonic quantum state can be occupied by several bosons at the same time.

The minimization problem (1.3.4), together with the time dependent Schrödinger equation

$$\mathrm{i}\frac{\partial \Psi}{\partial t} = H_{\mathrm{e}}\Psi \tag{1.3.7}$$

and its perturbations by an external field can be used to model any molecular system and to derive its subsequent macroscopic properties. However, the wave-function $\Psi$ belongs to the space $\mathrm{L}^2(\mathbb{R}^{3N_{\mathrm{el}}})$: trying to solve directly such equations would yield numerical methods with degrees of freedom that scale exponentially with the number of electrons. This difficulty has a deep physical origin: each electron interacts with every other electrons through the Coulomb interaction and it is not possible to describe the state of one of them without knowing the states of all the others. To give a better idea of this entanglement issue, let us imagine that we want to compute the wave-function of a simple system, *e.g.* the caffeine molecule $\mathrm{C_8H_{10}N_4O_2}$. It is composed of 24 nuclei and 102 electrons, yielding a partial differential equation in dimension $3 \times 102$ for the electrons only. Trying to solve it for instance with the FEM would require at least $2^{306} > 10^{92}$ degrees of freedom, which is not even a conceivable quantity (the number of atoms in the universe is estimated to $10^{80}$). There is thus a real need to develop and study approximations of the "true" Schrödinger equation, as recognized already by Dirac [65], and we now present some of them.

**Dirac (1929)**

*The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation.*

**Remark 1.1** (Spin)**.** In chemistry, it is fundamental to take spin into account, which is an intrinsic component of particles (just as mass or electric charge) and has important theoretical and practical implications. Mathematically, this amounts to consider antisymmetric elements of the Hilbert space

$$\bigotimes_{n=1}^{N_{\mathrm{el}}} \mathrm{L}^2(\mathbb{R}^3 \times \{|\uparrow\rangle, |\downarrow\rangle\}, \mathbb{C}) \tag{1.3.8}$$

where $|\uparrow\rangle$ and $|\downarrow\rangle$ respectively stand for spin up and spin down. Most of the numerical simulations presented in this manuscript are performed with spinless or closed-shell (*i.e.* every quantum state is doubly occupied) molecular systems, so that taking into account spins is not an issue: they can be easily incorporated in practice and remarks like this one will be made when details about the spins are needed.

### 1.3.2 Approximation models

There are three main families of approximation methods in electronic structure calculation:

- *Wave-functions methods* are based on approximating the wave-function $\Psi$ by wave-functions of specific forms, most of the time led by physical intuition from simpler systems. Among these methods, we can cite the Hartree–Fock method, which is a variational approximation that we detail below, or more sophisticated ones like the Configuration Interaction or Coupled Cluster methods (see [95] for more details).
- *Quantum Monte–Carlo methods* use Monte–Carlo probabilistic algorithms to overcome the curse of dimensionality mentioned above by using the links between partial differential equations and stochastic differential equations *via* the Feynman–Kac formula. This type of method will not be further explored and we refer the interested reader to [88].
- *Density functional theory* takes as main object of interest the electronic density

$$\rho(\boldsymbol{r}) = N_{\mathrm{el}} \int_{\mathbb{R}^{3(N_{\mathrm{el}}-1)}} |\Psi(\boldsymbol{r}, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{N_{\mathrm{el}}})|^2 \mathrm{d}\boldsymbol{r}_2 \ldots \mathrm{d}\boldsymbol{r}_{N_{\mathrm{el}}}, \tag{1.3.9}$$

which is a nonnegative function of $\boldsymbol{r} \in \mathbb{R}^3$ only (and not $\mathbb{R}^{3N_{\mathrm{el}}}$), integrating to $N_{\mathrm{el}}$. It then reformulates the minimization problem (1.3.4) into a problem on $\rho$ only, going from $3N_{\mathrm{el}}$ degrees of freedom to 3, independently of the number of electrons. This gain is however compensated by a *nonlinearization* of the problem, with unknown functionals that require approximations. The most famous DFT method is the Kohn–Sham DFT [116], detailed below, as it offers a good compromise between accuracy and computational efficiency.

### Hartree–Fock method

The Hartree–Fock (HF) method consists in a variational approximation of the minimization problem (1.3.4) in which we restrict the minimization space $\{\Psi \in \mathcal{H}_{\mathrm{e}},\ \|\Psi\|_{\mathrm{L}^2} = 1\}$ to the space of *Slater determinants*, *i.e.* functions of the form

$$\Psi(\boldsymbol{r}_1, \dots, \boldsymbol{r}_{N_{\mathrm{el}}}) = \frac{1}{\sqrt{N_{\mathrm{el}}!}} \begin{vmatrix} \phi_1(\boldsymbol{r}_1) & \cdots & \phi_1(\boldsymbol{r}_{N_{\mathrm{el}}}) \\ \vdots & \ddots & \vdots \\ \phi_{N_{\mathrm{el}}}(\boldsymbol{r}_1) & \cdots & \phi_{N_{\mathrm{el}}}(\boldsymbol{r}_{N_{\mathrm{el}}}) \end{vmatrix}, \tag{1.3.10}$$

which we write in short $\Psi = \mathrm{Slater}(\Phi)$ for $\Phi = (\phi_n)_{1 \leqslant n \leqslant N_{\mathrm{el}}}$, where the $(\phi_n)_{1 \leqslant n \leqslant N_{\mathrm{el}}}$ are orthonormal in $\mathrm{L}^2(\mathbb{R}^3, \mathbb{C})$ and are called *molecular orbitals*. Such a function belongs to $\mathcal{H}_{\mathrm{e}}$ if and only if each $\phi_n$ belongs to the Sobolev space $\mathrm{H}^1(\mathbb{R}^3, \mathbb{C})$.

The origin of the Hartree–Fock model comes from the fact that it is exact for noninteracting systems of electrons. Indeed, let us consider the Hamiltonian

$$\widetilde{H}_{\mathrm{e}} = \sum_{n=1}^{N_{\mathrm{el}}} -\frac{1}{2}\Delta_{\boldsymbol{r}_n} + V(\boldsymbol{r}_n) \tag{1.3.11}$$

where $V$ represents the Coulomb interaction with the nuclei and we have neglected the electron-electron interaction. Then the ground-state of $\widetilde{H}_{\mathrm{e}}$ is $\Psi = \mathrm{Slater}(\Phi)$, with energy $\langle \Psi, \widetilde{H}_{\mathrm{e}}\Psi \rangle = \sum_{n=1}^{N_{\mathrm{el}}} \varepsilon_n$, where $\Phi = (\phi_n)_{1 \leqslant n \leqslant N_{\mathrm{el}}}$ and $(\varepsilon_n, \phi_n)_{n \in \mathbb{N}}$ are eigenstates of $\widetilde{H}_{\mathrm{e}}$. This operator has infinitely many eigenstates, with negative nondecreasing eigenvalues $(\varepsilon_1 \leqslant \varepsilon_2 \leqslant \varepsilon_3 \leqslant \cdots)$ accumulating in zero, and positive continuous spectrum. In particular, the $N_{\mathrm{el}}$ lowest energy states are successively occupied and a lower energy cannot be reached because of the fermionic nature of electrons, which prevents them from filling all together the lowest energy level. This is known in chemistry as the *Aufbau* principle, see page 10.

**Remark 1.2** (Spectral theory)**.** The above example shows the importance of studying the spectral properties of Schrödinger operators, in particular the one-body Hamiltonians of the form $-\Delta + V$. Such properties will not be presented here in a systematic manner, but they will be widely used and we refer to [128] for a recent textbook (in French) on the topic. For English readers, we refer to [61].

To derive the Hartree–Fock model, we simply plug the ansatz (1.3.10) in (1.3.4). To this end, we introduce two sets:

- the set of molecular orbitals

$$\mathcal{W}_{N_{\mathrm{el}}} = \left\{ \Phi = (\phi_n)_{1 \leqslant n \leqslant N_{\mathrm{el}}},\ \phi_n \in \mathrm{H}^1(\mathbb{R}^3, \mathbb{C}),\ \int_{\mathbb{R}^3} \phi_n^* \phi_m = \delta_{nm},\ 1 \leqslant n, m \leqslant N_{\mathrm{el}} \right\}; \tag{1.3.12}$$

- the set of Slater determinants

$$\mathcal{S}_{N_{\mathrm{el}}} = \{ \Psi \in \mathcal{H}_{\mathrm{e}},\ \exists\ \Phi = (\phi_n)_{1 \leqslant n \leqslant N_{\mathrm{el}}} \in \mathcal{W}_{N_{\mathrm{el}}},\ \Psi = \mathrm{Slater}(\Phi) \}. \tag{1.3.13}$$

With this formalism, the Hartree–Fock approximation of problem (1.3.4) can be rewritten as

$$\inf \left\{ \langle \Psi, H_{\mathrm{e}}\Psi \rangle_{\mathrm{L}^2},\ \Psi \in \mathcal{S}_{N_{\mathrm{el}}} \right\}. \tag{1.3.14}$$

Let $\Phi = (\phi_n)_{1 \leqslant n \leqslant N_{\mathrm{el}}} \in \mathcal{W}_{N_{\mathrm{el}}}$ and $\Psi \in \mathcal{S}_{N_{\mathrm{el}}}$ be its Slater determinant. After some standard computations [42, Section 1.3], one gets

$$\langle \Psi, H_{\mathrm{e}}\Psi \rangle_{\mathrm{L}^2} = \mathcal{E}^{\mathrm{HF}}(\Phi), \tag{1.3.15}$$

where

$$\mathcal{E}^{\mathrm{HF}}(\Phi) = \sum_{n=1}^{N_{\mathrm{el}}} \int_{\mathbb{R}^3} \frac{1}{2}|\nabla\phi_n|^2 + \int_{\mathbb{R}^3} \rho_\Phi V + \frac{1}{2}\int_{\mathbb{R}^3\times\mathbb{R}^3} \frac{\rho_\Phi(\boldsymbol{r})\rho_\Phi(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{r}' - \frac{1}{2}\int_{\mathbb{R}^3\times\mathbb{R}^3} \frac{|\gamma_\Phi(\boldsymbol{r},\boldsymbol{r}')|^2}{|\boldsymbol{r}-\boldsymbol{r}'|}\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{r}', \tag{1.3.16}$$

with $\rho_\Phi = \sum_{n=1}^{N_{\mathrm{el}}} |\phi_n|^2$ the electronic density and $\gamma_\Phi(\boldsymbol{r},\boldsymbol{r}') = \sum_{n=1}^{N_{\mathrm{el}}} \phi_n(\boldsymbol{r})\phi_n^*(\boldsymbol{r}')$ the density matrix of order 1. We have, in order:

- $\sum_{n=1}^{N_{\mathrm{el}}} \int_{\mathbb{R}^3} \frac{1}{2}|\nabla\phi_n|^2$: the kinetic energy of the orbitals;
- $\int_{\mathbb{R}^3} \rho_\Phi V$: the interaction of the electrons with the nuclei, which generate the Coulomb potential

$$V(\boldsymbol{r}) = -\sum_{k=1}^{N_{\mathrm{nuc}}} \frac{z_k}{|\boldsymbol{r}-\overline{\boldsymbol{r}}_k|}; \tag{1.3.17}$$

- $\frac{1}{2}\int_{\mathbb{R}^3\times\mathbb{R}^3} \frac{\rho_\Phi(\boldsymbol{r})\rho_\Phi(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{r}'$: the Hartree term, that can be seen as the classical Coulomb energy of the density $\rho_\Phi$;
- $\frac{1}{2}\int_{\mathbb{R}^3\times\mathbb{R}^3} \frac{|\gamma_\Phi(\boldsymbol{r},\boldsymbol{r}')|^2}{|\boldsymbol{r}-\boldsymbol{r}'|}\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{r}'$: the exchange term, which can be understood for $N_{\mathrm{el}} = 1$ as it exactly compensates the interaction of an electron with itself in the Hartree term.

Therefore, we can rewrite the Hartree–Fock problem in its simplest form:

$$\mathcal{E}_*^{\mathrm{HF}} = \inf\left\{\mathcal{E}^{\mathrm{HF}}(\Phi),\ \Phi \in \mathcal{W}_{N_{\mathrm{el}}}\right\}. \tag{1.3.18}$$

Note that if $\mathcal{E}_*$ is the solution to problem (1.3.4), then $\mathcal{E}_*^{\mathrm{HF}}$ is an approximation from above of $\mathcal{E}_*$. The difference between these two energies is called the *correlation energy* and more sophisticated *ab initio* models compute an approximation of this energy to improve the result. Such methods are called *post*-Hartree–Fock and we can cite among the most used ones: the Configuration Interaction method, the Coupled Cluster or the multi-configuration methods. The interested reader is referred to [42, Section 6.2.7] or [95, Chapter 5] for more details. Mathematically, a proof of the existence of solutions to the Hartree–Fock minimization problem can be found in [132, 135].

**Remark 1.3** (Spin)**.** As mentioned in Remark 1.1, we omitted here the spin degrees of freedom, describing actually what is known as the *spinless* Hartree–Fock model (there is no spin associated to the orbitals). The *restricted* Hartree–Fock model, in which every orbital is doubly occupied, has a form similar to (1.3.12), (1.3.16), (1.3.18) except that: $N_{\mathrm{el}} = 2N_{\mathrm{p}}$ where $N_{\mathrm{p}}$ represents the number of electron pairs, $\rho_\Phi = 2\sum_{n=1}^{N_{\mathrm{p}}} |\phi_n|^2$ and the prefactor in front of the kinetic term is 1 (instead of 1/2) while the one in front of the exchange term is 1/4 (instead of 1/2). Similar classes of methods exists and allow for more freedom on the spins, such as *unrestricted* Hartree–Fock models, that are better suited to describe open-shell systems.

### Kohn–Sham density functional theory

The first theoretical works on DFT date back to Hohenberg and Kohn [104]. We present here the approach from Levy [126, 127] and Lieb [131], which is based on the following, simple but powerful, statement: the electronic ground-state energy and density of a molecular system with $N_{\mathrm{el}}$ electrons can be obtained by solving the minimization problem

$$\inf\left\{F(\rho) + \int_{\mathbb{R}^3} \rho V,\ \rho \geqslant 0,\ \sqrt{\rho} \in \mathrm{H}^1(\mathbb{R}^3,\mathbb{R}),\ \int_{\mathbb{R}^3} \rho = N_{\mathrm{el}}\right\}, \tag{1.3.19}$$

where $V$ is the potential generated by the nuclei and $F$ is a universal functional of the electronic density $\rho$, in the sense that it does not depend on $V$. For every such $\rho$, $F(\rho)$ can be defined as

$$F(\rho) = \inf \left\{ \left\langle \Psi, H_{\mathrm{e}}^0 \Psi \right\rangle_{\mathrm{L}^2},\ \Psi \in \mathcal{H}_{\mathrm{e}},\ \|\Psi\|_{\mathrm{L}^2} = 1, \right. \\ \left. \rho = N_{\mathrm{el}} \int_{\mathbb{R}^{3(N_{\mathrm{el}}-1)}} |\Psi(\cdot, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{N_{\mathrm{el}}})|^2 \mathrm{d}\boldsymbol{r}_2 \ldots \mathrm{d}\boldsymbol{r}_{N_{\mathrm{el}}} \right\}, \tag{1.3.20}$$

where $H_{\mathrm{e}}^0$ is the Hamiltonian $H_{\mathrm{e}}$ defined in (1.3.6) with the potential $V$ generated by the nuclei set to 0. $F$ is known as the Levy–Lieb functional.

For the moment, the computation of the ground-state energy is still *exact*: no approximations have been made. Nonetheless, there is no known explicit formulation for the density functional $F$, which makes it untractable in practice. The very essence of DFT therefore lies in the approximation of the universal functional $F$. The first approximation was actually proposed in the 30s by Thomas and Fermi, long before the theoretical ground laid by Hohenberg and Kohn. It is inspired from the behaviour of homogeneous electron gases and it suggests for $F$ the following form:

$$F_{\mathrm{TF}}(\rho) = C_{\mathrm{TF}} \underbrace{\int_{\mathbb{R}^3} \rho^{5/3}}_{\text{kinetic energy}} + \frac{1}{2} \underbrace{\mathcal{D}(\rho, \rho)}_{\text{Coulomb interaction}}. \tag{1.3.21}$$

Here, $C_{\mathrm{TF}}$ is a universal nonempirical constant and $\mathcal{D}(\rho, \rho')$ denotes the classical Coulomb interaction energy

$$\mathcal{D}(\rho, \rho') = \int_{\mathbb{R}^3 \times \mathbb{R}^3} \frac{\rho(\boldsymbol{r}) \rho'(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} \mathrm{d}\boldsymbol{r} \mathrm{d}\boldsymbol{r}'. \tag{1.3.22}$$

From this first simple model, more sophisticated models were proposed, among which we can mention:

- the Thomas–Fermi–von Weizsäcker (TFW) model

$$F_{\mathrm{TFW}}(\rho) = C_{\mathrm{W}} \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}|^2 + C_{\mathrm{TF}} \int_{\mathbb{R}^3} \rho^{5/3} + \frac{1}{2} \mathcal{D}(\rho, \rho); \tag{1.3.23}$$

- the Thomas–Fermi–Dirac–von Weizsäcker model

$$F_{\mathrm{TFDW}}(\rho) = C_{\mathrm{W}} \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}|^2 + C_{\mathrm{TF}} \int_{\mathbb{R}^3} \rho^{5/3} - C_{\mathrm{D}} \int_{\mathbb{R}^3} \rho^{4/3} + \frac{1}{2} \mathcal{D}(\rho, \rho). \tag{1.3.24}$$

Thomas–Fermi (TF) type models are not widely used nowadays as more sophisticated models have been introduced since then. However, they are still interesting from a mathematical point of view because they have a simpler structure than Hartree–Fock or Kohn–Sham models and present interesting mathematical properties (*e.g.* they only depend on the density $\rho$, the functional $F$ is a convex function of the density $\rho$ for the TF and TFW models). See for instance [17, 48, 130, 135] for mathematical insights on such models, [31, 210] for their numerical analysis, and references therein.

Kohn–Sham (KS) DFT explicitly includes in the density functional the minimal kinetic energy of a cloud of noninteracting electrons and the Hartree term of Coulomb interaction of a cloud of electrons. What is left is called the *exchange-correlation* energy. Thus, it divides the functional $F$ into three different parts:

- the kinetic energy term $T(\rho) = \inf \left\{ \sum_{n=1}^{N_{\mathrm{el}}} \int_{\mathbb{R}^3} \frac{1}{2} |\nabla \phi_n|^2,\ \Phi = (\phi_n)_{1 \leqslant n \leqslant N_{\mathrm{el}}} \in \mathcal{W}_{N_{\mathrm{el}}},\ \rho_\Phi = \rho \right\}$;
- the Hartree term $\frac{1}{2} D(\rho, \rho)$;
- the exchange-correlation term $E_{\mathrm{xc}}(\rho)$.

The functional $F$ then reads

$$F(\rho) = T(\rho) + \frac{1}{2}D(\rho,\rho) + E_{\mathrm{xc}}(\rho) \tag{1.3.25}$$

where the unknown contributions to the universal functional $F$ are all gathered in the exchange-correlation energy $E_{\mathrm{xc}}$, which usually accounts for 10% of the total energy. In this setting, we can write the KS-DFT model as a minimization problem with the following form:

$$\mathcal{E}_*^{\mathrm{KS}} = \inf\left\{\mathcal{E}^{\mathrm{KS}}(\Phi),\ \Phi \in \mathcal{W}_{N_{\mathrm{el}}}\right\}, \tag{1.3.26}$$

where

$$\mathcal{E}^{\mathrm{KS}}(\Phi) = \sum_{n=1}^{N_{\mathrm{el}}} \int_{\mathbb{R}^3} \frac{1}{2}|\nabla\phi_n|^2 + \int_{\mathbb{R}^3} \rho_\Phi V + \frac{1}{2}\int_{\mathbb{R}^3\times\mathbb{R}^3} \frac{\rho_\Phi(\boldsymbol{r})\rho_\Phi(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{r}' + E_{\mathrm{xc}}(\rho_\Phi), \tag{1.3.27}$$

with $\rho_\Phi(\boldsymbol{r}) = \sum_{n=1}^{N_{\mathrm{el}}} |\phi_n(\boldsymbol{r})|^2$ the density associated to the molecular orbitals $\Phi$. The quality of Kohn–Sham models lies in the approximation of the exchange-correlation energy $E_{\mathrm{xc}}$. In the Hartree–Fock model (1.3.16), the correlation is absent from $E_{\mathrm{xc}}(\rho)$ where only the exchange term appears to balance the interaction of electrons with themselves. In Kohn–Sham DFT, both are mixed to take into account the correlation energy and several ways to approximate this energy exist.

The simplest model is the reduced Hartree–Fock (rHF) model, for which $E_{\mathrm{xc}} = 0$. This model is of a form similar to the Hartree–Fock model, but without the correction to compensate the self-interaction of electrons. This can lead to wrong predictions that do not agree with the experiment, see page 11. A more accurate KS-DFT model is given by the local density approximation (LDA), introduced by Kohn and Sham in 1965 [116] and still commonly used nowadays, in which the exchange-correlation functional is of the form

$$E_{\mathrm{xc}}(\rho) = \int_{\mathbb{R}^3} e_{\mathrm{xc}}(\rho(\boldsymbol{r}))\mathrm{d}\boldsymbol{r}, \tag{1.3.28}$$

where $e_{\mathrm{xc}} : \mathbb{R}^+ \to \mathbb{R}$ is the exchange-correlation energy of a homogeneous electron gas of density $\rho$. The simplest LDA functional, the $X\alpha$ functional [186], is extrapolated from homogeneous electron gases and thus uses the same additional term than the TFDW model:

$$e_{\mathrm{xc}}(\rho(\boldsymbol{r})) = -C_{\mathrm{D}}\rho(\boldsymbol{r})^{4/3}, \quad C_{\mathrm{D}} = \frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}. \tag{1.3.29}$$

There exist other ways to approximate the exchange-correlation functional, for instance the Generalized Gradient Approximation (GGA) in which a correction using the gradient of the electronic density is added:

$$E_{\mathrm{xc}}(\rho) = \int_{\mathbb{R}^3} e_{\mathrm{xc}}(\rho(\boldsymbol{r}), \nabla\rho(\boldsymbol{r}))\mathrm{d}\boldsymbol{r}. \tag{1.3.30}$$

There is a full spectrum of exchange-correlation functionals, going from the Hartree world to the chemical accuracy. This spectrum is sometimes known as "Jacob's ladder", LDA and GGA being the first rungs of the ladder. The interest reader is referred to [192] for a recent review of the topic, accessible to chemists and physicists as well as mathematicians. Let us also mention that the mathematical properties of LDA and GGA KS-DFT, in particular the existence of minimizers, are analysed for instance in [5].

### Kohn–Sham and Hartree–Fock equations

We now have all the tools to introduce the Kohn–Sham equations, a cornerstone of DFT. Writing the Euler–Lagrange equations of problem (1.3.26) yields that

$$\forall\ i = 1,\dots,N_{\mathrm{el}}, \quad H_{\rho_\Phi}\phi_n = \sum_{j=1}^{N_{\mathrm{el}}} \varepsilon_{nm}\phi_m, \tag{1.3.31}$$

where $\varepsilon_{nm}$ is the Lagrange multiplier associated to the constraint $\langle\phi_n,\phi_m\rangle_{\mathrm{L}^2} = \delta_{nm}$ and

$$H_\rho = -\frac{1}{2}\Delta + V + \left(\rho \star \frac{1}{|\cdot|}\right) + V_{\mathrm{xc}}(\rho) \tag{1.3.32}$$

is the self-consistent Hamiltonian generated by the density $\rho$. The term $\rho \star \frac{1}{|\cdot|}$ is also known as the Hartree potential $V_{\mathrm{H}}(\rho)$, which solves the Poisson equation $-\Delta V_{\mathrm{H}}(\rho) = 4\pi\rho$ on $\mathbb{R}^3$. $V_{\mathrm{xc}}(\rho) = \frac{\mathrm{d}E_{\mathrm{xc}}}{\mathrm{d}\rho}(\rho)$ is the exchange-correlation potential, defined as the gradient of the exchange-correlation energy. The Kohn–Sham energy $\mathcal{E}^{\mathrm{KS}}(\Phi)$ is actually invariant by unitary rotation: for any unitary matrix $U$ of order $N_{\mathrm{el}}$, $\rho_{\Phi U} = \rho_\Phi$ and thus $\mathcal{E}^{\mathrm{KS}}(\Phi U) = \mathcal{E}^{\mathrm{KS}}(\Phi)$, where $(\Phi U)_n = \sum_{m=1}^{N_{\mathrm{el}}} \phi_m U_{mn}$. The matrix $(\varepsilon_{nm})_{1\leqslant n,m\leqslant N_{\mathrm{el}}}$ being Hermitian, we can diagonalize it and rotate $\Phi$ in (1.3.31) to obtain an equivalent eigenvalue problem

$$\boxed{\forall\, n,m = 1,\dots,N_{\mathrm{el}}, \quad H_{\rho_\Phi}\phi_n = \varepsilon_n\phi_n, \quad \langle\phi_n,\phi_m\rangle_{\mathrm{L}^2} = \delta_{nm}, \quad \rho_\Phi = \sum_{n=1}^{N_{\mathrm{el}}}|\phi_n|^2,} \tag{1.3.33}$$

known as the *Kohn–Sham equations*. These equations are of uttermost importance: they translate, in a condensed form, the *self-consistent* nature of DFT as the Hamiltonian $H_{\rho_\Phi}$ depends on its eigenvectors $\phi_n$ through the density $\rho_\Phi$. The self-consistency in DFT is for instance well described in [202]. It can also be seen from a mean-field theory point of view: the electrons behave as noninteracting particles in the mean-field potential they create all together.

The Hartree–Fock equations are derived similarly. Writing the Euler–Lagrange equations of problem (1.3.18) yields, after diagonalization of the Lagrange multipliers matrix,

$$\boxed{\forall\, n,m = 1,\dots,N_{\mathrm{el}}, \quad H_\Phi\phi_n = \varepsilon_n\phi_n, \quad \langle\phi_n,\phi_m\rangle_{\mathrm{L}^2} = \delta_{nm},} \tag{1.3.34}$$

where the Fock operator $H_\Phi$ is defined as

$$(H_\Phi)\phi(\boldsymbol{r}) = -\frac{1}{2}\Delta\phi(\boldsymbol{r}) + V(\boldsymbol{r})\phi(\boldsymbol{r}) + \left(\rho_\Phi \star \frac{1}{|\cdot|}\right)(\boldsymbol{r})\phi(\boldsymbol{r}) - \int_{\mathbb{R}^3}\frac{\gamma_\Phi(\boldsymbol{r},\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\phi(\boldsymbol{r}')\mathrm{d}\boldsymbol{r}',$$

with $\rho_\Phi$ and $\gamma_\Phi$ defined in (1.3.16).

### *Aufbau* principle

The *Aufbau* principle (*Aufbauprinzip*, "building-up principle" in German) states that, in the ground-state of a molecular system, electrons fill energy levels in order, from the bottom up. This principle justifies in particular the empirical construction rules of elementary chemistry, such as the Klechkowski or the Madelung rules. The *Aufbau* principle is always satisfied for the Hartree–Fock model (this is due to the variational principle and the specific form of $\mathcal{E}^{\mathrm{HF}}$, see [42, Chapter 5] or [32, Section 22]). As a consequence, the eigenvalues $\varepsilon_n$ in (1.3.34) correspond to the $N_{\mathrm{el}}$ lowest eigenvalues of $H_\Phi$. For the Kohn–Sham model however, it is not known if it holds. It does in practice for most systems and it is always satisfied for the extended Kohn–Sham model where the orbitals are allowed to have fractional occupation numbers, see [32, Section 15] for more details. When, in addition, we have $\varepsilon_{N_{\mathrm{el}}+1} > \varepsilon_{N_{\mathrm{el}}}$, we say that the *strong Aufbau* principle is satisfied. The *Aufbau* principle is of high interest in practice as it gives, for most systems, an accurate heuristic for choosing the occupied eigenvalues along iterations of self-consistent field algorithms, see Section 1.5.

### Density matrices formulations

The formulations we presented above rely on the molecular orbitals $(\phi_n)_{1\leqslant n\leqslant N_{\mathrm{el}}}$ or the electronic density $\rho$, which have a clear physical meaning. However, one will notice that if a set $\Phi$ of orbitals minimizes the Hartree–Fock or Kohn–Sham energy, any unitary rotation of these orbital yields the same energy (the Slater determinant and the density remain unchanged) and there is thus no uniqueness of the solutions. A way to overcome this issue is to use *density matrices* and *density operators* of order 1. Indeed, given a set of orthonormal orbitals $\Phi = (\phi_n)_{1\leqslant n\leqslant N_{\mathrm{el}}}$, let $\gamma_\Phi$ be the orthogonal projector of rank $N_{\mathrm{el}}$ defined by

$$\gamma_\Phi = \sum_{n=1}^{N_{\mathrm{el}}}\phi_n\langle\phi_n,\cdot\rangle_{\mathrm{L}^2}. \tag{1.3.35}$$

Using Dirac bra-ket notation, such a projector can be written as

$$\gamma_\Phi = \sum_{n=1}^{N_{\mathrm{el}}} |\phi_n\rangle\langle\phi_n|. \tag{1.3.36}$$

$\gamma_\Phi$ is the density operator of order 1 associated to $\Phi$, with kernel the density matrix of order 1, still denoted by $\gamma_\Phi$,

$$\gamma_\Phi(\boldsymbol{r},\boldsymbol{r}') = \sum_{n=1}^{N_{\mathrm{el}}} \phi_n(\boldsymbol{r})\phi_n^*(\boldsymbol{r}'). \tag{1.3.37}$$

One can then show that, with $\mathcal{E}$ being either $\mathcal{E}^{\mathrm{HF}}$ or $\mathcal{E}^{\mathrm{KS}}$,

$$\mathcal{E}(\Phi) = \widetilde{\mathcal{E}}(\gamma_\Phi) = \mathrm{Tr}(h\gamma_\Phi) + \mathcal{E}_{\mathrm{nl}}(\gamma_\Phi), \tag{1.3.38}$$

where

- Tr stands for the trace of an operator (a good introduction to trace-class operators can be found in [172, Chapter 6]);
- $h = -\frac{1}{2}\Delta + V$ is the core Hamiltonian of the system;
- $\mathcal{E}_{\mathrm{nl}}$ depends on the chosen model:
  - for the Hartree–Fock model, $\mathcal{E}_{\mathrm{nl}}(\gamma) = \frac{1}{2}\mathrm{Tr}(\mathcal{A}(\gamma)\gamma)$ where for every $\phi \in \mathrm{H}^1(\mathbb{R}^3,\mathbb{C})$ and $\boldsymbol{r} \in \mathbb{R}^3$,
  $$(\mathcal{A}(\gamma)\phi)(\boldsymbol{r}) = \left(\rho_\gamma \star \frac{1}{|\cdot|}\right)(\boldsymbol{r})\phi(\boldsymbol{r}) - \int_{\mathbb{R}^3} \frac{\gamma(\boldsymbol{r},\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\phi(\boldsymbol{r}')\mathrm{d}\boldsymbol{r}', \tag{1.3.39}$$
  $\rho_\gamma(\boldsymbol{r}) = \gamma(\boldsymbol{r},\boldsymbol{r})$ being the electronic density associated to $\gamma$;
  - for the KS-DFT model, $\mathcal{E}_{\mathrm{nl}}(\gamma)$ depends on the choice of the exchange-correlation functional.

Thus, we can look at the Hartree–Fock problem (1.3.18) or the Kohn–Sham problem (1.3.26) in the density matrices formulation by solving the minimization problem

$$\inf\left\{\widetilde{\mathcal{E}}(\gamma),\ \gamma \in \mathcal{P}_{N_{\mathrm{el}}}\right\}, \tag{1.3.40}$$

where

$$\begin{aligned}\mathcal{P}_{N_{\mathrm{el}}} &= \{\gamma_\Phi,\ \Phi \in \mathcal{W}_{N_{\mathrm{el}}}\}\\ &= \left\{\gamma \in \mathfrak{S}_1\big(\mathrm{L}^2(\mathbb{R}^3,\mathbb{C})\big),\ \mathrm{Ran}\,(\gamma) \subset \mathrm{H}^1(\mathbb{R}^3,\mathbb{C}),\ \gamma^2 = \gamma^* = \gamma,\ \mathrm{Tr}(\gamma) = N_{\mathrm{el}}\right\},\end{aligned} \tag{1.3.41}$$

with $\mathfrak{S}_1\big(\mathrm{L}^2(\mathbb{R}^3,\mathbb{C})\big)$ the set of trace-class operators on the space $\mathrm{L}^2(\mathbb{R}^3,\mathbb{C})$.

### Limits of the present models

In this section, we presented some of the most famous models used in electronic structure theory, all of them being accurate in their own range of systems. We try to gather here the main limits of these models. First, the TF-type models are limited to the description of homogeneous electron gases. They are still of interest for their mathematical properties, in particular convexity for TF and TFW, and are nowadays more used as starting points for more sophisticated models. Then comes the rHF model, which is of high interest mathematically as it is a KS-DFT model with a convex functional $F$, which is not the case in general. Contrary to TF-type models, the rHF model reproduces qualitatively some properties such as the shell structure, but also leads to wrong qualitative results: for instance, in the rHF model, the $\mathrm{H}^-$ ion (two electrons, one proton) has a higher energy than that of the atom H and a free electron, leading to the wrong conclusion that the ion is not stable. Third, LDA and Hartree–Fock give satisfying quantitative results for static properties of many materials or molecules. From a mathematical point of view, these methods are a step above the first two models in terms of complexity as we loose convexity of the energy functional.

To describe more complex properties or materials, there are several solutions. For most materials, using advanced DFT functionals (sometimes with empirical data, therefore losing the *ab initio* aspect of the model) is sufficient in general. Dynamical properties, such as excited states, are intrinsically out of the scope of DFT, which focuses on the computation of ground-states. However, various models using ground-state DFT calculations as a starting point exist and allow for the computation of excitation energies and other dynamical properties (see for instance GW methods [7] or time-dependent DFT [27]). Some effects, such as dispersion, can be hard to describe with *ab initio* DFT, but a mixing of DFT and semi-empirical methods can help to obtain good numerical simulations. Let us also mention that another intrinsic limit of Hartree–Fock and KS-DFT with approximate exchange-correlation functionals is reached with *strongly correlated* materials. Such materials often have partially filled $d$ or $f$ orbitals, making it difficult to approximate the interaction of electrons with one-body or local description: each electron has a complex interaction with the others, which cannot be efficiently described with standard mean-field approximations. One significant example is given by the so-called Mott insulators [62, 152], which are expected to be good conductors according to standard DFT, but turn out to be insulators.

## 1.4 Plane-wave density functional theory

Now that the mathematical framework is set up and that we have introduced some of the most representative models, we present in this section some of the numerical tools that will be used in this manuscript to solve these equations on computers. As most of the numerical examples provided in the following chapters are concerned with DFT for crystals, we will focus in this section on the periodic KS-DFT equations, discretized with plane-wave bases, and detail the main approximations we use. Note that this is easily transposable to other models, such as the Hartree–Fock model we presented, as well as molecules, by using a large enough periodic cell to neglect the interaction of the molecule with its periodic neighbours.

A perfect crystal is a physical structure described by a specific disposition of atoms in a unit cell that is repeated periodically. Such a system is described by a Bravais lattice, defined for a (nonnecessarily orthonormal) basis $(\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3)$ of $\mathbb{R}^3$ as $\mathcal{R} = \mathbb{Z}\boldsymbol{a}_1 + \mathbb{Z}\boldsymbol{a}_2 + \mathbb{Z}\boldsymbol{a}_3$, with unit cell $\Omega = [0,1)\boldsymbol{a}_1 + [0,1)\boldsymbol{a}_2 + [0,1)\boldsymbol{a}_3$ and reciprocal lattice $\mathcal{R}^* = \mathbb{Z}\boldsymbol{b}_1 + \mathbb{Z}\boldsymbol{b}_2 + \mathbb{Z}\boldsymbol{b}_3$ with $\boldsymbol{a}_i \cdot \boldsymbol{b}_j = 2\pi\delta_{ij}$. We describe here the formalism of the Kohn–Sham equations with periodic boundary conditions, for a system with $N_{\mathrm{el}}$ electron in the unit cell $\Omega$. This model is somewhat artificial, but simpler than the more physical model of periodic Kohn–Sham equations for an infinite crystal with $N_{\mathrm{el}}$ electrons per unit cell (see Section 1.4.3).

We consider the following periodic functional space, endowed with its usual inner products,

$$\mathrm{L}^2_{\#}(\mathbb{R}^3, \mathbb{C}) = \left\{ f \in \mathrm{L}^2_{\mathrm{loc}}(\mathbb{R}^3, \mathbb{C}),\ f \text{ is } \mathcal{R}\text{-periodic} \right\}, \tag{1.4.1}$$

where a function $f$ is said to be $\mathcal{R}$-periodic if for any $\boldsymbol{x} \in \mathbb{R}^3$ and $\boldsymbol{R} \in \mathcal{R}$, $f(\boldsymbol{x} + \boldsymbol{R}) = f(\boldsymbol{x})$. An orthonormal basis of $\mathrm{L}^2_{\#}(\mathbb{R}^3, \mathbb{C})$ is given by the family $(e_{\boldsymbol{G}})_{\boldsymbol{G} \in \mathcal{R}^*}$ where

$$\forall\ \boldsymbol{r} \in \mathbb{R}^3,\ e_{\boldsymbol{G}}(\boldsymbol{r}) = \frac{1}{\sqrt{|\Omega|}} \mathrm{e}^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{r}}. \tag{1.4.2}$$

Then, we define the periodic Sobolev spaces of order $s \in \mathbb{R}$,

$$\mathrm{H}^s_{\#}(\mathbb{R}^3, \mathbb{C}) = \left\{ f \in \mathrm{L}^2_{\#}(\mathbb{R}^3, \mathbb{C}),\ \sum_{\boldsymbol{G} \in \mathcal{R}^*} \left(1 + |\boldsymbol{G}|^2\right)^s \left|\widehat{f}_{\boldsymbol{G}}\right|^2 < +\infty \right\}, \tag{1.4.3}$$

where $\widehat{f}_{\boldsymbol{G}} = \langle e_{\boldsymbol{G}}, f \rangle_{\mathrm{L}^2_{\#}}$ is the Fourier coefficient of $f$ with wave-vector $\boldsymbol{G} \in \mathcal{R}^*$. In this setting, the Kohn–Sham equations (1.3.33), with periodic boundary conditions and assuming the *Aufbau* principle, read: find $\phi_1, \dots, \phi_{N_{\mathrm{el}}} \in \mathrm{H}^1_{\#}(\mathbb{R}^3, \mathbb{C})$ such that

$$\begin{cases} H_{\rho_\Phi}\phi_n = \varepsilon_n \phi_n, \quad \varepsilon_1 \leqslant \cdots \leqslant \varepsilon_{N_{\mathrm{el}}}, \\ \langle \phi_n, \phi_m \rangle_{\mathrm{L}^2_{\#}} = \delta_{nm}, \quad n, m = 1, \dots, N_{\mathrm{el}}, \\ \rho_\Phi = \displaystyle\sum_{n=1}^{N_{\mathrm{el}}} |\phi_n|^2, \end{cases} \tag{1.4.4}$$

where $H_\rho = -\frac{1}{2}\Delta + V + V_{\mathrm{H}}(\rho) + V_{\mathrm{xc}}(\rho)$. Here, the Hartree potential $V_{\mathrm{H}}(\rho)$ is the unique zero-mean solution to the periodic Poisson equation $-\Delta V_{\mathrm{H}}(\rho) = 4\pi\big(\rho - \fint_\Omega \rho\big)$ and the exchange-correlation potential $V_{\mathrm{xc}}(\rho)$ depends on the chosen approximation of the exchange-correlation energy potential.

### 1.4.1 Plane-wave discretization

The plane-wave discretization method is a specific Galerkin approximation, which takes as variational approximation space

$$X_{N_{\mathrm{b}}} = \mathrm{Span}\left\{e_{\boldsymbol{G}},\ \boldsymbol{G} \in \mathcal{R}^*,\ \frac{1}{2}|\boldsymbol{G}|^2 \leqslant E_{\mathrm{cut}}\right\}, \tag{1.4.5}$$

where $N_{\mathrm{b}}$ is the dimension of the space, linked to the cut-off energy $E_{\mathrm{cut}}$. Denoting by $\Pi_{N_{\mathrm{b}}}$ the $\mathrm{L}^2_\#$-projection operator onto $X_{N_{\mathrm{b}}}$, we then solve the discrete problem: find $\phi_1, \dots, \phi_{N_{\mathrm{el}}} \in X_{N_{\mathrm{b}}}$ such that

$$\begin{cases} \Pi_{N_{\mathrm{b}}} H_{\rho_\Phi} \Pi_{N_{\mathrm{b}}} \phi_n = \varepsilon_n \phi_n, \quad \varepsilon_1 \leqslant \cdots \leqslant \varepsilon_{N_{\mathrm{el}}}, \\ \langle \phi_n, \phi_m \rangle_{\mathrm{L}^2_\#} = \delta_{nm}, \quad n, m = 1, \dots, N_{\mathrm{el}}, \\ \rho_\Phi = \displaystyle\sum_{n=1}^{N_{\mathrm{el}}} |\phi_n|^2. \end{cases} \tag{1.4.6}$$

The components of the core Hamiltonian matrix can be computed explicitly as

$$[H_0]_{\boldsymbol{G}\boldsymbol{G}'} = \langle e_{\boldsymbol{G}}, h e_{\boldsymbol{G}'} \rangle_{\mathrm{L}^2_\#} = \frac{|\boldsymbol{G}|^2}{2} \delta_{\boldsymbol{G}\boldsymbol{G}'} + \langle e_{\boldsymbol{G}}, V e_{\boldsymbol{G}'} \rangle_{\mathrm{L}^2_\#}. \tag{1.4.7}$$

Regarding the other terms of the Hamiltonian, the Hartree potential is obtained by solving the Poisson equation $-\Delta V_{\mathrm{H}}(\rho) = 4\pi\big(\rho - \fint_\Omega \rho\big)$ which, given a density $\rho$, is exactly solved in plane-wave bases and allows for an exact computation of the Hartree energy. The exchange-correlation energy cannot be computed exactly and requires to be approximated with quadrature rules to compute integrals. This discretization method for the Kohn–Sham equations has been analysed for instance in [31].

Using such a discretization method has several consequences, some of which we mention here. In the Fourier space, the Laplace operator is diagonal, which makes its application to a function of $X_{N_{\mathrm{b}}}$, as well as the application of its inverse, immediate. The Laplace operator being the higher-order derivative of the Hamiltonian, using it as a preconditioner is thus a standard practice in plane-wave DFT. This is also used for instance in [38, 69], where the authors propose a post-processing of the Kohn–Sham equations based on a plane-wave discretization method which relies heavily on this property. Note also that computing the density $\rho$ generated by orbitals in $X_{N_{\mathrm{b}}}$ requires real-space products, *i.e.* convolution of their Fourier coefficients. This can be avoided by using bigger Cartesian grids that contains all the $\boldsymbol{G} + \boldsymbol{G}'$ for $|\boldsymbol{G}|$, $|\boldsymbol{G}'| \leqslant \sqrt{2E_{\mathrm{cut}}}$. Finally, the orthogonality constraints imply that the orbitals oscillate rapidly close to the nuclei. This in turn requires a lot of Fourier modes to have accurate approximations. Similarly, exact orbitals usually have cusps (the $1s$ orbital of Hydrogen behaves like $\mathrm{e}^{-|\boldsymbol{r}|}$), which are difficult to approximate using plane-waves because of the link between the regularity of a function and the fast decay of its Fourier coefficients: the less regular the orbital, the slower the convergence of the plane-wave approximation. With methods like FEM, this can be dealt with by placing the atoms on the vertices of the mesh where there is only a continuity constraint. This trick is not possible any more for plane-wave discretizations and another way to overcome this issue is, for instance, to use pseudopotentials.

### 1.4.2 Pseudopotential approximation – Results from Chapter 5

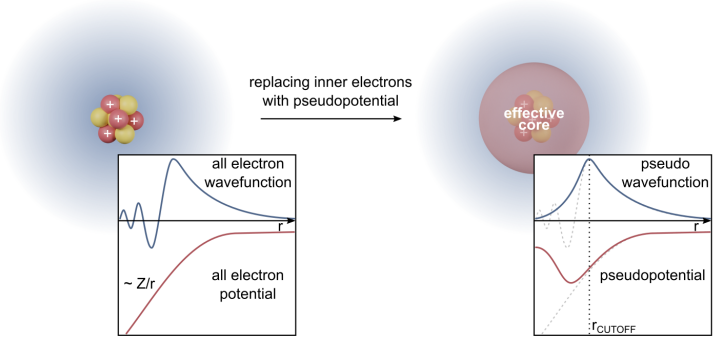
#### General overview

Pseudopotentials have been introduced (i) to remove the cusps that appear at the nuclei positions and (ii) to deal only with the valence electrons by treating the core electrons as frozen. This is justified as only valence electrons are strongly affected by the chemical environment and interact with electrons of other atoms or molecules. This approach allows to reduce the number of electrons explicitly taken into account.

In pseudopotentials methods, the Coulomb and exchange-correlation potentials generated by the core electrons are replaced by fixed, smoother approximations of the original operators. This results into smoother orbitals, that can be well approximated by plane-wave methods, at the price of an additional approximation made by the way the pseudopotentials have been built. A whole zoology of pseudopotentials exists, whose description is out of the scope of this manuscript, and we refer the interested reader to [67] for a detailed overview of pseudopotentials. Let us however mention a specific class of pseudopotentials, known as *norm-conserving* pseudopotentials, first introduced in [191] and then improved in [87]. These pseudopotentials are built such that the actual atomic orbital and the pseudo orbital, seen as radial functions, behave similarly above some radial cut-off . This is enforced by imposing preservation of the norm and some continuity conditions between the core and valence regions, see Figure 1.2. In [114], the authors introduced a specific form for these pseudopotentials, which replace the (local) Coulomb potential $V$ by two contributions, one local term $V_{\mathrm{loc}}$ (a multiplicative operator) and one nonlocal term $V_{\mathrm{nloc}}$ (a finite rank operator), gathering together the contributions from the nuclei and the core electrons. This is known as the Kleinmann–Bylander form of pseudopotentials, where the core Hamiltonian $h$ reads

$$h = -\frac{1}{2}\Delta + V_{\mathrm{loc}} + V_{\mathrm{nloc}}. \tag{1.4.8}$$

The interest of such a form is that the nonlocal contribution being of finite rank, it can be easily applied to (discrete) orbitals. Let us finally mention two specific pseudopotentials that are of interested in this manuscript: Troullier–Martins (TM) pseudopotentials [193] and Goedecker–Teter–Hutter (GTH) pseudopotentials [79, 91].




Source: *Wikipedia Commons*

FIGURE 1.2 – Representation of the pseudopotential approximation: the core electrons are considered as frozen. Note that the pseudo wave-function (or orbital) matches with the actual, all electrons, wave-function outside of some cut-off radius, and that continuity conditions are imposed between the two regions.

### Results from Chapter 5

The choice of pseudopotentials results into local and nonlocal functions of different regularities. As expected, the rate of convergence of the plane-wave discretization method is directly linked to the regularities of these functions. Indeed, this impacts the regularity of the solutions, which in turn impacts the rate of convergence of their plane-wave approximations: if $u \in \mathrm{H}^s_{\#}(\mathbb{R}, \mathbb{C})$, then $\|u - \Pi_N u\|_{\mathrm{H}^r_{\#}}$ goes to 0 as $N^{-(s-r)}$ for any $0 \leqslant r < s$. As a consequence, we expect that the more regular the pseudopotentials, the faster the convergence of the plane-wave discretization method.

The *a priori* error analysis of the plane-wave discretization of the periodic Kohn–Sham equations was performed in [31] for pseudopotentials with Sobolev regularity. It was proved in particular, for the LDA exchange-correlation functional (1.3.28), that if the local part of the pseudopotential and the range of the nonlocal part are in the periodic Sobolev space of order $s > 3/2$, then the Kohn–Sham orbitals $\phi_n$ and the density $\rho_\Phi$ are in the periodic Sobolev space of order $s + 2$, and (optimal) polynomial convergence rates were obtained in any Sobolev spaces of order $r$ with $-s < r < s + 2$. In addition, as for linear second-order elliptic eigenproblems, the error on the eigenvalues converges to zero as the square of the error on the eigenfunctions evaluated in $\mathrm{H}^1$-norm. The analysis in [31] covers for example the case of TM pseudopotentials, for which $s = \frac{7}{2} - \varepsilon$. On the other hand, these estimates are not sharp in the case of GTH pseudopotentials, for which the local and nonlocal contributions are periodic sums of Gaussian-polynomial

functions, and therefore have entire continuations to the whole complex plane. Such pseudopotentials are implemented in different DFT software, such as BigDFT [170], Quantum Espresso [78] or Abinit [84, 175], as well as DFTK, a recent electronic structure package in the `Julia` language [19, 101] (see Section 1.4.4).

We investigate this case in Chapter 5. While it has been known for a long time (see *e.g.* [18, 76, 163] and references therein for a historical overview or [21, 92] for more recent developments) that the solutions to elliptic equations on $\mathbb{R}^d$ with real-analytic data have an analytic continuation in a complex neighbourhood of $\mathbb{R}^d$, the size of this neighbourhood is *a priori* unknown. As already mentioned in the periodic case we are considering, the latter directly impacts the decay rate of the Fourier coefficients of the solution, hence the convergence rate of the plane-wave discretization method. For pedagogical reasons, we work in this chapter with one dimensional linear or nonlinear Schrödinger equations, because (i) it is easier to visualize analytic or entire continuations of functions originally defined on the real space $\mathbb{R}^d$ when $d = 1$, and (ii) exponential convergence rates of plane-wave discretization methods are easier to spot in 1D numerical simulations. We introduce a hierarchy of spaces $(\mathcal{H}_A)_{A>0}$ of complex-valued $2\pi$-periodic functions on the real line having analytic continuations to the strip $\mathbb{R} + \mathrm{i}(-A, A)$, with norm $\|\cdot\|_A$, and we prove the following theorem.

**Theorem 1.1.** *Let $B > 0$ and $V \in \mathcal{H}_B$ be real-valued and such that $V \geqslant 1$ on $\mathbb{R}$. Then, for all $0 < A < B$ and $f \in \mathcal{H}_A$, the unique solution $u$ of $-\Delta u + Vu = f$ is in $\mathcal{H}_A$. Moreover, we have the following estimate*

$$\exists\, C > 0 \text{ independent of } f \text{ such that } \|u\|_A \leqslant C\|f\|_A.$$

*As a consequence, if $V$ and $f$ are entire, then so is $u$.*

A similar result holds for the eigenvalue problem

$$\begin{cases} -\Delta u + Vu = \lambda u, \\ \|u\| = 1, \end{cases} \tag{1.4.9}$$

in the sense that if $V \in \mathcal{H}_B$, then $u \in \mathcal{H}_A$ for any $0 < A < B$. A direct consequence of these results is that the error between the variational solution in $X_{N_\mathrm{b}}$ and the exact solution converges faster than any exponential if the inputs of the problem are entire, which plays in favour of GTH pseudopotentials with plane-wave discretization for linear problems.

However, such results are not true in general for nonlinear models, and we exhibit in Chapter 5 a counter-example based on a 1D Gross–Pitaevskii equation, where we show, using a combination of analytical and numerical tools, that the solution to

$$-\varepsilon\Delta u_\varepsilon + u_\varepsilon + u_\varepsilon^3 = \mu \sin, \quad \varepsilon \geqslant 0, \tag{1.4.10}$$

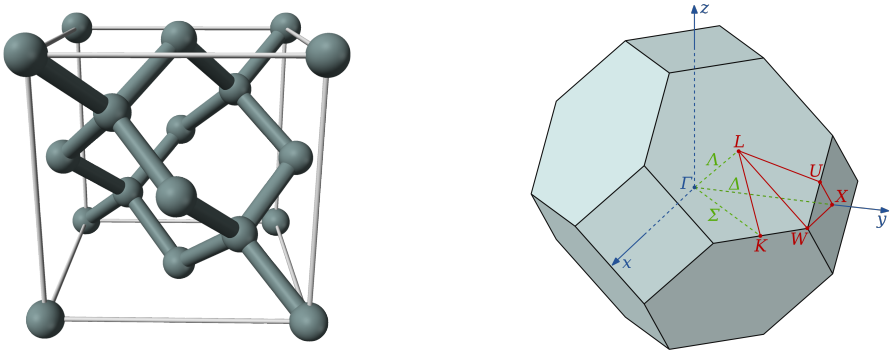
is not entire, even though the source term is.

### 1.4.3 Brillouin zone sampling

The periodic setting we described in this section is somewhat artificial. In practice, it is more relevant to consider the (more realistic) case of the periodic Kohn–Sham equations for an infinite crystal with $N_\mathrm{el}$ electrons per unit cell, which we briefly introduce here. This formalism leads to the study of $\mathcal{R}$-periodic Schrödinger-like operators, whose spectral properties can be deduced from Bloch theory. We introduce to this end the first Brillouin zone $\mathcal{B}$ which is the Voronoï cell of the reciprocal lattice $\mathcal{R}^*$ containing 0 (called the $\Gamma$-point in solid-state physics).

Bloch theory decomposes the Kohn–Sham Hamiltonian $H_\rho$, seen as an $\mathcal{R}$-periodic Schrödinger-like operator, into its Bloch fibers $H_{\rho,\boldsymbol{k}}$. The framework is then similar to what we described before, except that each Hamiltonian $H_{\rho,\boldsymbol{k}}$ needs to be treated separately. The Hamiltonians $H_{\rho,\boldsymbol{k}}$ are operators on $\mathrm{L}^2_\#(\mathbb{R}^3, \mathbb{C})$ with domain $\mathrm{H}^2_\#(\mathbb{R}^3, \mathbb{C})$ defined for any wave-vector $\boldsymbol{k}$ in $\mathbb{R}^3$ by

$$H_{\rho,\boldsymbol{k}} = \frac{1}{2}(-\mathrm{i}\nabla + \boldsymbol{k})^2 + V + V_\mathrm{H}(\rho) + V_\mathrm{xc}(\rho). \tag{1.4.11}$$




Source: *Wikipedia Commons*

FIGURE 1.3 – Unit cell of the FCC silicon crystal (left) and its first Brillouin zone in moments space (right).

$H_{\rho,\boldsymbol{k}}$ is bounded below with compact resolvent for every $\boldsymbol{k} \in \mathbb{R}^3$. Its spectrum is therefore composed of a nondecreasing sequence of eigenvalues $(\varepsilon_{n\boldsymbol{k}})_{n\in\mathbb{N}}$ diverging to $+\infty$: there exists some $\varepsilon_{\boldsymbol{k}} \in \mathbb{R}$ such that

$$\sigma(H_{\rho,\boldsymbol{k}}) = \{\varepsilon_{n\boldsymbol{k}},\ n \in \mathbb{N}\} \subset [\varepsilon_{\boldsymbol{k}}, +\infty). \tag{1.4.12}$$

Note that for any $n \in \mathbb{N}$, $\boldsymbol{k} \mapsto \varepsilon_{n\boldsymbol{k}}$ is Lipschitz continuous and $\mathcal{R}^*$-periodic. The (purely continuous) spectrum of $H_\rho$ can then be computed as the union of the (purely discrete) spectra of $H_{\rho,\boldsymbol{k}}$ for every $\boldsymbol{k} \in \mathcal{B}$

$$\sigma(H_\rho) = \bigcup_{\boldsymbol{k}\in\mathcal{B}} \sigma(H_{\rho,\boldsymbol{k}}). \tag{1.4.13}$$

If we denote by $u_{n\boldsymbol{k}}$ the associated eigenfunctions, then the ground-state density $\rho$ can then be computed as

$$\rho(\boldsymbol{r}) = \fint_{\mathcal{B}} \sum_{n=1}^{N_{\mathrm{el}}} |u_{n\boldsymbol{k}}(\boldsymbol{r})|^2 \mathrm{d}\boldsymbol{k}. \tag{1.4.14}$$

For more details on the Bloch theory, we refer for instance to [171, Section XIII.16] for general proofs of the statements we made.

We now introduce an important quantity in solid-state physics, known as the *Fermi level* $\varepsilon_{\mathrm{F}}$. It represents the highest energy attainable for electrons without violating the charge neutrality of the unit cell. Mathematically, it can be obtained implicitly through

$$\fint_{\mathcal{B}} \sum_{n=1}^{+\infty} \mathbf{1}_{\{\varepsilon_{n\boldsymbol{k}} \leqslant \varepsilon_{\mathrm{F}}\}} \mathrm{d}\boldsymbol{k} = N_{\mathrm{el}}. \tag{1.4.15}$$

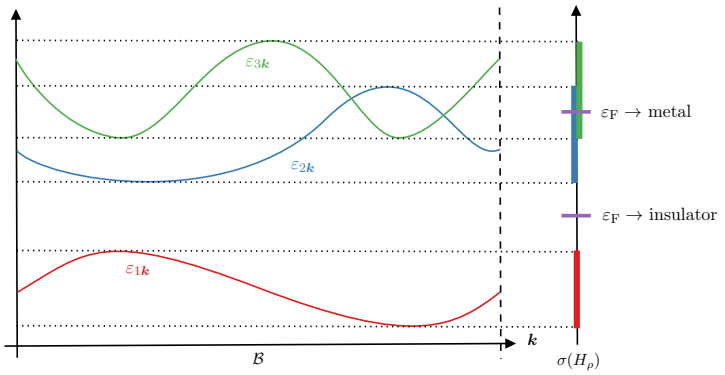
The Fermi level describes the electronic properties of materials: if it lies within a gap between two bands of the spectrum of $H_\rho$, the material is an insulator (or a semiconductor) whereas if it lies in a band of the spectrum of $H_\rho$, the material is a metal (Figure 1.4). For insulators, the Fermi level separates between occupied and virtual states, see Figure 1.5 (left).

Of course, it is not possible in practice to span the full Brillouin zone $\mathcal{B}$ or any continuous path in it. Instead, one usually considers a supercell $\Omega_{N_{\mathrm{cell}}}$ composed of $N_{\mathrm{cell}} = L_1 \times L_2 \times L_3$ copies of the unit cell $\Omega$ (that is $L_i$ copies in direction $i$) and containing $N_{\mathrm{cell}} N_{\mathrm{el}}$ electrons, as well as a finite subset $\mathcal{B}_{N_{\mathrm{cell}}} \subset \mathcal{B}$ made of the $N_{\mathrm{cell}}$ wave-vectors of $\mathcal{B}$ that are compatible with the periodic boundary conditions on $\Omega_{N_{\mathrm{cell}}}$. The generalized eigenfunctions of the periodic operator $H_\rho$ can then be taken as the Bloch waves

$$\phi_{n\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{N_{\mathrm{cell}}}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} u_{n\boldsymbol{k}}(\boldsymbol{r}),$$

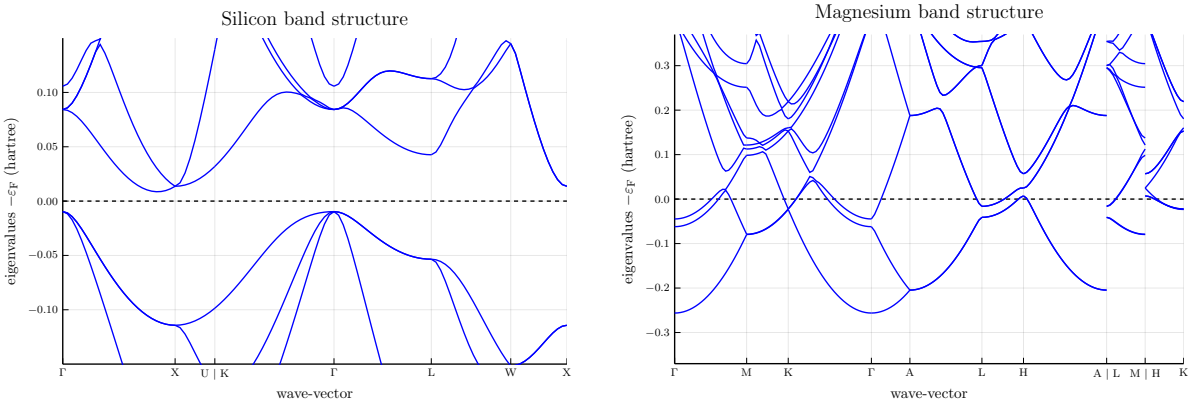
that have $\Omega_{N_{\mathrm{cell}}}$ periodicity for $\boldsymbol{k} \in \mathcal{B}_{N_{\mathrm{cell}}}$, and satisfy $\int_{\Omega_{N_{\mathrm{cell}}}} |\phi_{n\boldsymbol{k}}|^2 = 1$. An approximation of the ground-state density can then be recovered as

$$\rho(\boldsymbol{r}) \approx \sum_{\boldsymbol{k}\in\mathcal{B}_{N_{\mathrm{cell}}}} \sum_{n=1}^{N_{\mathrm{el}}} |\phi_{n\boldsymbol{k}}(\boldsymbol{r})|^2.$$




Source: Laurent Vidal

FIGURE 1.4 – Band diagram and characterization of insulators and metals depending on the location of the Fermi level $\varepsilon_F$ in the spectrum of $H_\rho$. Note that the ordering has not been conserved ($\varepsilon_{2\boldsymbol{k}}$ goes over $\varepsilon_{3\boldsymbol{k}}$) so that the eigenvalues are actually analytic. When the order is conserved, $\boldsymbol{k} \mapsto \varepsilon_{n\boldsymbol{k}}$ is analytic outside of crossings but only Lipschitz continuous at crossings.




Source: *plots generated with DFTK*

FIGURE 1.5 – Band structures of silicon (left) and magnesium (right). Silicon is a semiconductor and there is a clear gap between the first four bands (which are thus occupied) and the others. This is not the case for magnesium, which is a metal. The discontinuity in the band diagram of magnesium is due to a discontinuity in the Brillouin zone path.

In practice, not all $k$-points need to be used. Physical symmetries can be exploited to reduce the total number of $k$-points to a smaller amount of weighted *irreducible* $k$-points to make computations faster. DFT calculations with $k$-points discretizations can also very easily be parallelized: computations can be done for separate $k$-points on different processes, which only need to communicate to compute quantities such as the density $\rho$ or the Fermi level.

Using a discretization of the Brillouin zone introduces an error when it comes to recover quantities of interest from the Bloch fibers of the Hamiltonian. Numerically, Monkhorst and Pack observed in [150] that the uniform discretization of the Brillouin zone $\mathcal{B}$ into $\mathcal{B}_{N_{\text{cell}}}$ led to small errors for insulators and semiconductors, which is why this is a widely used discretization of the Brillouin zone, but it is not an easy task to quantify this error rigorously: for more insight on Brillouin zone, we refer to [80] for a mathematical study in the framework of the reduced Hartree–Fock model for insulators, or [39] for an extension to metallic systems.

### 1.4.4 DFTK: the Density Functional ToolKit

Most of the numerical results presented in this manuscript have been obtained with DFTK [101], a `Julia` package for plane-wave DFT available at https://dftk.org/ and which has been actively developed since 2019, mainly by Antoine Levitt and Michael F. Herbst. While many efficient DFT

codes have already been developed for decades, such as BigDFT [170], Quantum Espresso [78], Abinit [84, 175] or VASP [117], they do not allow for enough flexibility when it comes to the implementation of novel numerical methods. DFTK has been designed to overcome this issue, by allowing simulations from simple 1D toy models well suited for rigorous mathematical analysis to relevant physical systems up to 1,000 electrons. With such a flexibility in the range of systems it supports, DFTK is at the intersection between numerical analysis, high-performance computing and materials simulations, allowing for efficient collaborations between researchers from these fields. DFTK has been presented at JuliaCon 2021 [101] and some of the algorithms implemented were recently published [98, 99].

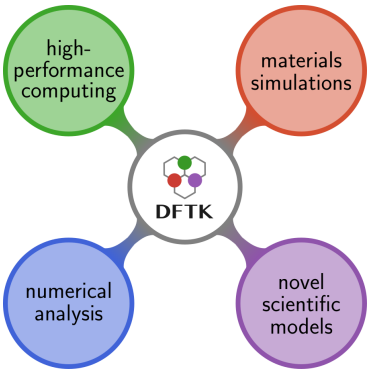
DFTK currently performs plane-wave simulations of periodic systems with GTH [79, 91] pseudopotentials and already supports a sizeable feature sets: 1D, 2D and 3D problems, different DFT models (standard LDA but also PBE exchange-correlation functionals [160]) and algorithms, Monkhorst–Pack uniform discretization of the Brillouin zone, MPI parallelism, custom analytic potentials, interface with established packages such as ASE [103] or pymatgen [157], numerical error control [GK2, 100], ... Coding with DFTK also follows the physical description of the systems we aim at simulating, see Figure 1.6 for an example of a code which computes the ground-state of the FCC silicon crystal.

```
using DFTK

# define the periodic lattice
a = 10.26  # silicon lattice constant in Bohr
lattice = a / 2 * [[0 1 1.];
                   [1 0 1.];
                   [1 1 0.]]
# load silicon information
Si = ElementPsp(:Si, psp=load_psp("hgh/lda/Si-q4"))
# define the atoms that make up the crystal and their positions
atoms     = [Si, Si]
positions = [ones(3)/8, -ones(3)/8]

# use a DFT model with LDA exchange-correlation functional
model = model_LDA(lattice, atoms, positions)
# build the plane-wave basis
basis = PlaneWaveBasis(model; Ecut=15, kgrid=[4, 4, 4])
# solve the KS-DFT equations
scfres = self_consistent_field(basis, tol=1e-8)
# post-process to plot the band diagram
plot_bandstructure(scfres)
```

FIGURE 1.6 – `Julia` code to compute, with DFTK, the ground-state of the FCC silicon crystal and generate Figure 1.5 (left). Notice how the different parameters we introduced up to now are set up.




Source: Michael F. Herbst https://dftk.org

Being written in the `Julia` language, DFTK is also fully composable with the underlying ecosystem. For instance, it natively supports arbitrary floating point precision. On more advanced topics, Automatic Differentiation (AD) [86] is currently being implemented, for both forward and backward modes, using the `Julia` packages that deal with AD. Specific differentiation rules were implemented to make the computations faster: for instance, results from Chapter 2 and Chapter 4 were used to implemented directly the differentiation of the solution of the Kohn–Sham equations without requiring to differentiate through the full numerical solver of these equations. This work highly benefited of the interdisciplinary nature of DFTK as it results from collaboration between chemists, mathematicians and computer scientists. To the

best of our knowledge, such an implementation of AD for DFT is only present in DFTK at the moment and it has been presented at JuliaCon 2022. Let us also mention that, recently, GPU calculations with DFTK have started to be investigated.

## 1.5 Computing the ground-state

### 1.5.1 General setting

In the previous sections, we presented the mathematical background of electronic structure calculation along with some models that were introduced, in their continuous and discrete forms, to approximate the ground-state of the electronic Schrödinger equation. This section is dedicated to the description of some algorithms that solve these models by converging to a fixed-point of (1.3.33). After choosing a discretization method (*e.g.* plane-wave or FEM) with an orthonormal basis, we discretize (1.3.38) into the following constrained minimization problem:

$$\boxed{\inf\{E(P),\ P \in \mathcal{M}_{N_{\mathrm{el}}}\}, \quad \text{with } E(P) = \mathrm{Tr}(H_0 P) + E_{\mathrm{nl}}(P),} \tag{1.5.1}$$

where $H_0$ is the discrete core Hamiltonian matrix and $E_{\mathrm{nl}} : \mathcal{M}_{N_{\mathrm{el}}} \to \mathbb{R}$ describes the interaction of the electrons with themselves, accordingly to the chosen model (for instance Hartree–Fock or KS-DFT). The discrete equivalent $\mathcal{M}_{N_{\mathrm{el}}}$ of $\mathcal{P}_{N_{\mathrm{el}}}$ in (1.3.40) is

$$\mathcal{M}_{N_{\mathrm{el}}} = \left\{P \in \mathbb{C}^{N_{\mathrm{b}} \times N_{\mathrm{b}}},\ P = P^*,\ \mathrm{Tr}(P) = N_{\mathrm{el}},\ P^2 = P\right\}. \tag{1.5.2}$$

It is diffeomorphic to the Grassmann manifold $\mathrm{Grass}(N_{\mathrm{el}}, N_{\mathrm{b}})$ [1] and, in particular, matrices in $\mathcal{M}_{N_{\mathrm{el}}}$ have eigenvalues 0 or 1. $\mathcal{M}_{N_{\mathrm{el}}}$ is a smooth manifold, its tangent space is defined for $P \in \mathcal{M}_{N_{\mathrm{el}}}$ by

$$\mathcal{T}_P\mathcal{M}_{N_{\mathrm{el}}} = \left\{X \in \mathbb{C}^{N_{\mathrm{b}} \times N_{\mathrm{b}}}_{\mathrm{herm}},\ PX + XP = X,\ \mathrm{Tr}(X) = 0\right\}, \tag{1.5.3}$$

and we call $\mathbf{\Pi}_P$ the orthogonal projection operator onto $\mathcal{T}_P\mathcal{M}_{N_{\mathrm{el}}}$ for the Frobenius inner product. Denoting by $H(P) = \nabla E(P) = H_0 + \nabla E_{\mathrm{nl}}(P)$ the Hamiltonian matrix, one immediately gets that the first-order condition satisfied by a solution $P_*$ to (1.5.1) is $\mathbf{\Pi}_{P_*} H(P_*) = 0$, which is equivalent to $[H(P_*), P_*] = H(P_*)P_* - P_* H(P_*) = 0$. The minimization problem (1.5.1) is compact but nonconvex: there exists at least one minimizer, but the minimizer might not be unique, and local minima might not be global ones.

Finally, we recall that, in plane-wave, finite differences or finite elements electronic structure calculation codes, the size $N_{\mathrm{b}}$ of the discretized space is in practice much larger than the number $N_{\mathrm{el}}$ of electrons. Therefore, it is not practical to store and manipulate the (dense) matrix $P$. Instead, these algorithms work on the discrete version of the orbitals $(\phi_n)_{1 \leqslant n \leqslant N_{\mathrm{el}}}$ introduced in (1.3.33). The density matrix $P$ is then recovered as

$$P = \sum_{n=1}^{N_{\mathrm{el}}} \phi_n \phi_n^*. \tag{1.5.4}$$

All the algorithms below are presented for the sake of clarity in the density matrix framework but, in practice, they are expressed in a way that avoids ever forming the density matrix [203]. Moreover, all the algorithms we mention are iterative solvers so that they are almost systematically improved by using *preconditioning*. Finding the best preconditioner, *i.e.* a good compromise between cost and efficiency, is an entire research field by itself but, in the particular case of plane-wave DFT, using a *kinetic* preconditioner is usually the default solution: the kinetic operator $-\frac{1}{2}\Delta$ is diagonal in Fourier modes, which makes its inverse almost free in comparison to other operations such as matrix-vector products, and often yields satisfying results.

### 1.5.2 Direct minimization algorithms

A first class of algorithms solves the minimization problem (1.5.1) by a direct minimization of the energy on the constraint manifold $\mathcal{M}_{N_{\mathrm{el}}}$. They read, in their simplest form, as Algorithm 1.1 (see Figure 1.7 for a schematic view).

**Algorithm 1.1** – Projected gradient descent

**Data:** $P^0 \in \mathcal{M}_{N_{\mathrm{el}}}$
**while** *convergence not reached* **do**
  $P^{k+1} = R\big(P^k - \beta \mathbf{\Pi}_{P^k}\big(\nabla E(P^k)\big)\big)$ ;
**end**

Here, $R$ is called the *retraction* and is used to ensure that the density matrix stays on the manifold $\mathcal{M}_{N_{\mathrm{el}}}$. $\beta$ is a free parameter that represents the step size in the opposite direction of the gradient at each iteration and can be chosen fixed or adapted to each iteration by performing efficient linesearches to minimize the energy in the gradient direction.

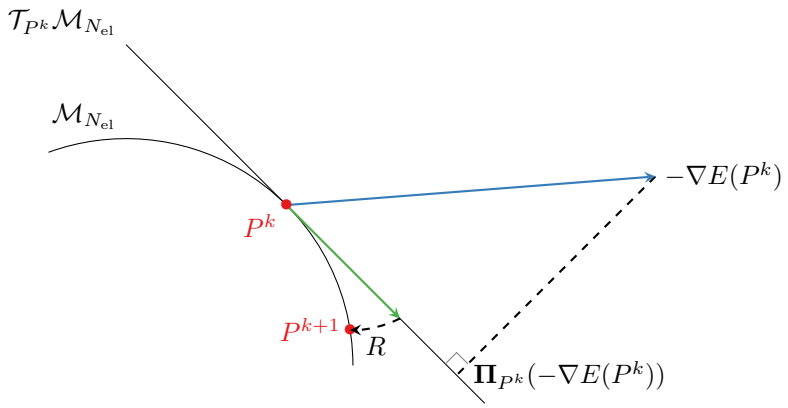



**Figure 1.7** – Schematic view of the projected gradient descent algorithm.

This method is barely used in practice but its simplicity makes it of mathematical interest. The minimization set $\mathcal{M}_{N_{\mathrm{el}}}$ being diffeomorphic to the Grassmann manifold $\mathrm{Grass}(N_{\mathrm{el}}, N_{\mathrm{b}})$, it is naturally equipped with the structure of a Riemannian manifold, allowing for the use of Riemann optimization algorithms [1, 4, 72]. More sophisticated methods have also been developed, among which we can mention Gradient-type [3, 57, 151, 184, 196, 207], and Newton-type [11, 49, 209] methods.

### 1.5.3 Self-consistent field algorithms

A second class of algorithms, known as *self-consistent field* (SCF) methods, is based on the interpretation of problem (1.5.1) as a nonlinear eigenvalue problem, similarly to (1.3.33),

$$\begin{cases} H(P)\phi_n = \varepsilon_n \phi_n, \ \varepsilon_1 \leqslant \cdots \leqslant \varepsilon_{N_{\mathrm{el}}} \\ \phi_n^* \phi_m = \delta_{nm}, \\ P = \displaystyle\sum_{n=1}^{N_{\mathrm{el}}} \phi_n \phi_n^*, \end{cases} \tag{1.5.5}$$

where $H(P) = H_0 + \nabla E_{\mathrm{nl}}(P)$ is the Hamiltonian matrix. The simplest version works, in its original version [146, 176], as follows: if $P^k$ is the current iterate of the algorithm, $P^{k+1}$ is found by solving the eigenproblem

$$H(P^k)\phi_n^k = \varepsilon_n^k \phi_n^k, \quad (\phi_n^k)^* \phi_m^k = \delta_{nm} \tag{1.5.6}$$

with the $\varepsilon_n^k$ sorted in nondecreasing order, and building $P^{k+1}$, assuming the strong *Aufbau* principle, as

$$P^{k+1} = \sum_{n=1}^{N_{\mathrm{el}}} \phi_n^k (\phi_n^k)^*. \tag{1.5.7}$$

Physically, this method (known as the Roothaan algorithm in the literature) can be seen as generating a first mean-field, computing the orbitals of the electrons in this mean-field, update the mean-field and iterate, until convergence. This is this aspect of the problem which makes it nonlinear, and it suggests fixed-point-like iterations where we successively solve eigenproblems until self-consistency is reached. SCF

algorithms can also be improved using preconditioners for the eigenproblem solved at each iteration. The kinetic preconditioner mentioned above is usually sufficient for this part of the algorithm. This basic procedure converges for systems where the nonlinearity is weak, but fails to converge otherwise (see [41] for a comprehensive mathematical analysis of this behaviour when the functional $E$ is a sum of a linear and a quadratic term in $P$, which is the case for the Hartree–Fock model, see also [124]).

A solution to overcome this convergence issue is to use a *damped* version of this algorithm, presented in Algorithm 1.2 and represented in Figure 1.8. Note that it assumes the strong *Aufbau* principle to hold and uses the same retraction $R$ than in Algorithm 1.1.

ALGORITHM 1.2 – Damped SCF algorithm

**Data:** $P^0 \in \mathcal{M}_{N_{\mathrm{el}}}$
**while** *convergence not reached* **do**

$$\text{solve } \begin{cases} H(P^k)\phi_n^k = \varepsilon_n^k \phi_n^k, \ \varepsilon_1^k \leqslant \cdots \leqslant \varepsilon_{N_{\mathrm{el}}}^k < \varepsilon_{N_{\mathrm{el}}+1}^k \\ (\phi_n^k)^* \phi_m^k = \delta_{nm}, \end{cases} ;$$

$$\widetilde{P}^k = \sum_{n=1}^{N_{\mathrm{el}}} \phi_n^k \left(\phi_n^k\right)^*;$$

$$P^{k+1} = R\Big(P^k + \beta \mathbf{\Pi}_{P^k}\Big(\widetilde{P}^k - P^k\Big)\Big);$$

**end**

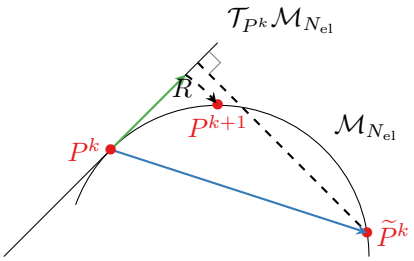



FIGURE 1.8 – Schematic view of the damped SCF algorithm.

Another solution to overcome the convergence issues of the Roothaan algorithm is to *mix* the iterates to accelerate convergence, most of the time combined with damping. This gives rise to a variety of SCF-type algorithms, among which Broyden-like and Anderson-like mixing algorithms [54, 108, 143, 169, 187], the Direct Inversion in the Iterative Space (DIIS) algorithm [118, 167, 168], the Optimal Damping Algorithm (ODA) [40], and the Energy-DIIS (EDIIS) algorithm combining the latter two approaches [119] are the most used nowadays.

### 1.5.4 Direct minimization or SCF? – Results from Chapter 2

While the convergence of several SCF and direct minimization algorithms has been analysed from a mathematical point of view (see *e.g.* [41, 54, 124, 138, 174, 194, 206] and references therein), the two approaches have not been compared in a systematic way. The purpose of Chapter 2 is to contribute to fill this gap, by focusing on very simple representatives of each class, namely the projected gradient descent (Algorithm 1.1) and the damped SCF iteration (Algorithm 1.2). We emphasize that neither of these two algorithms is a practical choice as is. The SCF iterations should be accelerated (for instance using the DIIS acceleration technique), and the gradient information in direct minimization methods should rather be used as part of a quasi-Newton method (such as the L-BFGS algorithm [154]). Depending on the exact problem at hand, all these methods should also be preconditioned to avoid issues related to small mesh sizes (which leads to a divergence of the kinetic energy term) and/or large computational domains (which can lead to a divergence of the Coulomb energy, or the confining potential).

In Chapter 2, we therefore study the minimization problem

$$\min_{P \in \mathcal{M}_{N_{\mathrm{el}}}} E(P), \tag{1.5.8}$$

with

$$\mathcal{M}_{N_{\rm el}} = \big\{P \in \mathcal{H},\ P = P^T,\ \mathrm{Tr}(P) = N_{\rm el},\ P^2 = P\big\}, \tag{1.5.9}$$

where $\mathcal{H} = \mathbb{R}^{N_{\rm b}\times N_{\rm b}}$ is endowed with the Frobenius inner product $\langle A, B\rangle_{\rm F} = \mathrm{Tr}(A^T B)$, the extension to complex matrices being immediate. $E : \mathcal{H} \to \mathbb{R}$ is typically of the form (1.5.1). We assume enough regularity on the functional $E$ and the existence of a nondegenerate minimizer $P_*$: there exists some positive constant $\eta$ such that

$$E(P) \geqslant E(P_*) + \eta\|P - P_*\|_{\rm F}^2 \quad \text{for } P \text{ in a neighbourhood of } P_*. \tag{1.5.10}$$

We then derive first- and second-order optimality conditions:

- The first-order condition reads, as mentioned before, $\mathbf{\Pi}_{P_*} H(P_*) = 0$, which is the same as $[H(P_*), P_*] = 0$, where $H(P) = \nabla E(P)$ is the Hamiltonian matrix. This condition traduces that $P_*$ is a minimizer on the constraint manifold: the energy cannot be decreased at first-order unless we allow leaving the manifold $\mathcal{M}_{N_{\rm el}}$.
- The second-order condition is obtained by linearization and reads

$$\forall\ X \in \mathcal{T}_{P_*}\mathcal{M}_{N_{\rm el}}, \quad \langle X, (\mathbf{\Omega}_* + \boldsymbol{K}_*)X\rangle_{\rm F} \geqslant \eta\|X\|_{\rm F}^2, \tag{1.5.11}$$

  where $\boldsymbol{K}_* = \mathbf{\Pi}_{P_*}\nabla^2 E(P_*)\mathbf{\Pi}_{P_*}$ is the Hessian of the energy projected onto $\mathcal{T}_{P_*}\mathcal{M}_{N_{\rm el}}$ and

$$\forall\ X \in \mathcal{T}_{P_*}\mathcal{M}_{N_{\rm el}}, \quad \mathbf{\Omega}_* X = -[P_*, [H(P_*), X]] \tag{1.5.12}$$

  represents the influence of the curvature of $\mathcal{M}_{N_{\rm el}}$. This condition translates the nondegeneracy of the minimizer $P_*$. Indeed, the second-order optimality of a nondegenerate unconstrained minimization problem is that the Hessian of the objective function is positive definite. Here this condition is modified by the constraints. Moreover, $\mathbf{\Omega}_*$ has the remarkable property that its smallest eigenvalue is the gap between the highest occupied and the lowest unoccupied eigenvalues of the self-consistent Hamiltonian $H(P_*)$. In the linear case, $\boldsymbol{K}_* = 0$ and the second-order optimality condition is therefore equivalent to the strong *Aufbau* principle. In general, the sign of $\boldsymbol{K}_*$ is not known and it is therefore difficult to derive the optimality condition from the strong *Aufbau* principle. However, for the reduced Hartree–Fock or the Gross–Pitaevskii models, we have $\langle X, \boldsymbol{K}_* X\rangle_{\rm F} \geqslant 0$ on $\mathcal{T}_{P_*}\mathcal{M}_{N_{\rm el}}$ and the strong *Aufbau* principle is, in these cases, a sufficient (but not necessary) condition for the second-order optimality condition.

Using these two optimality conditions, we examine the convergence of two simple representatives in the classes of direct minimization and SCF algorithms: the gradient descent described in Algorithm 1.1 and the damped SCF described in Algorithm 1.2. Under assumptions that are made precise in Chapter 2, we then prove the following theorems, which give the convergence rate of the algorithms we consider as the spectral radius $r$ of some operators.

**Theorem 1.2.** *Under suitable assumptions, if $P^0 \in \mathcal{M}_{N_{\rm el}}$ is close enough to $P_*$, Algorithm 1.1 linearly converges to $P_*$ for $\beta > 0$ small enough, with asymptotic rate $r(1 - \beta J_{\rm grad})$ where $J_{\rm grad} = \mathbf{\Omega}_* + \boldsymbol{K}_*$.*

**Theorem 1.3.** *Under suitable assumptions and if the strong* Aufbau *principle holds, then, for $\beta > 0$ small enough and $P^0 \in \mathcal{M}_{N_{\rm el}}$ close enough to $P_*$, Algorithm 1.2 linearly converges to $P_*$, with asymptotic rate $r(1 - \beta J_{\rm SCF})$ where $J_{\rm SCF} = 1 + \mathbf{\Omega}_*^{-1}\boldsymbol{K}_*$.*

In short, if the step, or damping parameter, $\beta$ is small enough, then both algorithms converge to $P_*$ if the initial point is close enough to it. In particular, we find that the convergence rates depend on the spectral radius of operators (acting on $\mathbb{R}^{N_{\rm b}\times N_{\rm b}}$) of the form $1 - \beta J$, where $J = J_{\rm grad} = \mathbf{\Omega}_* + \boldsymbol{K}_*$ for the gradient descent and $J = J_{\rm SCF} = 1 + \mathbf{\Omega}_*^{-1}\boldsymbol{K}_*$ for the SCF algorithm (under the additional assumption that the strong *Aufbau* principle is satisfied). These results have several consequences, which enable for a better understanding of the minimization problem (1.5.8).

In the linear case (*i.e.* $E_{\rm nl} = 0$), we have that $\boldsymbol{K}_* = 0$ and the SCF algorithm converges in one iteration: we only need to diagonalize the Hamiltonian once because the self-consistency nature of the problem disappears, which is consistent with $J_{\rm SCF} = 1$. Regarding the Gradient Descent, we have in this case $J_{\rm grad} = \mathbf{\Omega}_*$ and the conditioning of the system is linked to the highest eigenvalue of $\mathbf{\Omega}_*$, which

blows up when the discretization is refined. This is a common issue in numerical linear algebra that one usually solves by preconditioning. When adding the nonlinear term $E_{\text{nl}}$, the situation is summarized in Table 1.2. If no link could be drawn *a priori* between SCF and Gradient Descent, it appears that the SCF can actually be interpreted as a "matrix splitting" of the Gradient Descent (in linear algebra, a matrix splitting consists in solving $(A+B)x = b$ by solving $(1 + A^{-1}B)x = A^{-1}b$). Moreover, while the Gradient Descent is sensitive to the spectral radius of $\boldsymbol{\Omega}_*$, the SCF seems to be sensitive to the inverse of the smallest eigenvalue of $\boldsymbol{\Omega}_*$, *i.e.* the gap. This explains why SCF algorithms struggle to converge for systems with small gap, a well-known issue in quantum chemistry.

| Problem | characteristic matrix |
|---|---|
| Linear eigenvalue problem | $\boldsymbol{\Omega}_*$ |
| Damped SCF | $1 + \boldsymbol{\Omega}_*^{-1}\boldsymbol{K}_*$ |
| Gradient Descent | $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ |

TABLE 1.2 – Condition matrices of the different problems we consider.

We then consider the following question: *In practice, should the SCF or direct minimization class of algorithms be preferred?* The answer depends not only on the convergence rate studied in this chapter, but also on the cost of each step, and the robustness of the algorithm. We examine two prototypical situations.

In quantum chemistry using Gaussian basis sets [161] (see also Chapter 6 for more details on basis sets in quantum chemistry) to solve the Hartree–Fock model or Kohn–Sham DFT using hybrid functionals, the rate-limiting step is often the computation of the Hamiltonian matrix $H(P)$. In this case, an iteration of a gradient descent and a damped SCF algorithm are of roughly equal cost. In most cases, solutions for isolated molecules satisfy the *Aufbau* principle, and the damped SCF algorithm, suitably robustified (for instance using the ODA algorithm) and accelerated (for instance with the DIIS algorithm), converges reliably and efficiently towards a solution. Direct minimization algorithms are then only useful in the cases where local or semilocal functionals are used [177] and the *Aufbau* principle is violated, or when SCF algorithms tend to converge to saddle points (for instance for computations involving spin).

In condensed-matter physics using plane-wave basis sets to solve Kohn–Sham DFT with local or semilocal functionals, the matrices $P$ and $H$ are not stored explicitly. Solving the linear eigenproblem is then done using iterative block eigensolvers, which can be understood as specialized direct minimization algorithms in the case of a linear energy functional $E(P) = \operatorname{Tr}(H_0 P)$. In this case, direct minimization algorithms effectively merge the two loops of the SCF and linear eigensolver, and should therefore be more efficient. Another interest of direct minimization algorithms is their robustness, as the choice of a step size can be made in order to minimize the energy. Despite this, direct minimization algorithms are rarely used in condensed-matter physics. The main reason seems to be that challenging problems are often metallic in character, and require the introduction of a finite temperature. Direct minimization algorithms then need to optimize over the occupations as well as the orbitals, a significantly more complex task, see for instance [28, 56, 75, 145]. A thorough comparison of the performance and robustness of direct minimization and self-consistent approach for these systems would be an interesting topic of inquiry. A number of implementation "tricks" commonly used to accelerate the convergence of iterative eigensolvers (for instance, using a block size larger than the number of electrons) might also play a big role in performance comparison for the two classes of algorithms: understanding how to generalize these to direct minimization would be interesting.

## 1.6 Estimating the error

One of the main challenges in modern computational chemistry (and, more generally, numerical simulation of physical systems) is the estimation of the error from the results of numerical simulations: *knowing that the output of the simulation is only an approximation of the real solution, how can we estimate the actual error we make? Can we also estimate the error we commit on quantities of interest, such as the energy or the interatomic forces?* Some answers to these questions can be found with what is known as *a posteriori* error estimates in mathematics. *A posteriori* estimates differ from *a priori* estimates as they should be computable without any knowledge of the actual, continuous or discrete,

solution, for a cost not significantly more important than that of the computation performed to obtain the current approximation. They should ideally also be guaranteed, meaning that we can rigorously prove that they hold under not too strong assumptions. In this section, we briefly describe the state of the art of *a posteriori* estimates for linear and nonlinear eigenvalue problems, with a particular focus on DFT. We then describe the results in Chapter 3, which are a first step towards the estimation of errors on quantities of interest for nonlinear KS-DFT models.

### 1.6.1 Sources of error

When trying to estimate the error between the output of a numerical simulation and the actual solution we are looking for, it is useful to recall that there are different sources of error, each of them requiring a different approach. We give here a list of the different sources of error that arise in DFT calculations:

**Model error** It corresponds to the error made by the choice of the model we use: even before discretizing and solving the equations, there is already an error due to the modelling choices. In KS-DFT, it mainly comes from the choice of the exchange-correlation energy functional and pseudopotentials.

**Discretization error** This error source is due to the transformation of the continuous equations into discrete ones. In the case of KS-DFT for crystals, it is the error we make by using finite approximations with Fourier modes and discretized Brillouin zone.

**Algorithmic error** This is the error made when we stop for instance eigenvalue solvers or SCF algorithms when some convergence threshold is reached.

**Numerical error** This error is a consequence of floating point arithmetic.

There are of course other sources of error (bugs, hardware failures, ...), but these lay out of the scope of numerical analysis. Knowing the contribution of each sources of error is a crucial step towards more efficient and robust, adaptive, algorithms: as long as the main source of error is due for instance to the size of the discretization space, there is no need to choose tight convergence thresholds. Such adaptive algorithms are nowadays an active field of research, and we mention some recent advances below.

### 1.6.2 The linear case

For general elliptic linear source problems, efficient *a posteriori* estimates have been introduced since the 50s, based for instance on the theory of equilibrated fluxes from Prager and Synge [166] (see [26, 64, 73, 121] and references therein) or on gradient recovery type estimate (see [204] and references therein). When it comes to estimating the error for eigenvalue problems, the computation of guaranteed error bounds seems more difficult. Following works from Kato [110], Forsythe [74], Weinberger [201] or Bazley and Fox [16], several works in the last decades presented estimations of simple eigenvalues. See for instance [47, 68, 106, 107, 122, 137, 141] and references therein. *A posteriori* estimates for both eigenvalues and eigenvectors can also be found in [35] for conforming discretizations and in [36] for a more general framework, including nonconforming discretizations. See also [153, Chapter 10] for a recent monograph on the subject. With specific focus on electronic structure calculation, guaranteed error bounds for linear eigenvalues equations are presented in [100].

However, the above only holds for simple eigenvalues. As degenerate, or near-degenerate, eigenvalues often appear in practice (in particular in quantum chemistry and related fields due to symmetries), dealing with multiples eigenvalues is a fundamental task. In [35, 36], the estimates depend on the gap between the estimated eigenvalue and the surrounding ones, which deteriorates the estimates for (near) degenerate estimates. *A posteriori* estimates for clusters of eigenvalues have been proposed for instance in [24] for Crouzeix–Raviart nonconforming finite elements or in [77] for the discontinuous Galerkin method. *A posteriori* error estimates for conforming approximations of eigenvalue clusters of second-order self-adjoint elliptic operators with compact resolvent have also been derived in [37].

There is thus a substantial literature on the *a posteriori* error estimation of linear eigenvalue problems. For nonlinear problems, the situation is more difficult and we mention below some of the few existing results.

### 1.6.3 The nonlinear case

For nonlinear eigenvalue equations, such as the Kohn–Sham model where the Hamiltonian to diagonalize depends on the orbitals themselves, no rigorous and certified error bounds are known at the moment to estimate the discretization error due to the plane-wave approximation. A few results still exist for simpler models or other discretizations.

The simplest nonlinear equation one can consider in quantum physics is the 1D Gross–Pitaevskii equation, which is of high interest in the study of Bose–Einstein condensates. It is the particular case of (1.4.4) with $H_\rho = -\Delta + V + \rho$ and $N_{\rm el} = 1$. It therefore reads,

$$\begin{cases} -\Delta\phi + V\phi + |\phi|^2\phi = \varepsilon\phi, \\ \|\phi\|_{\mathrm{L}^2_\#} = 1. \end{cases} \tag{1.6.1}$$

It is well known that, if $V$ is in $\mathrm{H}^s_\#(\mathbb{R}^3, \mathbb{C})$ for $s > d/2$, then there is a unique, real-valued, positive solution to the Gross–Pitaevskii equation, which belongs in addition to $\mathrm{H}^{s+2}_\#(\mathbb{R}^3, \mathbb{C})$ (see [30, Section 3 and Appendix]). It also solves the minimization problem

$$\mathcal{E}^{\rm GP}_* = \min\Big\{\mathcal{E}^{\rm GP}(\psi),\ \psi \in \mathrm{H}^1_\#(\mathbb{R}^3, \mathbb{C}),\ \|\psi\|_{\mathrm{L}^2_\#} = 1\Big\}, \tag{1.6.2}$$

where

$$\mathcal{E}^{\rm GP}(\psi) = \int_{\mathbb{R}^3} |\nabla\psi|^2 + \int_{\mathbb{R}^3} V|\psi|^2 + \frac{1}{2}\int_{\mathbb{R}^3} |\psi|^4. \tag{1.6.3}$$

In [70], an error bound is developed, but the computational cost of evaluating this error bound in this contribution is quite extensive. We also refer to [53], where *a posteriori* error estimates are developed for finite elements approximations of a class of nonlinear eigenvalue problems, including the Gross–Pitaevskii equation. In [34], a rigorous error bound on the Gross–Pitaevskii energy is proposed. It has the particularity that it is valid at each step of the discrete self-consistent iterations and it can therefore be used to design an adaptive algorithm which automatically refines the discretization space along the iterations when the error due to the discretization becomes larger than the error due to the iterative procedure.

Such adaptive methods have also been developed in the last decade for linear and nonlinear models with finite elements approximations, see [52, 59, 142, 205] and references therein. In the context of plane-wave discretization, see [58] for an adaptive method for linear elliptic eigenvalue problems and [136] for a recent adaptive method for Kohn–Sham models. To refine automatically the discretization space along the iterations, robust and guaranteed error bounds for Kohn–Sham models are still missing. Most results are also limited to the estimation of the error on the orbitals or the energy, and no results exist at the moment on the estimation of the error for quantities that are of practical interest, such as the interatomic forces. Chapter 3 proposes a first step in these directions; we summarize the main results below.

### 1.6.4 Practical error estimates for plane-wave KS-DFT – Results from Chapter 3

This chapter focuses on providing practical error estimates for the discretization error of numerical approximations of electronic structure calculation. To this end, we use a general approach based on a linearization of the Kohn–Sham equations. It is instructive to start by comparing our approach to those used in a general context. Assume we want to find $x \in \mathbb{R}^n$ such that $f(x) = 0$, for some nonlinear function $f : \mathbb{R}^n \to \mathbb{R}^n$ (the residual). Near a solution $x_*$, we have $f(x) \approx f'(x)(x - x_*)$, and therefore, if $f'(x)$ is invertible, we have the error-residual relationship

$$x - x_* \approx f'(x)^{-1} f(x). \tag{1.6.4}$$

This is the same approximation that leads to the Newton algorithm. Assume now that we want to compute a real-valued quantity of interest $A(x_*)$, where $A : \mathbb{R}^n \to \mathbb{R}$ is a $C^1$ function (*e.g.* the energy, a component of the interatomic forces, of the density, ...); then we have the approximate equality with computable right-hand side:

$$A(x) - A(x_*) \approx \nabla A(x) \cdot \big(f'(x)^{-1} f(x)\big). \tag{1.6.5}$$

From here, we obtain the simple first estimate

$$|A(x) - A(x_*)| \leqslant |\nabla A(x)| \big\|f'(x)^{-1}\big\|_{\mathrm{op}} |f(x)|, \tag{1.6.6}$$

where $|\cdot|$ is any chosen norm on $\mathbb{R}^n$, and $\|\cdot\|_{\mathrm{op}}$ is the induced operator norm on $\mathbb{R}^{n\times n}$ (note that $\nabla A(x) \in \mathbb{R}^n$ and $f'(x) \in \mathbb{R}^{n\times n}$). This approximate bound can be turned into a rigorous one using information on the second derivatives of $f$; see for instance [181]. In extending this approach to Kohn–Sham models, we encountered various difficulties which led to several findings.

First, the structure of our problem is not easily formulated as above because of the presence of constraints and degeneracies. We solve this using the geometrical framework of Chapter 2 to identify the appropriate analogue to the Jacobian $f'(x)$: we have proved that the operator $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ defined in (1.5.11) is the Jacobian of the residual map $R : P \mapsto \boldsymbol{\Pi}_P H(P)$, which vanishes at $P = P_*$. Our approach therefore relies on the first-order approximation

$$P - P_* \approx (\boldsymbol{\Omega}_* + \boldsymbol{K}_*)^{-1} R(P), \tag{1.6.7}$$

where $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ plays the role of $f'$ in the general context described above, $P_*$ is a reference solution in a large plane-wave reference space (ideally the exact one) and $P$ is an approximate solution from a smaller variational space. This approximation is found to be actually very good, even for energy cut-offs as small as 5 hartree, but not suitable in practice as the inversion of $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ in the reference space cannot be performed in a reasonable computational time.

Then, choosing the right norm is not obvious in this context. For problems involving partial differential equations, it is natural to consider Sobolev-type norms, with the aim of making the Jacobian a bounded operator between the relevant function spaces. We explore different choices and their impacts on the error bounds. However, in our case, the naive operator norm inequalities

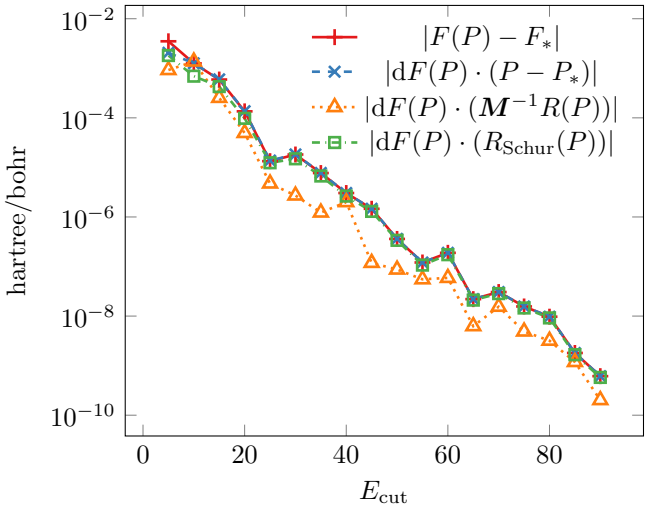
$$\big|\nabla A(x) \cdot \big(f'(x)^{-1} f(x)\big)\big| \leqslant |\nabla A(x)| \big\|f'(x)^{-1}\big\|_{\mathrm{op}} |f(x)|, \tag{1.6.8}$$

where $A$ represents the interatomic forces, are very often largely suboptimal (by more than five orders of magnitude), even with appropriate norms, and we quantify this on representative examples. The reason is that, for plane-wave calculations, the discretization error is mostly made up of high frequency components, whereas $\nabla A$ is mostly supported on low frequencies for interatomic forces. This results into the bound in (1.6.8) being very suboptimal.

We therefore follow another natural idea which consists in replacing the error by the (easily computable) preconditioned residual to compute estimates similar to (1.6.5), instead of computing upper bounds. This yields reasonable estimates of the errors on the quantities of interest we investigate, but they are not systematic upper bounds, nor asymptotically valid. This is again due to the error and the preconditioned residual differing mainly on low frequencies, making the approximation (1.6.5) not accurate enough. We then build a Schur complement approach based on a low/high frequency splitting to approximate the inverse of $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$. This systematically improves the estimation of the error on the low frequencies and gives reliable error estimates on $A$ at reasonable cost: the Jacobian $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ only needs to be inverted on the low frequencies (instead of the full reference space), the high frequency components being approximated with a (diagonal) kinetic preconditioner and then coupled through a Schur complement. This adds a computational work no more expensive than the SCF algorithm used to compute $P$.

In short, the main result of Chapter 3 lies in the derivation of an efficient, asymptotically accurate, way of approximating $\nabla A(x) \cdot \big(f'(x)^{-1} f(x)\big)$ using the specific structure of the residual $f(x)$ in a plane-wave discretization, where $A$ represents the interatomic forces of the system (see Figure 1.9). This approximation can then be used either to approach the actual error $A(x) - A(x_*)$ or to improve $A(x)$ by computing $A(x) - \nabla A(x) \cdot \big(f'(x)^{-1} f(x)\big)$, which is a better approximation of $A(x_*)$. This work is a first

step towards robust and guaranteed error estimates for Kohn–Sham DFT calculations. There are still limits to it, the main one being the nonguaranteed nature of our estimates, even though their asymptotic accuracy still allows for practically useful results. They also require defining a "coarse" grid on which the main SCF calculations are performed and the low frequency components of the error are approximated, as well as a "fine" grid to perform the Schur splitting on high frequencies. The choice of the good ratio between these two grids is not an easy task and, at the moment, it is most of the time made empirically. Note also that the inversion of $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ on the low frequency space is only defined for gapped system. However, it is possible to extend its definition to metallic systems (see for instance [99] or Chapter 4), making possible the extension of this work to such systems.




Source: Chapter 3

FIGURE 1.9 – Estimation of the error on the interatomic forces for the FCC silicon crystal, when the plane-wave cut-off $E_{\text{cut}}$ is increased. Here, $\mathrm{d}F$ represents $\nabla A$. (Solid line) The actual error we want to estimate. (Crosses) The linearization of the error on the forces with the real error $P - P*$ matches rapidly. (Triangles) The linearization with the preconditioned residual $\boldsymbol{M}^{-1}R(P)$ fails because of the low frequency support of $\mathrm{d}F(P)$. (Squares) We can recover the low frequency components of the error with a Schur complement, yielding an efficient approximation of the error $F(P) - F_*$.

## 1.7 Density functional perturbation theory

According to what we described in the previous sections of this introduction, KS-DFT aims at computing the electronic ground-state of a given system of interest. Only few quantities of interest (*e.g.* the ground-state density and energy) do not require the computation of derivatives of the ground-state with respect to external perturbations. The others, such as interatomic forces, (hyper)polarizabilities, magnetic susceptibilities, phonons spectra, or transport coefficients, correspond physically to the response of the ground-state to nuclear positions or external electromagnetic fields, and their mathematical expressions *a priori* involve derivatives of the ground-state with respect to these parameters. More recent applications, such as the design of machine-learned exchange-correlation energy functionals, also require the computation of derivatives of the ground-state with respect to parameters, such as the ones defining the exchange-correlation functional [109, 113, 129].

Thanks to the Hellmann–Feynman theorem [96], computing the interatomic forces as the derivative of the energy with respect to the atomic displacements can be done directly from the ground-state orbitals (see Chapter 3). However, general quantities of interest are based on more involved types of derivatives and require the response of the orbitals with respect to an external potential perturbation. This is usually done *via* standard first-order perturbation theory, a framework known in the field as density functional perturbation theory (DFPT) [15, 81, 82, 85], with applications detailed for instance in [14] for phonons in solid-state physics or in [155] for quantum chemistry. See also [45] for a mathematical analysis of DFPT within the reduced Hartree–Fock approximation. Although the practical implementation of first- and higher-order derivatives computed by DFPT in electronic structure calculation software can

be greatly simplified by Automatic Differentiation techniques [86], the efficiency of the resulting code crucially depends on the efficiency of a key building block: the computation of the linear response $\delta\rho$ of the ground-state density to an infinitesimal variation $\delta V$ of the total Kohn–Sham potential. Achieving efficient calculations of $\delta\rho$ for metallic systems is the main subject of Chapter 4, and, before presenting the results of this chapter, we briefly introduce the framework in which we work.

### 1.7.1 The Kohn–Sham equations at finite temperature

For the sake of clarity, we detail the equations for a periodic system of $N_{\mathrm{el}}$ electrons without interactions, at finite temperature $T > 0$ and taking spin into account:

$$H\phi_n = \varepsilon_n\phi_n, \quad \varepsilon_1 \leqslant \varepsilon_2 \leqslant \cdots, \quad \langle\phi_n, \phi_m\rangle_{\mathrm{L}^2_\#} = \delta_{nm}, \quad \rho(\boldsymbol{r}) = \sum_{n=1}^{+\infty} f_n|\phi_n(\boldsymbol{r})|^2, \quad \sum_{n=1}^{+\infty} f_n = N_{\mathrm{el}}, \tag{1.7.1}$$

where $H = -\frac{1}{2}\Delta + V$. Here, every orbital $\phi_n$ has occupation number $f_n$ and energy $\varepsilon_n$. At finite temperature $T > 0$, $f_n$ is a real number in the interval $[0, 2]$ and we have

$$f_n = f\left(\frac{\varepsilon_n - \varepsilon_{\mathrm{F}}}{T}\right), \tag{1.7.2}$$

where $f$ is a fixed analytic *smearing* function (for instance $f(x) = 2/(1 + \mathrm{e}^x)$ is twice the Fermi–Dirac distribution). The Fermi level $\varepsilon_{\mathrm{F}}$ is then uniquely defined by the constraint $\sum_{n=1}^{+\infty} f_n = N_{\mathrm{el}}$. When $T \to 0$, $f((\cdot - \varepsilon_{\mathrm{F}})/T) \to 2 \times \mathbf{1}_{\{\cdot < \varepsilon_{\mathrm{F}}\}}$ in the sense of distributions, and only the first $N_{\mathrm{p}} = N_{\mathrm{el}}/2$ energy levels for which $\varepsilon_n < \varepsilon_{\mathrm{F}}$ are occupied by two electrons of opposite spins (see Figure 1.10): $f_n = 2$ for $n \leqslant N_{\mathrm{p}}$ and $f_n = 0$ for $n > N_{\mathrm{p}}$. The need to introduce a numerical temperature $T$ (usually much higher than the physical temperature) arises when dealing with the Brillouin zone discretization of metallic systems, see [39, 125] for more details.




FIGURE 1.10 – The occupation numbers $f_n$ for $T = 0$ (left) and $T > 0$ (right).

Finally, note that only a finite number $N$ of orbitals needs to be computed. At zero temperature, which is the relevant choice for insulators, $N$ is the number of electron pairs and the $N$ first energy levels are occupied by two electrons of opposite spins ($f_n = 2$). At finite temperature, which is the right setting for metals, every orbital has a fractional occupation number $f_n \in [0, 2]$ but one usually assumes that $N$ can be chosen so that the orbitals $\phi_n$ with $n > N$ have a small enough occupancy to be discarded from the computations.

### 1.7.2 Density functional perturbation theory

DFPT aims at computing the response $\delta\rho$ of the ground-state density with respect to an infinitesimal perturbation of the external potential $\delta V$. Denoting by $F$ the potential-to-density map which, given a potential $V$, associates the density $\rho$ satisfying (1.7.1), we write $\rho = F(V)$. Then, we obtain that

$$\delta\rho = F'(V) \cdot \delta V, \tag{1.7.3}$$

where $F'(V)$ is the derivative of $F$ computed at $V$. In DFT, this operator is known as the *independent-particle susceptibility* operator, denoted by $\chi_0$. It maps $\delta V$ to the first-order variation of the density $\delta\rho$. Denoting $A_{mn} \coloneqq \langle\phi_m, A\phi_n\rangle_{\mathrm{L}^2_\#}$ for a given operator $A$, it holds

$$\delta\rho(\boldsymbol{r}) = (\chi_0\delta V)(\boldsymbol{r}) = \sum_{n=1}^{+\infty}\sum_{m=1}^{+\infty} \frac{f_n - f_m}{\varepsilon_n - \varepsilon_m}\phi_n^*(\boldsymbol{r})\phi_m(\boldsymbol{r})(\delta V_{mn} - \delta\varepsilon_{\mathrm{F}}\delta_{mn}), \tag{1.7.4}$$

where $\delta_{mn}$ is the Kronecker delta, $\delta\varepsilon_{\mathrm{F}}$ is the induced variation in the Fermi level and we use the following convention

$$\frac{f_n - f_n}{\varepsilon_n - \varepsilon_n} = \frac{1}{T} f'\left(\frac{\varepsilon_n - \varepsilon_{\mathrm{F}}}{T}\right) = f'_n. \tag{1.7.5}$$

This formula is formally derived in [14]. We refer also to [44, 98, 125], where this formula is rigourously proven using contour integrals. Charge conservation leads to

$$\int_\Omega \delta\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} = 0 \quad \Rightarrow \quad \delta\varepsilon_{\mathrm{F}} = \frac{\sum_{n=1}^{+\infty} f'_n \delta V_{nn}}{\sum_{n=1}^{+\infty} f'_n}. \tag{1.7.6}$$

The infinite sums in (1.7.4) make this formula unusable as it stands. However, recalling that in practice, only a finite number $N$ of orbitals are computed, one can represent $\delta\rho$ using variations of the occupied orbitals $(\delta\phi_n)_{1\leqslant n\leqslant N}$ and their occupation numbers $(\delta f_n)_{1\leqslant n\leqslant N}$, along with appropriate ansatz and gauge choices. Indeed, a formal differentiation of the relation $\rho(\boldsymbol{r}) = \sum_{n=1}^{N} f_n|\phi_n(\boldsymbol{r})|^2$ yields the ansatz

$$\delta\rho(\boldsymbol{r}) = \sum_{n=1}^{N} 2f_n \times \mathrm{Re}\left(\phi_n^*(\boldsymbol{r})\delta\phi_n(\boldsymbol{r})\right) + \delta f_n|\phi_n(\boldsymbol{r})|^2. \tag{1.7.7}$$

Due to the rotational invariance of the orbitals, there are different ways of choosing the $\delta\phi_n$'s and $\delta f_n$'s, that are all valid as long as the resulting $\delta\rho$ coincides with (1.7.4). A gauge choice therefore has to be made here. This question is the main subject of Chapter 4, which we detail below, after a short remark on interacting systems.

**Remark 1.4** (Self-consistent response and links with Chapter 2)**.** We briefly mention the case of a system with interactions: $\rho$ satisfies the fixed-point equation

$$\rho = F(V + V_{\mathrm{Hxc}}(\rho)) \tag{1.7.8}$$

where $V_{\mathrm{Hxc}}(\rho)$ is the Hartree-exchange-correlation potential. The chain rule yields the implicit equation

$$\delta\rho = F'(V + V_{\mathrm{Hxc}}(\rho)) \cdot (\delta V + K_{\mathrm{Hxc}}(\rho)\delta\rho), \tag{1.7.9}$$

where the Hartree-exchange-correlation kernel $K_{\mathrm{Hxc}}(\rho)$ is the derivative of the map $\rho \mapsto V_{\mathrm{Hxc}}(\rho)$. It is directly linked to the second derivative of the Kohn–Sham energy functional, and thus to the operator $\boldsymbol{K}_*$ we introduced in Chapter 2 (see Section 1.5.4). In addition, $\chi_0$ is also related to $-\boldsymbol{\Omega}_*^{-1}$ at zero temperature (see (1.5.11)). This gives a natural extension of $\boldsymbol{\Omega}_*$ for metallic systems, see [98]. The response of the density to the variation of the *total* potential is finally computed through

$$\delta\rho = \chi_0(\delta V + K_{\mathrm{Hxc}}(\rho)\delta\rho) \quad \Leftrightarrow \quad \delta\rho = (1 - \chi_0 K_{\mathrm{Hxc}}(\rho))^{-1}\chi_0\delta V, \tag{1.7.10}$$

where $1 - \chi_0 K_{\mathrm{Hxc}}(\rho)$ can be proved to be invertible similarly to $1 + \boldsymbol{\Omega}_*^{-1}\boldsymbol{K}_*$. Computing $\delta\rho$ can thus be done using iterative solvers to invert $1 - \chi_0 K_{\mathrm{Hxc}}(\rho)$, which requires efficient and robust applications of the linear operator $\chi_0$, giving another motivation for Chapter 4.

### 1.7.3 Calculations of response properties for metals – Results from Chapter 4

We propose in Chapter 4 a new approach which splits $\delta\phi_n$ into two contributions:

$$\delta\phi_n = \delta\phi_n^P + \delta\phi_n^Q, \tag{1.7.11}$$

where, for $n \leqslant N$, $P$ is the orthogonal projector onto $\mathrm{Span}\,(\phi_m)_{1\leqslant m\leqslant N}$, $Q = 1 - P$ and

- $\delta\phi_n^P \in \mathrm{Ran}\,(P) = \mathrm{Span}\,(\phi_m)_{1\leqslant m\leqslant N}$ are the occupied-occupied contributions, which can be directly computed *via* a sum-over-state formula. Note that this contribution vanishes at zero temperature but, when dealing with finite temperature, gauge choices have to be made;
- $\delta\phi_n^Q \in \mathrm{Ran}\,(Q) = \mathrm{Span}\,(\phi_m)_{m>N}$ are the unoccupied-occupied contributions. These contributions cannot be computed similarly to $\delta\phi_n^P$ but one can show that $\delta\phi_n^Q$ is the unique solution of the so-called Sternheimer equation [188]:

$$Q(H - \varepsilon_n)Q\delta\phi_n^Q = -Q\delta V\phi_n. \tag{1.7.12}$$

  As $n \leqslant N$, this equation is well-posed but possibly very ill-conditioned for $n = N$ if $\varepsilon_{N+1} - \varepsilon_N$ is too small.

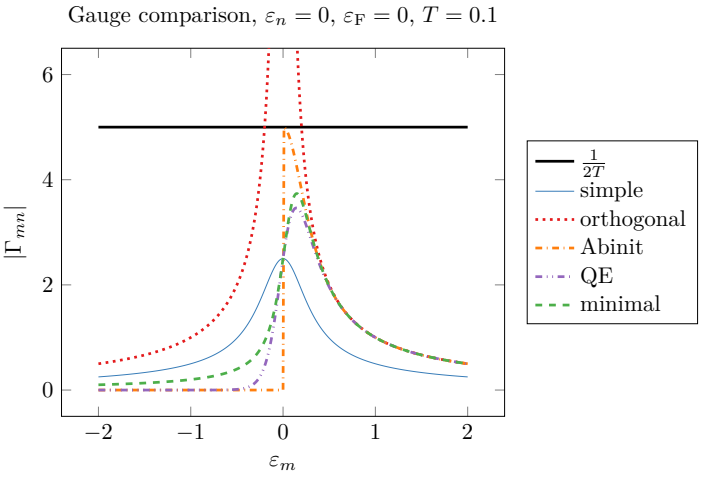
With regard to the calculation of the first contribution $\delta\phi_n^P$, we identify in Chapter 4 that the gauge choice is based on $\Gamma_{mn} = \langle \phi_m, f_n \delta\phi_n \rangle_{\mathrm{L}^2_\#}$ for $1 \leqslant n, m \leqslant N$. We detail them in Chapter 4 and present a summary in Figure 1.11. We review in particular the gauge choices made in Quantum Espresso (QE) [14, 78] and Abinit [84, 175]. We also introduce the minimal gauge, which aims at minimizing the contributions $\langle \phi_m, \delta\phi_n \rangle_{\mathrm{L}^2_\#} = \Gamma_{mn}/f_n$ and is implemented by default in DFTK. From (1.7.4), we can see that the growth of $\delta\rho$ with respect to $\delta V$ cannot be smaller than that of $f_n'(\delta V_{nn} - \delta\varepsilon_{\mathrm{F}})$ with respect to $\delta V$, which is of order $\max_{x\in\mathbb{R}} \frac{1}{T}|f'(x)| = \frac{1}{2T}$. This is the intrinsic limit on the conditioning of the problem and this bound is also achieved by all the gauge choices but the orthogonal one, which is inspired from the zero temperature case. Indeed, all the gauge choices except the orthogonal one are of the form $\Gamma_{mn} = \alpha_{mn}\Delta_{mn}$ with $\alpha_{mn} \in [0, 1]$ and

$$\Delta_{mn} = \frac{f_n - f_m}{\varepsilon_n - \varepsilon_m} \delta V_{mn}. \tag{1.7.13}$$

They thus satisfy

$$|\Gamma_{mn}| \leqslant |\Delta_{mn}| \leqslant \max_{x\in\mathbb{R}} \frac{1}{T}|f'(x)||\delta V_{mn}| = \frac{1}{2T}|\delta V_{mn}|, \tag{1.7.14}$$

and if we make an error on $\delta V$, it is amplified at most by a factor of the order $\frac{1}{2T}$.




| Gauge choice | $\Gamma_{mn}$ | Stability |
|---|---|---|
| simple | $\frac{1}{2}\Delta_{mn}$ | yes |
| orthogonal | $\frac{f_n}{\varepsilon_n - \varepsilon_m}\delta V_{mn}$ | no |
| QE | $f_{\mathrm{FD}}\left(\frac{\varepsilon_n - \varepsilon_m}{T}\right)\Delta_{mn}$ | yes |
| Abinit | $\mathbf{1}_{\{f_n > f_m\}}\Delta_{mn}$ | yes |
| minimal | $\frac{f_n^2}{f_n^2 + f_m^2}\Delta_{mn}$ | yes |

FIGURE 1.11 – Comparing different gauge choices for $\delta V_{mn} = 1$ for any $m$, $n$. Except the orthogonal gauge, all contributions $\Gamma_{mn}$ are bounded by $\max_{x\in\mathbb{R}} \frac{1}{T}|f'(x)| = \frac{1}{2T}$.

We now detail the computation of $\delta\phi_n^Q$. The same strategy as the one we used for $\delta\phi_n^P$ cannot be applied as this time the $\phi_m$'s for $m > N$ are unknown. However, some of them are. These $N_{\mathrm{ex}}$ extra bands can be divided into two categories:

- some have been discarded because they have a too small occupation number but they are exact eigenvectors (up to the solver tolerance). This is typically the case for the lower-energy extra states.
- the others (typically the higher-energy extra bands) have not been fully converged but have been used to enhance the successive diagonalizations of the SCF algorithm. Adding such unconverged extra bands is also not very expensive when the diagonalizations are performed with block-based algorithms, such as LOBPCG [115].

This additional information can be used to accelerate the computation of $\delta\phi_n^Q$ as follows. A direct approach solves the Sternheimer equation

$$Q(H - \varepsilon_n)Q\delta\phi_n^Q = -Q\delta V\phi_n, \tag{1.7.15}$$

with iterative solvers restricted to $\mathrm{Ran}\,(Q)$. However, as we already mentioned, conditioning issues can arise for $n = N$ if the difference $\varepsilon_{N+1} - \varepsilon_N$ is too small. We propose here a new solution to overcome this issue, based on a Schur complement and the usage of the extra unoccupied states. We assume that the number of computed bands $N + N_{\mathrm{ex}}$ is larger than the number of occupied states $N$ and that we

trust $\Phi = (\phi_1, \dots, \phi_N)$ but not $\widetilde{\Phi} = (\widetilde{\phi}_{N+1}, \dots, \widetilde{\phi}_{N+N_{\text{ex}}})$ to be eigenvectors. We assume in addition that $(\Phi, \widetilde{\Phi})$ forms an orthonormal family and that $\langle \widetilde{\Phi}, H\widetilde{\Phi} \rangle_{\mathrm{L}^2_\#}$ is a diagonal matrix whose elements, denoted by $E_{\text{ex}} = (\widetilde{\varepsilon}_n)_{n=N+1,\dots,N+N_{\text{ex}}}$, are not all exact eigenvalues. This is for instance the case if the successive eigenproblems of the SCF are solved with the LOBPCG algorithm [115]. Then, $Q(H - \varepsilon_n)Q$ can be written in the decomposition $\mathrm{Ran}\,(Q) = \mathrm{Ran}\,(T) \oplus \mathrm{Ran}\,(R)$ with $T$ the orthogonal projector onto $\mathrm{Span}\,(\widetilde{\phi}_m)_{N<m\leqslant N+N_{\text{ex}}}$ and $R = Q - T$, as

$$Q(H - \varepsilon_n)Q = \begin{pmatrix} E_{\text{ex}} - \varepsilon_n & RHT \\ THR & R(H - \varepsilon_n)R \end{pmatrix} \tag{1.7.16}$$

where $\varepsilon_n$ does not appear in the off diagonal terms because $RT = 0$. See Figure 1.12 for a graphical representation.

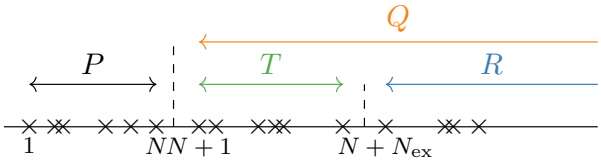
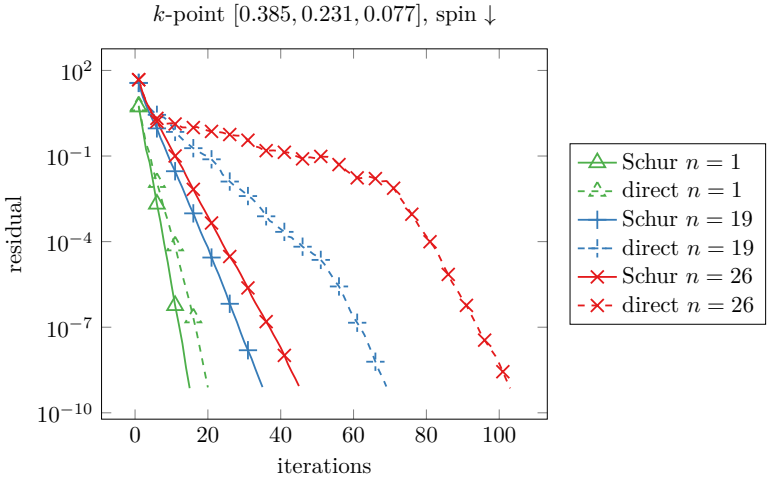



FIGURE 1.12 – Graphical representation of the Schur decomposition to solve the Sternheimer equation. $P$ is the orthogonal projector onto the occupied states. $Q$ is the orthogonal projector onto the unoccupied states, and we decompose is as the sum of $T$ (extra states which we can use) and $R$ (remaining states).

Therefore, the Sternheimer equation (1.7.15) can be solved *via* a Schur complement method, where the inversion of $E_{\text{ex}} - \varepsilon_n$ is free because it is a diagonal matrix, and $R(H - \varepsilon_n)R$ is hopefully better conditioned than $Q(H - \varepsilon_n)Q$ for $n = N$ as $\varepsilon_{N+N_{\text{ex}}+1} - \varepsilon_N > \varepsilon_{N+1} - \varepsilon_N$. This approach is tested in Chapter 4 on various systems. It reveals itself to be particularly efficient on the numerically challenging Heusler compounds: these transition metals behave like a metal on one spin channel and like an insulator on the other. Response calculations are thus particularly though for such systems. Using the Schur complement to compute $\delta\phi_n^Q$ in these cases reduces the total number of Hamiltonian applications by 40%. We plot in Figure 1.13 the convergence of the Sternheimer solver, with and without the Schur complement, for one particular $k$-point (the behaviour being similar for all the others).




Source: Chapter 4

FIGURE 1.13 – Resolution of the Sternheimer equation (1.7.15) with and without the Schur complement method for a particular $k$-point of $Fe_2MnAl$ (each line correspond to the convergence of the solver for a given eigenvalue $\varepsilon_n$, $n \leqslant N$, the slowest one being associated with the highest selected eigenvalue $\varepsilon_N$). When solving the Sternheimer equation for $n = N$ with a direct method, a plateau clearly appears where the solver encounters difficulties to converge the first eigenvectors because of the small gap $\varepsilon_{N+1} - \varepsilon_N$.

In summary, there are two main contributions in Chapter 4. First, we derive a new common framework for the computation of the response $\delta\rho$ to an infinitesimal perturbation $\delta V$ at finite temperature. Relying on the independent-particle susceptibility $\chi_0$, we show how $\delta\rho$ can be represented by perturbations of the orbitals $\delta\phi_n$ and the occupation numbers $\delta f_n$. Then, we split $\delta\phi_n = \delta\phi_n^P + \delta\phi_n^Q$ into two

contributions. We show how the occupied-occupied contributions $\delta\phi_n^P$ can be explicitly computed with a sum-over-state formula while ensuring numerical stability. Regarding $\delta\phi_n^Q$, it can be computed as the solution of a linear system, called the Sternheimer equation, which is possibly very ill-conditioned at finite temperature. Using extra information on additional bands that were discarded because of their small occupation numbers, we propose to improve the resolution of this linear system *via* a Schur complement method. This leads to very satisfying results where the number of Hamiltonian applications (which is the most costly step in the calculation for small to medium-sized systems) is reduced by 40%, even for numerically challenging systems. We also address in this chapter how to choose appropriately the number of extra bands, paving the way for future works.

Chapter 2

# Convergence analysis of direct minimization and self-consistent iterations

This chapter has been published in the article [GK1]:

*Eric Cancès, Gaspard Kemlin and Antoine Levitt. Convergence analysis of direct minimization and self-consistent iterations. SIAM Journal on Matrix Analysis and Applications, 42(1):243–274 (2021).* https://arxiv.org/abs/2004.09088.

**Abstract** This article is concerned with the numerical solution of subspace optimization problems, consisting of minimizing a smooth functional over the set of orthogonal projectors of fixed rank. Such problems are encountered in particular in electronic structure calculation (Hartree–Fock and Kohn–Sham density functional theory – DFT – models). We compare from a numerical analysis perspective two simple representatives, the damped self-consistent field (SCF) iterations and the gradient descent algorithm, of the two classes of methods competing in the field: SCF and direct minimization methods. We derive asymptotic rates of convergence for these algorithms and analyse their dependence on the spectral gap and other properties of the problem. Our theoretical results are complemented by numerical simulations on a variety of examples, from toy models with tunable parameters to realistic Kohn–Sham computations. We also provide an example of chaotic behaviour of the simple SCF iterations for a nonquadratic functional.

## Contents

**2.1 Introduction** . . . 34
**2.2 Optimization on Grassmann manifolds** . . . 36
2.2.1 First-order condition . . . 37
2.2.2 Second-order condition . . . 38
2.2.3 Fixed-point iterations on a manifold . . . 39
**2.3 Algorithms and analysis of convergence** . . . 39
2.3.1 Direct minimization . . . 39
2.3.2 Damped self-consistent field . . . 42
2.3.3 Comparison . . . 45
**2.4 Numerical tests** . . . 46
2.4.1 The retraction . . . 47
2.4.2 A toy model with tunable spectral gap . . . 47
2.4.3 Chaos in SCF iterations . . . 49
2.4.4 Local convergence for a 1D nonlinear Schrödinger equation . . . 50
2.4.5 Kohn–Sham density functional theory . . . 55
**2.5 Conclusion** . . . 57

## 2.1 Introduction

This chapter is concerned with the convergence behaviour of algorithms to solve the *subspace optimization problem*

$$\min\left\{E(P)\;\middle|\;P\in\mathbb{R}^{N_\mathrm{b}\times N_\mathrm{b}},\;P^2=P=P^*,\;\mathrm{Tr}(P)=N\right\}\tag{2.1.1}$$

consisting of optimizing a $C^2$ function $E:\mathbb{R}^{N_\mathrm{b}\times N_\mathrm{b}}\to\mathbb{R}$ over the set of rank-$N$ orthogonal projectors $P$. Here $P^*$ denotes the adjoint (transpose) of $P$. This problem can also be reformulated as

$$\min\left\{E\left(\sum_{i=1}^N\phi_i\phi_i^*\right)\;\middle|\;\phi_i\in\mathbb{R}^{N_\mathrm{b}},\;\phi_i^*\phi_j=\delta_{ij}\quad\forall\,i,j\in\{1,\dots,N\}\right\},\tag{2.1.2}$$

using an orthonormal basis $(\phi_i)_{i=1,\dots,N}$ for the subspace $\mathrm{Ran}\,(P)$. This problem is of interest in a number of contexts, such as matrix approximation, computer vision [1], and electronic structure theory [32, 95, 133, 134, 144, 179], the latter of which being the main motivation for this work.

Let $H(P)=\nabla E(P)$. The first-order conditions for problem (2.1.1) is

$$PH(P)(1-P)=(1-P)H(P)P=0.$$

Up to an appropriate choice for the orthonormal basis $(\phi_i)_{i=1,\dots,N}$ of $\mathrm{Ran}\,(P)$, this yields

$$H(P)\phi_i=\varepsilon_i\phi_i,\tag{2.1.3}$$

which reveals an alternative interpretation of this problem as a *nonlinear eigenvector problem* (to be distinguished from *nonlinear eigenvalue problems* of the form $A(\varepsilon)\phi=0$, where $A:\mathbb{R}\to\mathbb{R}^{N_\mathrm{b}\times N_\mathrm{b}}$). In the case when $E(P)=\mathrm{Tr}(H_0P)$ for a fixed symmetric matrix $H_0$, one recovers the classical eigenvalue problem $H_0\phi_i=\varepsilon_i\phi_i$. At a minimizer of (2.1.1), the $(\varepsilon_i)_{i=1,\dots,N}$ are the lowest eigenvalues of $H_0$, counting multiplicities.

Problems of the form (2.1.1) are found in the Hartree–Fock and Kohn–Sham theories of electronic structure [95, 144], both approximations of the many-body Schrödinger equation. In this context, the $\phi_i$ are (discretized) *orbitals*, the projector $P$ is the *density matrix*, and the energy $E(P)$ includes linear contributions from the kinetic and external potential energy of the electrons, as well as nonlinear terms arising from electron-electron interaction. Another notable problem of this form is the nonlinear Schrödinger or Gross–Pitaevskii equation for Bose–Einstein condensates [12], where $N=1$. In all these cases, the first-order condition (2.1.3) is interpreted as a *self-consistent* or *mean-field* equation: the particles behave as independent particles in an effective Hamiltonian $H(P)$ (also known as the Fock matrix) involving the mean-field they create. In the rest of this chapter, we will work on the formulation (2.1.1) without specifying $E$ for generality.

The minimization problem (2.1.1) is compact but nonconvex: there exists at least one minimizer, but the minimizer might not be unique, and local minima might not be global ones. Solving this optimization problem is of considerable practical interest, and algorithms for doing so date back to the early days of quantum mechanics [90]. The first introduced and still most popular approach is the *self-consistent field* (SCF) method, which, in its original version [146, 176], works as follows: if $P^k$ is the current iterate of the algorithm, $P^{k+1}$ is found by solving (2.1.3) for the fixed matrix $H(P^k)$:

$$H(P^k)\phi_i^k=\varepsilon_i^k\phi_i^k,\quad(\phi_i^k)^*\phi_j^k=\delta_{ij}$$

with the $\varepsilon_i^k$ sorted in nondecreasing order, and building $P^{k+1}$ as

$$P^{k+1}=\sum_{i=1}^N\phi_i^k(\phi_i^k)^*.$$

This algorithm assumes the *Aufbau* property, which is that at a minimum $P_*$ we have $P_*=\sum_{i=1}^N\phi_i\phi_i^*$ with $\phi_i$ a system of orthogonal eigenvectors associated with the lowest $N$ eigenvalues of $H(P_*)$. This property holds for the (spin-unconstrained) Hartree–Fock model [9] and the Gross–Pitaevskii models without magnetic field [30], usually holds for molecular systems in the Kohn–Sham model, but does not hold in general for Gross–Pitaevskii models with strong magnetic fields.

This basic procedure converges for systems where the nonlinearity is weak, but fails to converge otherwise (see [41] for a comprehensive mathematical analysis of this behaviour when the functional $E$ is a sum of a linear and a quadratic term in $P$, which is the case for the Hartree–Fock model). A solution is to *damp* this procedure, and *mix* the iterates to accelerate convergence. This gives rise to a variety of SCF algorithms, among which Broyden-like and Anderson-like mixing algorithms [108, 143, 169, 187], the Direct Inversion in the Iterative Space (DIIS) algorithm [118, 167, 168], the Optimal Damping Algorithm [40] (ODA), and the Energy-DIIS (EDIIS) algorithm combining the latter two approaches [119].

A second class of algorithms solves the minimization problem (2.1.1) directly. The minimization set $\{P \in \mathbb{R}^{N_\mathrm{b}\times N_\mathrm{b}},\ P^2 = P^* = P,\ \mathrm{Tr}P = N\}$ is diffeomorphic to the Grassmann manifold of the $N$-dimensional vector subspaces of $\mathbb{R}^{N_\mathrm{b}}$. This set is naturally equipped with the structure of a Riemannian manifold, and this allows the use of Riemann optimization algorithms [1, 72]. Direct minimization algorithms are preferred for the Gross–Pitaevskii model with magnetic fields [6, 60, 94, 97], for which the *Aufbau* principle is not satisfied in general. Gradient-type [3, 57, 151, 196, 207], Newton-type [11, 49, 209], and trust-region methods have also been designed to solve (2.1.1) for larger values of $N$. At the time of writing, direct minimization algorithms are less popular than SCF algorithms in electronic structure calculation, where $N$ can be very large, but it is not clear whether this is for sound scientific reasons or because SCF algorithms have been implemented and optimized for decades in the main production codes, which has not been the case for direct minimization algorithms.

While the convergence of several SCF and direct minimization algorithms has been analysed from a mathematical point of view (see *e.g.* [54, 124, 138, 174, 194, 206] and references therein), the two approaches have not been compared in a systematic way to our knowledge. The purpose of this chapter is to contribute to fill this gap, by focusing on very simple representatives of each class, namely the damped SCF iteration and the gradient descent. We emphasize that neither of these two algorithms is a practical choice as is. The SCF iteration should be accelerated (for instance using the Anderson acceleration technique), and the gradient information in direct minimization methods should rather be used as part of a quasi-Newton method (such as the limited-memory BFGS algorithm [1]). Depending on the exact problem at hand, all these methods should be preconditioned to avoid issues related to small mesh sizes (which leads to a divergence of the kinetic energy term) and/or large computational domains (which can lead to a divergence of the Coulomb energy, or the confining potential). We refer to [203] for a recent review in the context of the Kohn–Sham equations for solids. Rather, in this chapter, we aim to focus on the very simplest representative of each general strategy (SCF and direct minimization). The investigation of these two basic algorithms is informative on the strengths and weaknesses of the two classes, and is a first step in the analysis of more complex methods.

The chapter is organized as follows. In Section 2.2, we recall some results about optimization on Grassmann manifolds, in particular the first and second-order optimality conditions, and prove preparatory lemmas. In Section 2.3, we present the two algorithms that are in the scope of this chapter: a fixed-step gradient descent and a damped SCF algorithm. We prove their local convergence as long as the step is small enough and we derive convergence rates. We find that the convergence rates depend on the spectral radius of operators (acting on $\mathbb{R}^{N_\mathrm{b}\times N_\mathrm{b}}$) of the form $1-\beta J$, with $\beta$ the fixed step and $J = \boldsymbol{\Omega}_* + \boldsymbol{K}_*$ for the gradient descent, $J = 1 + \boldsymbol{\Omega}_*^{-1}\boldsymbol{K}_*$ for the SCF algorithm, where the operators $\boldsymbol{\Omega}_*$ and $\boldsymbol{K}_*$ are specified in the next section. Let us just mention at this stage that the lowest eigenvalue of $\boldsymbol{\Omega}_*$ is equal to the spectral gap between the $N^{\mathrm{th}}$ and $(N+1)^{\mathrm{st}}$ eigenvalues of $H(P_*)$, allowing us to analyse the convergence rates of the algorithms in terms of natural quantities of the problem. This also shows that the damped SCF algorithm can be seen as a matrix splitting of the fixed-step gradient descent algorithm.

In Section 2.4, we compare the two algorithms on several test problems. First, we focus on a toy model for which we can easily tune the gap and observe some fundamental differences between SCF and direct minimization algorithms, in agreement with the mathematical results established in Section 2.3. We also provide an example of chaos in SCF iterations, complementing the results of [41, 124] in the case of a nonquadratic objective functional $E$. Then, we analyse a 1D Gross–Pitaevskii model ($N = 1$) and its fermionic counterpart for $N = 2$, for which we investigate the behaviour of the algorithms when the gap closes. We conclude with an example from electronic structure calculation: a silicon crystal, in the framework of Kohn–Sham DFT, where we show in particular that accelerated SCF algorithms are less sensitive to small gaps than the simple damped SCF. Finally, in Section 2.5 we draw conclusions and outline perspectives for future work.

## 2.2 Optimization on Grassmann manifolds

We focus in this chapter on the case of real symmetric matrices, but the study can be easily extended to complex hermitian matrices. Let $\mathcal{H} \coloneqq \mathbb{R}^{N_\mathrm{b}\times N_\mathrm{b}}_{\mathrm{sym}}$ be the vector space of $N_\mathrm{b}\times N_\mathrm{b}$ real symmetric matrices endowed with the Frobenius inner product $\langle A, B\rangle_\mathrm{F} \coloneqq \mathrm{Tr}(AB)$. Let

$$\mathcal{M} \coloneqq \big\{P \in \mathcal{H} \;\big|\; P^2 = P\big\} \text{ and } \mathcal{M}_N \coloneqq \big\{P \in \mathcal{H} \;\big|\; P^2 = P,\ \mathrm{Tr}(P) = N\big\}.$$

From a geometrical point of view, $\mathcal{M}$ is a compact subset of $\mathcal{H}$ with $N_\mathrm{b}+1$ connected components $\mathcal{M}_N$, $N = 0, \dots, N_\mathrm{b}$, each of them being characterized by the value of $\mathrm{Tr}(P)$, namely the rank of the orthogonal projector $P$, and being diffeomorphic to the Grassmann manifold $\mathrm{Grass}(N, N_\mathrm{b})$ [1]. From now on, we fix the number of electrons $N$ and we seek the local minimizers of the problem

$$\min_{P\in\mathcal{M}_N} E(P), \tag{2.2.1}$$

where $E : \mathcal{H} \to \mathbb{R}$ is a discretized energy functional, for which some examples are given below.

**Example 2.1.** As an example, we study a discrete Gross–Pitaevskii model in Section 2.4.4. Other models from electronic structure can be considered, such as the discretized Hartree–Fock or Kohn–Sham models, where the energy is of the form

$$E(P) \coloneqq \mathrm{Tr}(H_0 P) + E_\mathrm{nl}(P)$$

with $H_0$ being the core Hamiltonian (representing the kinetic energy and the external potential) and $E_\mathrm{nl}$ a nonlinear energy functional depending on the model (representing the interaction between electrons). For instance, for the Hartree–Fock model,

$$E_\mathrm{nl}(P) \coloneqq \frac{1}{2}\mathrm{Tr}(G(P)P) \qquad \text{where} \qquad (G(P))_{ij} \coloneqq \sum_{k,l=1}^{N_\mathrm{b}} A_{ijkl}P_{kl} \quad \forall\, i,j = 1,\dots,N_\mathrm{b},$$

with $A$ a symmetric tensor of order 4. For more details on these models or electronic structure in general, we refer to [32, 134, 179].

In plane-wave, finite differences, finite elements or wavelets electronic structure calculation codes, the size $N_\mathrm{b}$ of the discretized space is in practice much larger than the number $N$ of electrons. Therefore, it is not practical to store and manipulate the (dense) matrix $P$. Instead, algorithms work on the orbitals $(\phi_i)_{i=1,\dots,N}$ introduced in (2.1.3). The density matrix $P$ is then recovered as

$$P = \sum_{i=1}^{N} \phi_i \phi_i^*.$$

All the results in this chapter are presented in the density matrix framework. However, the algorithms we study can be expressed in a way that avoids ever forming the density matrix. We refer to [203] for details.

We will need two assumptions for our results.

**Assumption 2.1.** The energy functional $E : \mathcal{H} \to \mathbb{R}$ is of class $C^2$ (twice continuously differentiable).

Assumption 2.1 is true for Hartree–Fock models. For Kohn–Sham models, it is true when the density $\rho = \sum_{i=1}^N |\phi_i|^2$ is uniformly bounded away from zero, which is the case for instance in condensed phase systems. Most of the results presented in this chapter are local in nature, and therefore this assumption can be relaxed to local regularity.

**Assumption 2.2.** $P_* \in \mathcal{M}_N$ is a nondegenerate local minimizer of (2.2.1) in the sense that there exists some $\eta > 0$ such that, for $P \in \mathcal{M}_N$ in a neighbourhood of $P_*$, we have

$$E(P) \geqslant E(P_*) + \eta\|P - P_*\|_\mathrm{F}^2.$$

It is very hard in most practical situations to check this assumption, but it seems to be verified in practice. Notable exceptions are systems invariant with respect to continuous symmetry groups, in which case $E(P) = E(P_*)$ for all $P$ in the orbit of $P_*$ along the symmetry group. In this case, the assumption cannot be true, and $\|P - P_*\|_{\mathrm{F}}$ must be replaced by the distance from $P$ to the orbit of $P_*$. Our results can be extended to this case up to quotienting $\mathcal{H}$ by the symmetry group.

Throughout the chapter, we will use the following notation:

- $H(P) \coloneqq \nabla E(P)$ is the gradient, and $H_* \coloneqq H(P_*)$;
- $\boldsymbol{K}(P) \coloneqq \boldsymbol{\Pi}_P \nabla^2 E(P) \boldsymbol{\Pi}_P$ is the Hessian projected onto the tangent space at $P$, and $\boldsymbol{K}_* \coloneqq \boldsymbol{K}(P_*)$ (the projection $\boldsymbol{\Pi}_P$ is defined below in Proposition 2.1).

### 2.2.1 First-order condition

To study the first-order optimality conditions, we start by recalling some classical results about the geometry of the manifold $\mathcal{M}_N$.

**Proposition 2.1.** *$\mathcal{M}_N$ is a smooth real manifold and its tangent space $\mathcal{T}_P\mathcal{M}_N$ at $P \in \mathcal{M}_N$ is given by*

$$\mathcal{T}_P\mathcal{M}_N = \{X \in \mathcal{H} \mid PX + XP = X,\ \mathrm{Tr}(X) = 0\} = \{X \in \mathcal{H} \mid PXP = (1-P)X(1-P) = 0\}.$$

*The orthogonal projection $\boldsymbol{\Pi}_P$ on $\mathcal{T}_P\mathcal{M}_N$ for the Frobenius inner product is*

$$\forall\ X \in \mathcal{H}, \quad \boldsymbol{\Pi}_P(X) = PX(1-P) + (1-P)XP = [P,[P,X]], \tag{2.2.2}$$

*where $[A,B] \coloneqq AB - BA$.*

*Proof.* The tangent space is given by the kernel of $\mathrm{d}g(P)$ where $g(P) = P^2 - P$ is the constraint which defines the manifold. A simple computation shows that

$$\forall\ X \in \mathcal{H}, \quad \mathrm{d}g(P)X = PX + XP - X, \tag{2.2.3}$$

which gives the first definition of the tangent space. The second one follows by multiplying by $P$ or $(1-P)$ on the right and the left in $\mathrm{d}g(P)X = 0$ and using that $P^2 = P \Leftrightarrow P(1-P) = 0$. The equalities (2.2.2) straightforwardly follow from the decomposition (2.2.4) below. □

Using the fact that $\mathcal{H} = \mathrm{Ran}\,(P) \oplus \mathrm{Ran}\,(1-P)$ and the induced decomposition of $P \in \mathcal{M}_N$ and $X \in \mathcal{H}$ as

$$P = \begin{bmatrix} I_N & 0 \\ 0 & 0 \end{bmatrix}, \quad X = \begin{bmatrix} (X)_{\mathrm{oo}} & (X)_{\mathrm{ov}} \\ (X)_{\mathrm{vo}} & (X)_{\mathrm{vv}} \end{bmatrix}, \tag{2.2.4}$$

the projection $\boldsymbol{\Pi}_P$ is given by

$$\boldsymbol{\Pi}_P(X) = \begin{bmatrix} 0 & (X)_{\mathrm{ov}} \\ (X)_{\mathrm{vo}} & 0 \end{bmatrix}.$$

Here the subscript "o" (resp. "v") stand for *occupied* (resp. *virtual*).

The first-order optimality condition at $P_*$ is $\boldsymbol{\Pi}_{P_*}(H_*) = 0$, which can be formulated as follows:

$$\boxed{\textbf{First-order optimality condition: } P_*H_*(1-P_*) = (1-P_*)H_*P_* = 0.} \tag{2.2.5}$$

Note that this condition can be rewritten as $[H_*, P_*] = 0$, showing that $H_*$ and $P_*$ can be codiagonalized. Let $(\phi_k)_{1 \leqslant k \leqslant N_{\mathrm{b}}}$ be an orthonormal basis of eigenvectors of $H_*$ associated with the eigenvalues $(\varepsilon_k)_{1 \leqslant k \leqslant N_{\mathrm{b}}}$ sorted in ascending order. Then $P = \sum_{i \in \mathcal{I}} \phi_i \phi_i^*$, where $\mathcal{I} \subset \{1, \dots, N_{\mathrm{b}}\}$, $|\mathcal{I}| = N$ is the set of occupied orbitals. The minimizer $P_*$ is said to satisfy

- the *Aufbau* principle if $\mathcal{I} = \{1, \dots, N\}$;
- the strong *Aufbau* principle if $\mathcal{I} = \{1, \dots, N\}$ and if in addition $\varepsilon_N < \varepsilon_{N+1}$, in which case $P_* = \sum_{i=1}^{N} \phi_i \phi_i^*$.

### 2.2.2 Second-order condition

We derive here the second-order optimality condition from the nondegeneracy of the minimum (Assumption 2.2).

Let $X \in \mathcal{T}_{P_*}\mathcal{M}_N$, $I$ be a real interval containing 0 and $\gamma : I \to \mathcal{M}_N$ be a smooth path such that $\gamma(0) = P_*$ and $\dot\gamma(0) = X$. An example of a possible $\gamma$ is given in Section 2.4.1. We expand

$$\begin{aligned} E(\gamma(t)) &= E(P_*) + t\langle H_*, X\rangle_{\mathrm{F}} + \frac{t^2}{2}\Big(\langle H_*, \ddot\gamma(0)\rangle_{\mathrm{F}} + \big\langle X, \nabla^2 E(P_*)X\big\rangle_{\mathrm{F}}\Big) + o(t^2) \\ &= E(P_*) + \frac{t^2}{2}\Big(\langle H_*, \ddot\gamma(0)\rangle_{\mathrm{F}} + \langle X, \boldsymbol{K}_* X\rangle_{\mathrm{F}}\Big) + o(t^2) \end{aligned}$$

as $H_*$ is orthogonal to $\mathcal{T}_{P_*}\mathcal{M}_N$ at the minimum. Differentiating the relation $\gamma(t)^2 = \gamma(t)$ at $t = 0$, we get

$$P_*\ddot\gamma(0) + \ddot\gamma(0)P_* + 2X^2 = \ddot\gamma(0),$$

from which we obtain the following two relations on the diagonal blocks of $\ddot\gamma(0)$ in the decomposition (2.2.4):

$$\frac{1}{2}(\ddot\gamma(0))_{\mathrm{oo}} = -\big(X^2\big)_{\mathrm{oo}} = -(X)_{\mathrm{ov}}(X)_{\mathrm{vo}}, \qquad \frac{1}{2}(\ddot\gamma(0))_{\mathrm{vv}} = \big(X^2\big)_{\mathrm{vv}} = (X)_{\mathrm{vo}}(X)_{\mathrm{ov}}.$$

Thus, since $(H_*)_{\mathrm{vo}} = (H_*)^*_{\mathrm{ov}} = 0$ at the minimum, we have

$$\begin{aligned} \langle H_*, \ddot\gamma(0)\rangle_{\mathrm{F}} &= \mathrm{Tr}\left(\begin{bmatrix} (H_*)_{\mathrm{oo}} & 0 \\ 0 & (H_*)_{\mathrm{vv}} \end{bmatrix}\begin{bmatrix} (\ddot\gamma(0))_{\mathrm{oo}} & (\ddot\gamma(0))_{\mathrm{ov}} \\ (\ddot\gamma(0))_{\mathrm{vo}} & (\ddot\gamma(0))_{\mathrm{vv}} \end{bmatrix}\right) \\ &= 2\mathrm{Tr}\Big( - (H_*)_{\mathrm{oo}}(X)_{\mathrm{ov}}(X)_{\mathrm{vo}}\Big) + 2\mathrm{Tr}\Big((H_*)_{\mathrm{vv}}(X)_{\mathrm{vo}}(X)_{\mathrm{ov}}\Big) \\ &= 2\mathrm{Tr}\Big( - (X)_{\mathrm{vo}}(H_*)_{\mathrm{oo}}(X)_{\mathrm{ov}}\Big) + 2\mathrm{Tr}\Big((X)_{\mathrm{ov}}(H_*)_{\mathrm{vv}}(X)_{\mathrm{vo}}\Big) \\ &= \mathrm{Tr}\Big(X(\boldsymbol{\Omega}_* X)\Big), \end{aligned}$$

where the operator $\boldsymbol{\Omega}_* : \mathcal{T}_{P_*}\mathcal{M}_N \to \mathcal{T}_{P_*}\mathcal{M}_N$ is defined as

$$\begin{aligned} \boldsymbol{\Omega}_* X &\coloneqq P_* X(1 - P_*)H_* - H_* P_* X(1 - P_*) + \mathrm{sym} \\ &= \begin{bmatrix} 0 & (X)_{\mathrm{ov}}(H_*)_{\mathrm{vv}} - (H_*)_{\mathrm{oo}}(X)_{\mathrm{ov}} \\ (H_*)_{\mathrm{vv}}(X)_{\mathrm{vo}} - (X)_{\mathrm{vo}}(H_*)_{\mathrm{oo}} & 0 \end{bmatrix}, \end{aligned} \tag{2.2.6}$$

where "sym" stands for the transpose of the previous expression. Introducing the operator

$$\boldsymbol{\Omega}_* + \boldsymbol{K}_* : \mathcal{T}_{P_*}\mathcal{M}_N \to \mathcal{T}_{P_*}\mathcal{M}_N, \tag{2.2.7}$$

one gets in the end

$$E(\gamma(t)) = E(P_*) + \frac{t^2}{2}\langle X, (\boldsymbol{\Omega}_* + \boldsymbol{K}_*)X\rangle_{\mathrm{F}} + o(t^2).$$

At the critical point $P_*$, the second-order expansion of $E(\gamma(t))$ only depends on $X = \dot\gamma(0)$, a common feature in constrained optimization. The operator $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ can be interpreted as the Hessian of the energy on the manifold, or alternatively as the partial Hessian of the Lagrangian on $\mathcal{H}$. The operator $\boldsymbol{\Omega}_*$ represents the influence of the curvature of the manifold on the Hessian of $E$.

As $P_*$ is a nondegenerate minimum in the sense of Assumption 2.2, we have the

$$\boxed{\textbf{Second-order optimality condition: } \forall\ X \in \mathcal{T}_{P_*}\mathcal{M}_N, \quad \langle X, (\boldsymbol{\Omega}_* + \boldsymbol{K}_*)X\rangle_{\mathrm{F}} \geqslant \eta\|X\|_{\mathrm{F}}^2.} \tag{2.2.8}$$

**Remark 2.1** (Structure of $\boldsymbol{\Omega}_*$ and link with the *Aufbau* principle)**.** Let $P_*$ be a nondegenerate minimizer of (2.2.1) in the sense of Assumption 2.2. Denoting by $A_{kl}$ the component along $\phi_k\phi_l^*$ of the matrix $A \in \mathcal{H}$, the operator $\boldsymbol{\Omega}_*$ defined in (2.2.6) can alternatively be defined by

$$\forall\ X \in \mathcal{T}_{P_*}\mathcal{M}_N, \quad (\boldsymbol{\Omega}_* X)_{ia} = (\varepsilon_a - \varepsilon_i)X_{ia} \text{ and } (\boldsymbol{\Omega}_* X)_{ai} = (\varepsilon_a - \varepsilon_i)X_{ai} \text{ for } i \in \mathcal{I},\ a \notin \mathcal{I},$$

where we have used the standard notation in chemistry of using the subscripts $i$ for occupied and $a$ for virtual orbitals ($\mathcal{I}$ is the set of occupied orbitals).

In the case when $E(D) = \mathrm{Tr}(HD)$ for some fixed symmetric matrix $H \in \mathcal{H}$ (linear eigenvalue problem), then $\boldsymbol{K}_* = 0$ and so (2.2.8) is equivalent to the *Aufbau* principle. This equivalence does not hold true in general for nonlinear models: (2.2.8) is independent of the Aufbau principle, and $\eta$ is unrelated to the gap $\nu = \min_{a \notin \mathcal{I}} \varepsilon_a - \max_{i \in \mathcal{I}} \varepsilon_i$ (equal to the lowest eigenvalue of the operator $\boldsymbol{\Omega}_*$). However, in specific cases, such as the reduced Hartree–Fock or Gross–Pitaevskii model, where $\boldsymbol{K}_* \geqslant 0$, we have $\eta \geqslant \nu$ and a positive gap is a sufficient (but not necessary) condition for optimality.

**Remark 2.2** (Link with the Liouvillian)**.** Another way to understand $\boldsymbol{\Omega}_*$ is to use the Liouvillian $\mathcal{L}_{H_*}$ associated to $H_*$, which is defined by:

$$\forall\, A \in \mathcal{H}, \quad \mathcal{L}_{H_*} A \coloneqq [H_*, A].$$

The action of $\mathcal{L}_{H_*}$ has a simple expression in the eigenvector decomposition $(\varepsilon_k, \phi_k)_{1 \leqslant k \leqslant N_\mathrm{b}}$ of $H_*$:

$$\forall\, 1 \leqslant k, l \leqslant N_\mathrm{b}, \quad \mathcal{L}_{H_*}(\phi_k \phi_l^*) = (\varepsilon_k - \varepsilon_l)\phi_k \phi_l^*. \tag{2.2.9}$$

Thus, we have

$$\forall\, i \in \mathcal{I},\ a \notin \mathcal{I}, \quad \boldsymbol{\Omega}_*(\phi_i \phi_a^* + \phi_a \phi_i^*) = (\varepsilon_a - \varepsilon_i)(\phi_i \phi_a^* + \phi_a \phi_i^*).$$

Hence, one can easily check that, using again the decomposition (2.2.4), we have

$$\forall\, X \in \mathcal{T}_{P_*}\mathcal{M}_N, \quad \boldsymbol{\Omega}_* X = -[P_*, \mathcal{L}_{H_*} X] = -[P_*, [H_*, X]]. \tag{2.2.10}$$

This definition also provides a canonical way to extend the operator $\boldsymbol{\Omega}_*$, originally defined on $\mathcal{T}_{P_*}\mathcal{M}_N$, to the whole space $\mathcal{H}$.

### 2.2.3 Fixed-point iterations on a manifold

Finally, we state a general abstract result that we will use to study the convergence of optimization algorithms on manifolds.

**Lemma 2.1.** *Let $\mathcal{M}$ be a smooth finite dimensional Riemannian manifold. Let $P_* \in \mathcal{M}$ and $f : U \to \mathcal{M}$ be a continuously differentiable mapping on a neighbourhood $U$ of $P_*$ such that $f(P_*) = P_*$. Let $\mathrm{d}f(P_*) : \mathcal{T}_{P_*}\mathcal{M} \to \mathcal{T}_{P_*}\mathcal{M}$ be the derivative of $f$ at $P_*$. If $\mathrm{d}f(P_*)$ verifies $r(\mathrm{d}f(P_*)) < 1$ where $r(\mathrm{d}f(P_*))$ is the spectral radius of $\mathrm{d}f(P_*)$, then, for $P^0$ close enough to $P_*$, the fixed-point iteration $P^{k+1} = f(P^k)$ linearly converges to $P_*$ with asymptotic rate $r(\mathrm{d}f(P_*))$, in the sense that for all $\theta > 0$ there exists $C_\theta > 0$ such that, for all $P^0$ close enough to $P_*$,*

$$\left\| P^k - P_* \right\|_\mathrm{F} \leqslant C_\theta (r(\mathrm{d}f(P_*)) + \theta)^k \left\| P^0 - P_* \right\|_\mathrm{F}.$$

*Proof.* We use the notation presented in [148, Chapter 1]. Up to a restriction of $U$ to a smaller neighbourhood of $P_*$, there exists a neighbourhood $V$ of 0 in $\mathcal{T}_{P_*}\mathcal{M}$ and $g : V \to U$ a diffeomorphic parametrization of the manifold such that $g(0) = P_*$ and $\mathrm{d}g(0) = \mathrm{Id}$ (take for instance the restriction to $V$ of the exponential map). Therefore, as $f$ is continuously differentiable, there exists a neighbourhood $\widetilde{V} \subset V$ of 0 in $\mathcal{T}_{P_*}\mathcal{M}$ such that $F \coloneqq g^{-1} \circ f \circ g : \widetilde{V} \to V$ is a continuously differentiable map with fixed-point 0 and $\mathrm{d}F(0) = \mathrm{d}f(P_*)$. As $r(\mathrm{d}F(0)) = r(\mathrm{d}f(P_*)) < 1$, we can find a neighbourhood $V' \subset V$ of 0 in $\mathcal{T}_{P_*}\mathcal{M}$ such that $F$ is a contraction in $V'$ for some norm $\|\cdot\|_\theta$, with contraction factor $r(\mathrm{d}f(P_*)) + \theta$, $\theta > 0$ (see [105] for more details). Therefore, we can apply the Banach fixed-point theorem to $F$ and we get that, for $x^0$ close enough to 0, $x^{k+1} = F(x^k)$ converges to 0. Finally, for $P^0 = g(x^0)$, $P^{k+1} = g(x^{k+1}) = g(F(x^k)) = f(g(x^k)) = f(P^k)$ converges to $P_* = g(0)$, with asymptotic rate $r(\mathrm{d}f(P_*))$. □

## 2.3 Algorithms and analysis of convergence

### 2.3.1 Direct minimization

The gradient descent algorithm consists in following the steepest descent direction with a fixed step $\beta$ at each iteration point. As the iterations are constrained to stay on the manifold, we have to

1. project the gradient on the tangent space with $\mathbf{\Pi}_{P^k}$ to bring the steepest descent line $P^k - \beta\mathbf{\Pi}_{P^k}\big(\nabla E(P^k)\big)$ back to the manifold at first-order;

2. retract the steepest descent line defined in the tangent space onto the manifold $\mathcal{M}_N$ by a nonlinear retraction $R$ mapping a neighbourhood of $\mathcal{M}_N$ in $\mathcal{H}$ to $\mathcal{M}_N$.

An example of retraction is given in Section 2.4.1 and we will assume that the retraction $R$ satisfies

**Assumption 2.3.** $R : \mathcal{H} \to \mathcal{H}$ is of class $C^2$ and for all $P \in \mathcal{M}_N$ and $X \in \mathcal{H}$ small enough,

$$R(P+X) \in \mathcal{M}_N \text{ and } R(P+X) = P + \mathbf{\Pi}_P(X) + O(X^2).$$

These two successive operations are sketched in Figure 2.1 and the gradient descent algorithm is presented in Algorithm 2.1.

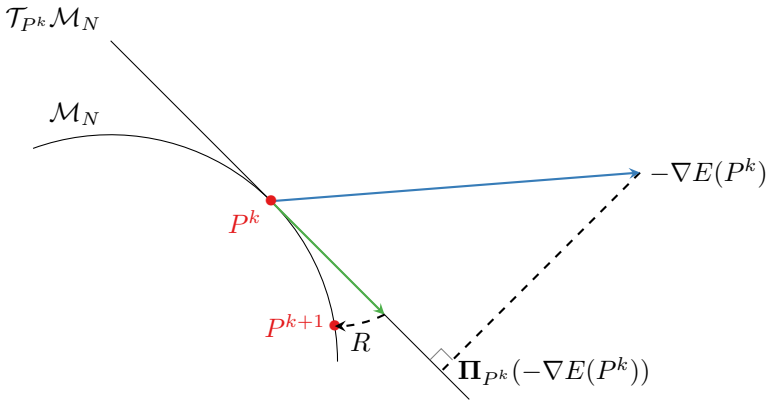



FIGURE 2.1 – Projection on the tangent space for the gradient descent, and retraction to the manifold.

ALGORITHM 2.1 – Gradient descent

**Data:** $P^0 \in \mathcal{M}_N$
**while** *convergence not reached* **do**
  $P^{k+1} := R\big(P^k - \beta\mathbf{\Pi}_{P^k}\big(\nabla E(P^k)\big)\big)$ ;
**end**

At the continuous level, this algorithm can be seen as the discretization of the flow $\dot{P} = -\mathbf{\Pi}_P \nabla E(P)$. Note that, by the use of the retraction $R$ and Assumption 2.3, the projection step has no influence on the convergence of the algorithm for $\beta$ small. Indeed, by Assumption 2.3,

$$\begin{aligned} \forall\, P \in \mathcal{M}_N, \quad R(P - \beta\mathbf{\Pi}_P(\nabla E(P))) &= P - \beta\mathbf{\Pi}_P(\mathbf{\Pi}_P(\nabla E(P))) + O(\beta^2) \\ &= P - \beta\mathbf{\Pi}_P(\nabla E(P)) + O(\beta^2) \end{aligned}$$

and thus the first-order expansion is the same with or without the projection step. The reason we use this projection step is that it is convenient to interpret $\mathbf{\Pi}_{P^k}\nabla E(P^k)$ as a residual.

The following theorems state that, for $\beta$ small enough, Algorithm 2.1 globally converges in the sense that $\mathbf{\Pi}_{P^k}\nabla E(P^k) \to 0$ and locally converges in the sense that $P^k \to P_*$ if $P^0$ is close enough to $P_*$.

**Theorem 2.1.** *Let $E : \mathcal{H} \to \mathbb{R}$ satisfy Assumption 2.1 and $R : \mathcal{H} \to \mathcal{H}$ satisfy Assumption 2.3. There exists $\beta_0 > 0$ such that for all $0 < \beta \leqslant \beta_0$ and all $P^0 \in \mathcal{M}_N$, the iterations*

$$P^{k+1} := R\big(P^k - \beta\mathbf{\Pi}_{P^k}\big(\nabla E(P^k)\big)\big)$$

*satisfy the following properties:*

1. *$(E(P^k))_{k\in\mathbb{N}}$ is a nonincreasing sequence converging to some critical value $E_c$ of $E$ on $\mathcal{M}_N$;*

2. *when $k$ goes to infinity, $\mathbf{\Pi}_{P^k}\nabla E(P^k) \to 0$, $\|P^{k+1}-P^k\|_{\mathrm{F}} \to 0$ and $d(P^k, A_{\mathrm{c}}) \to 0$ where $A_{\mathrm{c}}$ is one of the connected components of $C(E_{\mathrm{c}}) \coloneqq \{P \in \mathcal{M}_N \,|\, E(P) = E_{\mathrm{c}} \text{ and } \mathbf{\Pi}_P(\nabla E(P)) = 0\}$ .*

*Proof.* As $E : \mathcal{H} \to \mathbb{R}$ and $R : \mathcal{H} \to \mathcal{H}$ are $C^2$, and $\mathcal{M}_N$ is compact, we can use the expansion of Assumption 2.3 and obtain that there exists a constant $C \in \mathbb{R}_+$ such that for all $0 \leqslant \beta \leqslant 1$,

$$\forall\, k \in \mathbb{N}, \quad E(P^{k+1}) \leqslant E(P^k) - \beta\big\|\mathbf{\Pi}_{P^k}\nabla E(P^k)\big\|_{\mathrm{F}}^2 + C\beta^2\big\|\mathbf{\Pi}_{P^k}\nabla E(P^k)\big\|_{\mathrm{F}}^2$$

Therefore, we have for $\beta > 0$ small enough,

$$\forall\, k \in \mathbb{N}, \quad E(P^{k+1}) \leqslant E(P^k) - \frac{\beta}{2}\big\|\mathbf{\Pi}_{P^k}\nabla E(P^k)\big\|_{\mathrm{F}}^2.$$

This shows that the sequence $(E(P^k))_{k\in\mathbb{N}}$ is nonincreasing. As $E$ is continuous on the compact set $\mathcal{M}_N$, $(E(P^k))_{k\in\mathbb{N}}$ is bounded and hence converges to some $E_{\mathrm{c}} \in \mathbb{R}$. Moreover,

$$\sum_{k\in\mathbb{N}}\big\|\mathbf{\Pi}_{P^k}\nabla E(P^k)\big\|_{\mathrm{F}}^2 < \infty,$$

which implies that $\mathbf{\Pi}_{P^k}\nabla E(P^k) \to 0$ when $k \to \infty$. It follows that $\|P^{k+1}-P^k\|_{\mathrm{F}} \to 0$ when $k \to \infty$.

Let $B$ be the nonempty compact set of accumulation points of $(P^k)_{k\in\mathbb{N}}$. By continuity of $E$ and $P \mapsto \mathbf{\Pi}_P\nabla E(P)$, it follows that $B \subset C(E_{\mathrm{c}})$. Assuming that $d(P^k, B)$ does not go to zero, we can extract a subsequence at finite distance of $B$ which converges to a point in $B$, a contradiction. Assume that $B$ is disconnected: it is then the union of two compact subsets $B_1$ and $B_2$ at positive distance from each other. Since $P^{k+1}-P^k \to 0$, there is an infinite number of points in $(P^k)_{k\in\mathbb{N}}$ at distance greater or equal to $\eta > 0$ from both $B_1$ and $B_2$, from which we can extract a point in $B$, a contradiction. It follows that $B$ is connected, hence the result. □

This result implies in particular the convergence of the sequence $(P^k)_{k\in\mathbb{N}}$ in the generic case where critical points are isolated. If this is not the case but $E$ and $R$ are analytic, convergence can be shown following the approach in [124] based on Lojasiewicz inequality.

**Theorem 2.2.** *Let $E : \mathcal{H} \to \mathbb{R}$ satisfy Assumption 2.1 and Assumption 2.2 with $P_*$ a local minimizer of (2.2.1). Let $R : \mathcal{H} \to \mathcal{H}$ satisfy Assumption 2.3. Then, if $P^0 \in \mathcal{M}_N$ is close enough to $P_*$, the iterations*

$$P^{k+1} \coloneqq R\big(P^k - \beta\mathbf{\Pi}_{P^k}\big(\nabla E(P^k)\big)\big)$$

*linearly converge to $P_*$ for $\beta > 0$ small enough, with asymptotic rate $r(1-\beta J_{\mathrm{grad}})$ where $J_{\mathrm{grad}} \coloneqq \mathbf{\Omega}_* + \boldsymbol{K}_*$.*

*Proof.* In order to prove convergence, one can apply Lemma 2.1 to the function $f : \mathcal{M}_N \to \mathcal{M}_N$ defined by

$$f(P) \coloneqq R(P - \beta\mathbf{\Pi}_P(\nabla E(P))),$$

for which we know by the first-order optimality condition that $P_*$ is a fixed-point.

We compute explicitly $\mathrm{d}f(P_*)$ using the second-order optimality condition (2.2.8). To this end, take $X \in \mathcal{T}_{P_*}\mathcal{M}_N$ and a smooth path $\gamma : I \to \mathcal{M}_N$ defined on a real interval $I$ containing 0 such that $\gamma(0) = P_*$ and $\dot\gamma(0) = X$. We want to expand to the first-order in $t$ the following expression:

$$f(\gamma(t)) = R\big(\gamma(t) - \beta\mathbf{\Pi}_{\gamma(t)}(\nabla E(\gamma(t)))\big).$$

First, we focus on the projection of $H(\gamma(t))$ on $\mathcal{T}_{\gamma(t)}\mathcal{M}_N$:

$$\begin{aligned}
\mathbf{\Pi}_{\gamma(t)}H(\gamma(t)) &= \gamma(t)H(\gamma(t))(1-\gamma(t)) + \mathrm{sym} \\
&= (P_* + tX)\big(H_* + t\big(\nabla^2 E(P_*)X\big)\big)(1 - P_* - tX) + \mathrm{sym} + O(t^2) \\
&= t\big[P_*\big(\nabla^2 E(P_*)X\big)(1-P_*) + \mathrm{sym}\big] + t\big[XH_*(1-P_*) - P_*H_*X + \mathrm{sym}\big] + O(t^2) \\
&= t(\boldsymbol{K}_* + \mathbf{\Omega}_*)X + O(t^2).
\end{aligned}$$

Inserting this into the expansion of $f(\gamma(t))$, using Assumption 2.3 and the fact that $\mathbf{\Pi}_{\gamma(t)}X = \mathbf{\Pi}_{P_*}X + O(t^2)$, gives

$$f(\gamma(t)) = R\big(\gamma(t) - \beta t(\mathbf{\Omega}_* + \boldsymbol{K}_*)X + O(t^2)\big) = P_* + t\big(X - \beta(\mathbf{\Omega}_* + \boldsymbol{K}_*)X\big) + O(t^2).$$

Therefore,

$$\mathrm{d}f(P_*)X = \big(1 - \beta(\boldsymbol{\Omega}_* + \boldsymbol{K}_*)\big)X.$$

As the second-order optimality condition (2.2.8) shows that $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ is positive definite on $\mathcal{T}_{P_*}\mathcal{M}_N$, for $\beta$ small enough, the spectral radius $r(\mathrm{d}f(P_*))$ of the derivative $\mathrm{d}f(P_*)$, is less than 1, which concludes the proof. □

### 2.3.2 Damped self-consistent field

The damped SCF algorithm is a damped version of the Roothaan algorithm [41, 124] and is presented in Algorithm 2.2, under the assumption that the strong *Aufbau* principle is satisfied, and represented in Figure 2.2. Note that it is well-defined only if $\varepsilon_N^k < \varepsilon_{N+1}^k$ for all $k \in \mathbb{N}$. We introduce the nonlinear operators:

1. $A(H) \coloneqq \mathbf{1}_{(-\infty,\varepsilon_N(H)]}(H)$, with $\varepsilon_N(H)$ the lowest $N^{\text{th}}$ eigenvalue of $H$ and where we recall that $\mathbf{1}_{(-\infty,\mu]}(H) \coloneqq \sum_{\varepsilon_i \leqslant \mu} \phi_i \phi_i^*$, the $\phi_i$'s being orthonormal eigenvectors of $H$ associated to the eigenvalues $\varepsilon_i$;
2. $\Phi(P) = A(H(P))$ or, equivalently, $\Phi(P) \coloneqq \sum_{i=1}^{N} \phi_i \phi_i^*$ where the $\phi_i$'s are orthonormal eigenvectors associated to the lowest $N$ eigenvalues of $H(P)$.

**Algorithm 2.2** – Damped SCF algorithm

**Data:** $P^0 \in \mathcal{M}_N$
**while** *convergence not reached* **do**
  solve $\begin{cases} H(P^k)\phi_i^k = \varepsilon_i^k \phi_i^k, \ \varepsilon_1^k \leqslant \cdots \leqslant \varepsilon_N^k < \varepsilon_{N+1}^k \leqslant \cdots \leqslant \varepsilon_{N_\mathrm{b}}^k \\ (\phi_i^k)^* \phi_j^k = \delta_{ij}, \end{cases}$ ;
  $\Phi(P^k) \coloneqq \sum_{i=1}^{N} \phi_i^k \big(\phi_i^k\big)^*$;
  $P^{k+1} \coloneqq R\big(P^k + \beta \boldsymbol{\Pi}_{P^k}\big(\Phi(P^k) - P^k\big)\big)$;
**end**

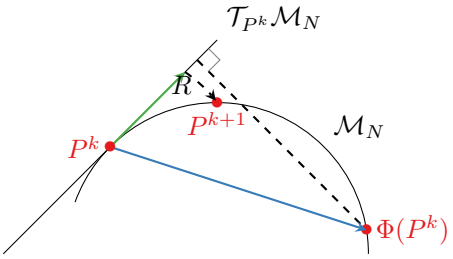



Figure 2.2 – Retraction for the damped SCF algorithm.

The following theorem states that, under the condition that there is a gap between the smallest $N^{\text{th}}$ and $(N+1)^{\text{st}}$ eigenvalues of the Hamiltonian $H_*$, Algorithm 2.2 locally converges for $\beta$ small enough.

**Theorem 2.3.** *Let $E : \mathcal{H} \to \mathbb{R}$ and $P_* \in \mathcal{M}_N$ satisfy Assumption 2.1 and Assumption 2.2 and $R : \mathcal{H} \to \mathcal{H}$ satisfy Assumption 2.3. Assume that $P_*$ satisfies the strong* Aufbau *principle*

$$\Phi(P_*) = P_* \text{ and } \nu \coloneqq \varepsilon_{N+1} - \varepsilon_N > 0,$$

*where $(\varepsilon_i)_{1 \leqslant i \leqslant N_\mathrm{b}}$ are the eigenvalues of $H_*$ ranked in nondecreasing order.*

*Then, for $\beta > 0$ small enough and $P^0 \in \mathcal{M}_N$ close enough to $P_*$, the iterations*

$$P^{k+1} \coloneqq R\big(P^k + \beta \boldsymbol{\Pi}_{P^k}\big(\Phi(P^k) - P^k\big)\big)$$

*are well-defined and $P^k$ linearly converges to $P_*$, with asymptotic rate $r(1-\beta J_{\mathrm{SCF}})$ where $J_{\mathrm{SCF}} \coloneqq 1+\boldsymbol{\Omega}_*^{-1}\boldsymbol{K}_*$.*

*Proof.* In order to prove convergence, we apply Lemma 2.1 to the function $f : \mathcal{M}_N \to \mathcal{M}_N$ defined by

$$f(P) \coloneqq R(P + \beta\boldsymbol{\Pi}_P(\Phi(P) - P)),$$

for which $P_*$ is a fixed-point.

First, we compute the derivative of $\Phi = A \circ H$ at the minimizer $P_*$ to get

$$\mathrm{d}\Phi(P_*) = \mathrm{d}A(H_*)\nabla^2 E(P_*).$$

Now, to compute $\mathrm{d}A(H_*)$, note that, as there is a gap $\varepsilon_{N+1} > \varepsilon_N$ at the minimum, we can find a contour $\mathcal{C}$ in the complex plane enclosing the lowest $N$ eigenvalues of $H_*$ (Figure 2.3) such that

$$A(H_*) = \frac{1}{2\pi\mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - H_*}\, \mathrm{d}z \tag{2.3.1}$$

(see [102, 111] for more details on spectral calculus, contour integrals and perturbation theory for functions of matrices). By continuity, we also have

$$A(H) = \frac{1}{2\pi\mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - H}\, \mathrm{d}z$$

for $H$ in a neighbourhood of $H_*$.

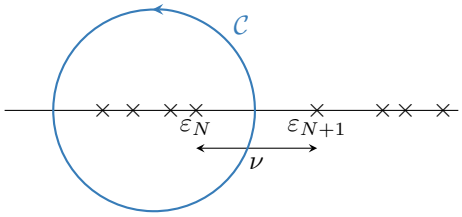



FIGURE 2.3 – Definition of $A$ and graphical interpretation of the *Aufbau* principle and the existence of a gap.

Then, one can use the expression (2.3.1) of $A$ and the expansion for $H$ in a neighbourhood of $H_*$

$$\forall\, z \in \mathcal{C}, \quad \frac{1}{z - H} = \frac{1}{z - H_*}(H - H^*)\frac{1}{z - H_*} + O(\|H - H_*\|_F^2)$$

to get

$$\begin{aligned}\forall\, h \in \mathcal{H}, \quad \mathrm{d}A(H_*)h &= \frac{1}{2\pi\mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - H_*} h \frac{1}{z - H_*}\, \mathrm{d}z \\ &= \sum_{k=1}^{N_\mathrm{b}} \sum_{l=1}^{N_\mathrm{b}} \left( \frac{1}{2\pi\mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - \varepsilon_k} h_{kl} \frac{1}{z - \varepsilon_l}\, \mathrm{d}z \right) \phi_k \phi_l^*,\end{aligned}$$

where $h_{kl} = \phi_k^* h \phi_l$. Now, let us denote by $1 \leqslant i \leqslant N$ the occupied orbitals ($\varepsilon_i$ is inside $\mathcal{C}$) and by $N+1 \leqslant a \leqslant N_\mathrm{b}$ the virtual ones ($\varepsilon_a$ is outside $\mathcal{C}$). Then,

$$\oint_{\mathcal{C}} \frac{1}{z - \varepsilon_i} \frac{1}{z - \varepsilon_a}\, \mathrm{d}z = \begin{cases} \dfrac{1}{\varepsilon_i - \varepsilon_a} & \text{if } 1 \leqslant i \leqslant N < a \leqslant N_\mathrm{b}; \\ 0 \quad \text{otherwise.} \end{cases}$$

Thus, the sum becomes

$$\mathrm{d}A(H_*)h = \sum_{i=1}^{N} \sum_{a=N+1}^{N_\mathrm{b}} \frac{1}{\varepsilon_i - \varepsilon_a} \left( h_{ia}\phi_i\phi_a^* + h_{ai}\phi_a\phi_i^* \right) = -\boldsymbol{\Omega}_*^{-1}\boldsymbol{\Pi}_{P_*} h,$$

and we finally get

$$\forall\, X \in \mathcal{T}_{P_*}\mathcal{M}_N, \quad \mathrm{d}\Phi(P_*)X = -\boldsymbol{\Omega}_*^{-1}\boldsymbol{K}_* X.$$

Now, we compute the derivative of $f$ at point $P_*$. Let $X \in \mathcal{T}_{P_*}\mathcal{M}_N$ and $\gamma : I \to \mathcal{M}_N$ a smooth path defined on a real interval $I$ containing 0 such that $\gamma(0) = P_*$ and $\dot\gamma(0) = X$. First, we expand $\gamma(t)$ around 0 and $\Phi$ around $\gamma(0) = P_*$ to obtain

$$f(\gamma(t)) = P_* + t\mathbf{\Pi}_{P_*}((1-\beta) + \beta \mathrm{d}\Phi(P_*))X + O\big(t^2\big).$$

Thus, for $X \in \mathcal{T}_{P_*}\mathcal{M}_N$,

$$\mathrm{d}f(P_*)X = \big((1-\beta) - \beta\mathbf{\Omega}_*^{-1}\boldsymbol{K}_*\big)X.$$

To conclude this proof, we compute the spectral radius of

$$\mathrm{d}f(P_*) = (1-\beta) - \beta\mathbf{\Omega}_*^{-1}\boldsymbol{K}_* = 1 - \beta\big(1 + \mathbf{\Omega}_*^{-1}\boldsymbol{K}_*\big).$$

First, notice that

$$1 + \mathbf{\Omega}_*^{-1}\boldsymbol{K}_* = 1 + \mathbf{\Omega}_*^{-1/2}\Big(\mathbf{\Omega}_*^{-1/2}\boldsymbol{K}_*\mathbf{\Omega}_*^{-1/2}\Big)\mathbf{\Omega}_*^{1/2}$$

and thus, $1 + \mathbf{\Omega}_*^{-1}\boldsymbol{K}_*$ and the symmetric operator $1 + \mathbf{\Omega}_*^{-1/2}\boldsymbol{K}_*\mathbf{\Omega}_*^{-1/2}$ have the same eigenvalues. Moreover, using the second-order optimality condition (2.2.8), with $X = \mathbf{\Omega}_*^{-1/2}Y$, we get

$$\begin{aligned} \forall\, Y \in \mathcal{T}_{P_*}\mathcal{M}_N, \quad \Big\langle Y, \Big(1 + \mathbf{\Omega}_*^{-1/2}\boldsymbol{K}_*\mathbf{\Omega}_*^{-1/2}\Big)Y\Big\rangle_{\mathrm{F}} &\geqslant \eta\langle Y, \mathbf{\Omega}_*^{-1}Y\rangle_{\mathrm{F}} \\ &\geqslant \frac{\eta}{\|\mathbf{\Omega}_*\|_{\mathrm{op}}}\|Y\|_{\mathrm{F}}^2, \end{aligned} \tag{2.3.2}$$

with $\|\mathbf{\Omega}_*\|_{\mathrm{op}}$ the operator norm associated to $\|\cdot\|_{\mathrm{F}}$. Thus, all the eigenvalues of $1+\mathbf{\Omega}_*^{-1/2}\boldsymbol{K}_*\mathbf{\Omega}_*^{-1/2}$, hence of $1 + \mathbf{\Omega}_*^{-1}\boldsymbol{K}$, are real and positive. Consequently, for $\beta$ small enough, the spectral radius $r(\mathrm{d}f(P_*))$ is less than 1 and we conclude by applying Lemma 2.1. □

**Remark 2.3** (Case when the *Aufbau* principle is not satisfied)**.** In the case when the minimizer $P_*$ does not verify the *Aufbau* principle, but does satisfy the condition that the eigenvalues of $1 + \mathbf{\Omega}_*^{-1}\boldsymbol{K}_*$ are positive (note that $\mathbf{\Omega}_*$ is not positive when the *Aufbau* principle is not verified, but $1 + \mathbf{\Omega}_*^{-1}\boldsymbol{K}_*$ might still have only positive eigenvalues), the damped SCF still converges locally to $P_*$ for $\beta > 0$ small enough if we change the way we select the occupied orbitals to build $\Phi(P)$ (in this case, we do not pick those associated to the smallest $N$ eigenvalues of $H(P)$, but those corresponding to the occupied orbitals of $P_*$).

We conclude this section by proving the local convergence of the nonretracted variant of Algorithm 2.2.

ALGORITHM 2.3 – Nonretracted damped SCF algorithm

**Data:** $P^0 \in \mathcal{M}_N$
**while** *convergence not reached* **do**
  solve $\begin{cases} H(P^k)\phi_i^k = \varepsilon_i^k\phi_i^k,\ \varepsilon_1^k \leqslant \cdots \leqslant \varepsilon_N^k < \varepsilon_{N+1}^k \leqslant \cdots \leqslant \varepsilon_{N_{\mathrm{b}}}^k \\ (\phi_i^k)^*(\phi_j^k) = \delta_{ij}, \end{cases}$ ;
  $\Phi(P^k) \coloneqq \displaystyle\sum_{i=1}^{N} \phi_i^k\big(\phi_i^k\big)^*$;
  $P^{k+1} \coloneqq P^k + \beta\mathbf{\Pi}_{P^k}\big(\Phi(P^k) - P^k\big)$;
**end**

**Theorem 2.4.** *Let $E : \mathcal{H} \to \mathbb{R}$ and $P_*$ satisfy Assumption 2.1 and Assumption 2.2. Moreover, assume that*

$$\Phi(P_*) = P_* \text{ and } \nu \coloneqq \varepsilon_{N+1} - \varepsilon_N > 0 \text{ (strong Aufbau principle)},$$

*where $(\varepsilon_i)_{1\leqslant i\leqslant N_{\mathrm{b}}}$ are the eigenvalues of $H_*$ ranked in nondecreasing order.*

*Then, for $\beta > 0$ small enough and $P^0 \in \mathcal{H}$ close enough to $P_*$ and with trace $N$, the iterations*

$$P^{k+1} \coloneqq P^k + \beta\big(\Phi(P^k) - P^k\big) \tag{2.3.3}$$

*are well-defined and $P^k$ linearly converges to $P_* \in \mathcal{M}_N$, with asymptotic rate $\max(r(1-\beta J_{\mathrm{SCF}}), 1-\beta)$ where $J_{\mathrm{SCF}} \coloneqq 1 + \mathbf{\Omega}_*^{-1}\boldsymbol{K}_*$.*

Note that the iterates $P^k$ defined by (2.3.3) have trace $N$ but do not lay on the manifold $\mathcal{M}_N$ in general.

*Proof.* The proof follows that of Theorem 2.3. This time, we need to compute the Jacobian matrix of $f : \mathcal{H} \ni P \mapsto P + \beta(\Phi(P) - P) \in \mathcal{H}$ at the minimizer $P_* \in \mathcal{M}_N$. As we work in the whole space $\mathcal{H}$, the Jacobian matrix has the form, in the decomposition $\mathcal{H} = \mathcal{T}_{P_*}\mathcal{M}_N \oplus (\mathcal{T}_{P_*}\mathcal{M}_N)^\perp$,

$$\mathrm{d}f(P_*) = \begin{bmatrix} 1 - \beta J_{\mathrm{SCF}} & \times \\ 0 & 1 - \beta \end{bmatrix},$$

where $J_{\mathrm{SCF}} = 1 + \boldsymbol{\Omega}_*^{-1}\boldsymbol{K}_*$ has been computed in the proof of Theorem 2.3. Hence, this time the algorithm converges to $P_* \in \mathcal{M}_N$ as long as $\beta$ is such that $\max(r(1 - \beta J_{\mathrm{SCF}}), 1 - \beta) < 1$. □

In LDA and GGA Kohn–Sham models [144], the mean-field Hamiltonian $H(P)$ is actually a function $\widetilde{H}(\rho_P)$ of the density $\rho_P$ associated with the density matrix $P$. Since the map $P \mapsto \rho_P$ is linear, (2.3.3) can be rewritten as

$$\rho^{k+1} = (1 - \beta)\rho^k + \beta\Psi(\rho^k),$$

where $\Psi(\rho) = \rho_{A(\widetilde{H}(\rho))}$. We can therefore interpret (2.3.3) as the equivalent density matrix formulation of this density mixing algorithm.

### 2.3.3 Comparison

In this section, we proved the local convergence of Algorithm 2.1 and Algorithm 2.2. and we obtained asymptotic convergence rates. On the tangent space, both Jacobian matrices are of the form $1 - \beta J$ where $J$ has positive real spectrum and

- for the gradient descent: $J_{\mathrm{grad}} = \boldsymbol{K}_* + \boldsymbol{\Omega}_*$, which is self-adjoint for the Frobenius inner product;
- for the damped SCF algorithm if the strong *Aufbau* principle is satisfied at $P_*$: $J_{\mathrm{SCF}} = 1 + \boldsymbol{\Omega}_*^{-1}\boldsymbol{K}_*$, which is self-adjoint for the inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{\Omega}_*} \coloneqq \langle \boldsymbol{\Omega}_* \cdot, \cdot \rangle_{\mathrm{F}}$.

One can notice that, *in the linear regime*, the SCF iterations correspond to a matrix splitting of the gradient iterations. Whether this results in a faster method or not depends not only on the relative conditioning of the iteration matrices but also on the relative cost of each step.

To have the fastest convergence, we want the eigenvalues of $1 - \beta J$ to be as close to 0 as possible. If we denote by $\lambda_1$ (resp. $\lambda_N$) the smallest (resp. largest) eigenvalue of $J$, the optimal step $\beta_*$ is the minimizer of $\min_\beta \max\{|1 - \beta\lambda_1|, |1 - \beta\lambda_N|\}$, which is given by

$$\beta_* = \frac{2}{\lambda_1 + \lambda_N}.$$

Then, the rate of convergence is, with $\kappa \coloneqq \lambda_N/\lambda_1$ the spectral condition number of $J$,

$$r = \frac{\kappa - 1}{\kappa + 1}.$$

Now, we can evaluate the conditioning of $J$ for the two algorithms:

- for the gradient descent, we have

$$\kappa(J_{\mathrm{grad}}) \leqslant \frac{\|\boldsymbol{\Omega}_*\|_{\mathrm{op}} + \|\boldsymbol{K}_*\|_{\mathrm{op}}}{\eta}, \tag{2.3.4}$$

  where $\eta$ is the coercivity constant in the nondegeneracy Assumption 2.2. First, the smaller $\eta$, the more difficult the convergence. Note however that there is no relationship in general between $\eta$ and the gap $\nu$. Second, the bigger $\|\boldsymbol{\Omega}_*\|_{\mathrm{op}} = \varepsilon_{N_{\mathrm{b}}} - \varepsilon_1$, the more difficult the convergence. In particular, for models arising from the discretization of partial differential equations, $\varepsilon_{N_{\mathrm{b}}} - \varepsilon_1 \to \infty$ when the discretization is refined. In practice, this issue is solved by preconditioning (see Remark 2.4).

- for the damped SCF algorithm, a naive bound would be

$$\kappa(J_{\mathrm{SCF}}) \leqslant \|\boldsymbol{\Omega}_*\|_{\mathrm{op}} \frac{1+\nu^{-1}\|\boldsymbol{K}_*\|_{\mathrm{op}}}{\eta}.$$

  In this bound the right-hand side diverges when $\|\boldsymbol{\Omega}_*\|_{\mathrm{op}} \to \infty$ as above, whereas the left-hand side may actually remain bounded. For instance, under the uniform coercivity assumption [31]

$$\forall\, X \in \mathcal{T}_{P_*}\mathcal{M}_N, \quad \langle X, (\boldsymbol{\Omega}_* + \boldsymbol{K}_*)X\rangle_{\mathrm{F}} \geqslant \widetilde{\eta}\langle \boldsymbol{\Omega}_* X, X\rangle_{\mathrm{F}},$$

  with $\widetilde{\eta}$ independent of $N_{\mathrm{b}}$ (which is often the case in practice), we have

$$\kappa(J_{\mathrm{SCF}}) \leqslant \frac{1+\nu^{-1}\|\boldsymbol{K}_*\|_{\mathrm{op}}}{\widetilde{\eta}}.$$

  In contrast with the bound (2.3.4), we can see that the smaller the gap $\nu$, the slower the convergence.

As a special case, if we consider the case where the Hessian $\nabla^2 E \equiv 0$, *i.e.* a linear eigenvalue problem. Then the SCF algorithm converges in one iteration, which is consistent with $J_{\mathrm{SCF}} = 1$. The gradient descent with optimal step locally converges with asymptotic rate $r = \frac{\kappa-1}{\kappa+1}$ where $\kappa = \frac{\varepsilon_{N_{\mathrm{b}}}-\varepsilon_1}{\varepsilon_{N+1}-\varepsilon_N}$.

The convergence rates we derived in Theorem 2.2 and Theorem 2.3 are consistent with well-known convergence issues, for instance the failure of the simple damped SCF algorithm to converge for systems with small gaps [177] (although Section 2.4.5 shows this is not necessarily true for more sophisticated acceleration methods).

**Remark 2.4** (Preconditioning)**.** We discuss here the extension of Theorem 2.2 to the preconditioned gradient descent:

$$P^{k+1} \coloneqq R\big(P^k - \beta \boldsymbol{\Pi}_{P^k} B \boldsymbol{\Pi}_{P^k}\big(\nabla E(P^k)\big)\big)$$

with $B : \mathcal{H} \to \mathcal{H}$ a symmetric positive definite preconditioner. If we denote by $\widetilde{B}_* \coloneqq \boldsymbol{\Pi}_{P_*} B \boldsymbol{\Pi}_{P_*}$ its restriction to the tangent plane, the Jacobian matrix of the gradient becomes $1 - \beta\widetilde{B}_*(\boldsymbol{\Omega}_* + \boldsymbol{K}_*)$ where $(\boldsymbol{\Omega}_* + \boldsymbol{K}_*)$ is positive definite (under Assumption 2.2) and the proof of local convergence for $\beta$ small enough follows exactly in the same way, using the positive definiteness of $\widetilde{B}_*$ to show that $\widetilde{B}_*(\boldsymbol{\Omega}_* + \boldsymbol{K}_*)$ has real positive spectrum. The same analysis holds true for the preconditioned SCF algorithm. In practice, preconditioning is a crucial tool to accelerate iterations, in particular in order to achieve mesh- and domain-size independence of the number of iterations for discretized partial differential equations. However, we are interested here in the intrinsic aspects of each algorithm (direct minimization *vs* SCF) and the influence of physical parameters (*e.g.* the gap $\nu$), so that the study of preconditioned algorithms is not in the scope of this chapter.

**Remark 2.5** (Dielectric operator)**.** In the context of Kohn–Sham density functional theory, the operator $\big(1 + \boldsymbol{K}_*\boldsymbol{\Omega}_*^{-1}\big)^{-1}$, the transpose of the inverse of the Jacobian of the simple SCF mapping, is known as the dielectric operator: it represents the infinitesimal change in the self-consistent Hamiltonian $H(P_*)$ in response to a change in the energy functional. Our results show that this operator is well-defined and has real positive spectrum, with no assumption on the sign of Hartree-exchange-correlation kernel $\boldsymbol{K}_*$, recovering in an algebraic framework the results of [63, 83] obtained using a different variational principle.

## 2.4 Numerical tests

We present here some numerical experiments to illustrate our theoretical results, explore their limits and investigate the global behaviour of the algorithms. First, we start by specifying the retraction $R$ that we use in our numerical tests. In Section 2.4.2, we use a simple toy model for which we can control the gap and analytically compute the exact minimizer: this allows us to study the impact of the gap on the convergence of Algorithm 2.1 and Algorithm 2.2. In Section 2.4.3, we show that simple (nondamped) SCF iterations can exhibit chaotic behaviour for some nonlinearities. Then, in Section 2.4.4, we report numerical tests for a 1D Gross–Pitaevskii model ($N = 1$) and its fermionic version for $N = 2$. Finally, in Section 2.4.5, we present results obtained with a more realistic case: silicon in the framework of the Kohn–Sham DFT.

### 2.4.1 The retraction

We choose the following algorithm: for a given symmetric matrix $\widetilde{P}$ close to $\mathcal{M}_N$ with eigendecomposition $\widetilde{P} = V\widetilde{D}V^*$ with $\widetilde{D}$ diagonal and $V$ orthogonal, we set the diagonal matrix $D$ as

$$D_{ii} = \begin{cases} 1 & \text{if } \widetilde{D}_{ii} > 0.5 \\ 0 & \text{otherwise} \end{cases}.$$

and $R(\widetilde{P}) = VDV^*$. When $\widetilde{P}$ is close to $\mathcal{M}_N$, its eigenvalues are close to either 0 or 1. Given a contour $\mathcal{C}$ enclosing only the eigenvalues close to 1 , $R$ has the following explicit expression

$$R(P) = \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - P} \, \mathrm{d}z. \tag{2.4.1}$$

Therefore, it follows from arguments similar to those used in the proof of Theorem 2.3 that $R$ is analytic and satisfies Assumption 2.3: if $X$ is small enough,

$$R(P + X) = R(P) + \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - P} X \frac{1}{z - P} \, \mathrm{d}z + o(X).$$

Therefore, the Jacobian matrix is

$$\begin{aligned} \mathrm{d}R(P)X &= \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - P} X \frac{1}{z - P} \, \mathrm{d}z \\ &= \sum_{k=1}^{N_\mathrm{b}} \sum_{l=1}^{N_\mathrm{b}} \left( \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - \lambda_k} (X)_{kl} \frac{1}{z - \lambda_l} \, \mathrm{d}z \right) \phi_k \phi_l^* \\ &= \sum_{k=1}^{N_\mathrm{b}} \sum_{l=1}^{N_\mathrm{b}} (X)_{kl} \left( \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - \lambda_k} \frac{1}{z - \lambda_l} \, \mathrm{d}z \right) \phi_k \phi_l^*, \end{aligned}$$

where $(A)_{kl}$ denotes the coefficient of the operator $A$ along the direction $\phi_k \phi_l^*$. By denoting by $i = 1, \ldots, N$ the occupied state ($\lambda_i = 1$) and $a > N_\mathrm{b}$ the virtual ones ($\lambda_a = 0$), we finally get, as the contour integral is 0 if $(k, l)$ are both occupied or both virtual, with sym being the hermitian conjugate,

$$\begin{aligned} \mathrm{d}R(P)X &= \sum_{i=1}^{N} \sum_{a=N+1}^{N_\mathrm{b}} (X)_{ia} \left( \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - \lambda_i} \frac{1}{z - \lambda_a} \, \mathrm{d}z \right) \phi_i \phi_a^* + \mathrm{sym} \\ &= \sum_{i=1}^{N} \sum_{a=N+1}^{N_\mathrm{b}} (X)_{ia} \phi_i \phi_a^* + \mathrm{sym}, \end{aligned}$$

as $\lambda_i = 1$ and $\lambda_a = 0$. Hence, if $X \in \mathcal{T}_P \mathcal{M}_N$, then

$$X = \sum_{i=1}^{N} \sum_{a=N+1}^{N_\mathrm{b}} (X)_{ia} \phi_i \phi_a^* + \mathrm{sym} = \mathrm{d}R(P)X,$$

so that we do have $\mathrm{d}R(P) = \mathbf{\Pi}_P$.

### 2.4.2 A toy model with tunable spectral gap

We work here in the very simple framework of real density matrices of order 2, *i.e.* the $2 \times 2$ real matrices $P$ such that $P^* = P$, $P^2 = P$ and $\mathrm{Tr}(P) = 1$. Then, we consider the following energy functional

$$E_\varepsilon(P) \coloneqq \mathrm{Tr}\left( \left( P - \begin{bmatrix} 1 & \varepsilon \\ \varepsilon & 0 \end{bmatrix} \right)^2 \right),$$

for parameters $\varepsilon \geqslant 0$. The gradient and Hessian of $E$ are

$$\begin{aligned} H_\varepsilon(P) &= 2\left( P - \begin{bmatrix} 1 & \varepsilon \\ \varepsilon & 0 \end{bmatrix} \right), \\ \nabla^2 E_\varepsilon(P) &= 2. \end{aligned}$$

Simple computations show that the set of rank-1 projectors on $\mathbb{R}^2$ can be parameterized as

$$\mathcal{M}_1 \coloneqq \left\{ P(a,b) = \begin{bmatrix} 1-a & b \\ b & a \end{bmatrix} \;\middle|\; a \in [0,1],\ b = \pm\sqrt{a(1-a)} \right\}.$$

The eigenvalues of $H_\varepsilon$ at $P(a,b) \in \mathcal{M}_1$ are $\pm 2\sqrt{a^2 + (b-\varepsilon)^2}$. The gap is thus $\nu(a,b) \coloneqq 4\sqrt{a^2 + (b-\varepsilon)^2}$.

**The case $\varepsilon = 0$**

Here, the unique minimum is clearly

$$P(0,0) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \in \mathcal{M}_1$$

and the gap is zero. Since $\nabla^2 E = 2$, this minimum satisfies Assumption 2.2 with $\eta = 2$.

**The case $\varepsilon > 0$**

We compute

$$E_\varepsilon(P(a,b)) = 2(a + \varepsilon^2 - 2\varepsilon b),$$

and therefore

$$E_\varepsilon\Big(P(a, \sqrt{a(1-a)})\Big) \leqslant E_\varepsilon\Big(P(a, -\sqrt{a(1-a)})\Big).$$

Hence, to compute the minimizer of the energy, we can restrict ourselves to the one-dimensional manifold

$$P(a) = \begin{bmatrix} 1-a & \sqrt{a-a^2} \\ \sqrt{a-a^2} & a \end{bmatrix}$$

with $a \in [0,1]$. Then, the energy is

$$E_\varepsilon(P(a)) = 2\Big(a + \varepsilon^2 - 2\varepsilon\sqrt{a(a-1)}\Big). \tag{2.4.2}$$

The first-order condition yields

$$a = \frac{1 \pm \sqrt{1 - \frac{4\varepsilon^2}{1+4\varepsilon^2}}}{2},$$

with the lowest energy achieved at

$$a(\varepsilon) \coloneqq \frac{1 - \sqrt{1 - \frac{4\varepsilon^2}{1+4\varepsilon^2}}}{2}.$$

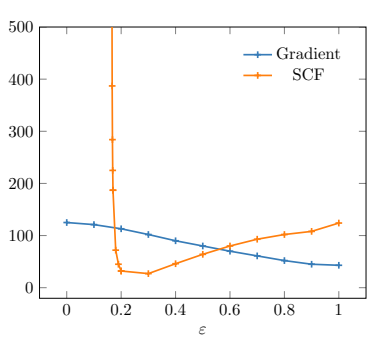
The gap $\nu(\varepsilon) \coloneqq 4\sqrt{a(\varepsilon)^2 + \left(\sqrt{a(\varepsilon)(1-a(\varepsilon))} - \varepsilon\right)^2}$ goes to 0 monotonically when $\varepsilon \to 0$. In particular, for $\varepsilon \approx 0$ we have $a(\varepsilon) \approx \varepsilon^2$ and $\nu(\varepsilon) \approx 4\varepsilon^2$. This model can thus be used to study the influence of the gap on the convergence of the two algorithms.

**Influence of $\varepsilon$ on the convergence**

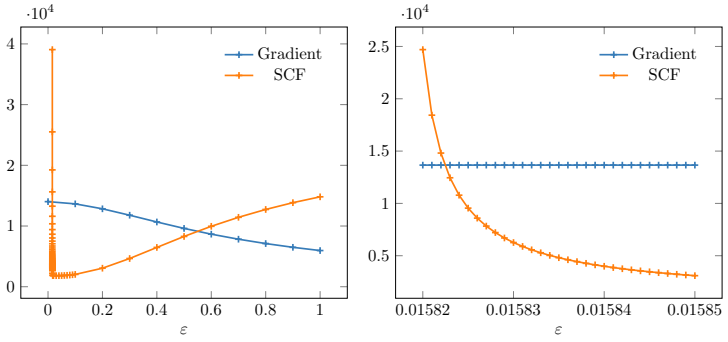
We run Algorithm 2.1 and Algorithm 2.2 with fixed $\beta$ on this system. We start from a random point on the manifold $\mathcal{M}$. We take as convergence criterion $\left\|P^k - P(a(\varepsilon))\right\|_{\mathrm{F}} \leqslant 10^{-12}$, and consider the algorithm has failed if convergence was not achieved after $50,000$ iterations.

On Figure 2.4, we plotted the number of iterations to achieve convergence for each algorithm as a function of $\varepsilon$ (without changing the starting point), for two different values of $\beta$: $10^{-1}$ and $10^{-3}$. The results confirm the theory we developed in Section 2.3.3: the gap has a strong influence on the convergence behaviour of the SCF algorithm. Indeed, as the gap decreases, smaller and smaller damping

parameters must be used, and the number of iterations increases. In fact for this system, $1 + \boldsymbol{\Omega}_*^{-1}\boldsymbol{K}_*$ has a single eigenvalue equal to $1 + \frac{2}{\nu(\varepsilon)} \approx 1 + \frac{1}{2\varepsilon^2}$ for $\varepsilon$ small. Thus, we expect convergence for $\beta < 4\varepsilon^2$, and therefore a critical $\varepsilon_{\mathrm{c}}$ of $\approx 0.158$ for $\beta = 10^{-1}$ and 0.0158 for $\beta = 10^{-3}$, with a number of iterations proportional to $\frac{1}{\varepsilon - \varepsilon_{\mathrm{c}}}$ when $\varepsilon > \varepsilon_{\mathrm{c}}$. The numerical results are in perfect agreement with this prediction. By contrast, the gradient algorithm is much less affected by the smallness of the gap, and converges in an essentially constant number of iterations: our prediction for the convergence rate of that method is $r = 1 - \beta(\nu(\varepsilon) + 2) \approx 1 - 2\beta$ for $\varepsilon$ small, and therefore a number of iterations for convergence to $10^{-12}$ of 124 for $\beta = 10^{-1}$ and $1.3 \times 10^4$ for $\beta = 10^{-3}$, again in perfect agreement with the numerical results.




(a) Number of iterations to reach convergence for both algorithms as a function of $\varepsilon$ for $\beta = 10^{-1}$.




(b) Number of iterations to reach convergence for both algorithms as a function of $\varepsilon$ for $\beta = 10^{-3}$. On the left is a global view of the convergence for $\varepsilon \in [0, 1]$ and on the right, we zoom in the neighbourhood of $\varepsilon_{\mathrm{c}}$.

FIGURE 2.4 – Comparison of the convergence of both algorithms depending on $\varepsilon$ for two different values of the step $\beta$.

### 2.4.3 Chaos in SCF iterations

We consider in this section another toy model a model inspired from the one proposed in [139, Section 2.1]. Let $h \in \mathbb{R}^{3\times3}_{\mathrm{sym}}$ and $J \in \mathbb{R}^{3\times3}$ be the matrices defined by

$$h \coloneqq \begin{bmatrix} 1.4299 & -0.2839 & -0.4056 \\ -0.2839 & 1.1874 & 0.2678 \\ -0.4056 & 0.2678 & 2.3826 \end{bmatrix} \quad \text{and} \quad J \coloneqq h^{-1}$$

and the energy functional $E_{c_1,c_2} : \mathbb{R}^{N_{\mathrm{b}}\times N_{\mathrm{b}}} \to \mathbb{R}$ defined by

$$E_{c_1,c_2}(P) = \mathrm{Tr}(hP) + \frac{c_1}{2}\rho_P^* J \rho_P - c_2 \sum_{j=1}^{3} \rho_{P,j}^{4/3}, \tag{2.4.3}$$

where $c_1, c_2 \in \mathbb{R}_+$ are nonnegative real parameter and where the density $\rho_P \in \mathbb{R}^3$ associated with the density matrix $P$ is given by $\rho_{P,j} = P_{jj}$ for $j = 1, 2, 3$. This model is reminiscent of a Kohn–Sham LDA model, with $h$ playing the role of the core Hamiltonian, $J$ of the Hartree operator and the third term in $E_{c_1,c_2}$ of the exchange-correlation energy. We seek the minimizers of $E_{c_1,c_2}$ on $\mathcal{M}_1$.

We study the behaviour of the simple SCF (Roothaan) iterations $P^k = \Phi(P^k)$ with initial guess $P^0 = \phi_0\phi_0^*$ with $\phi_0$ a random vector of norm 1.

#### The case $c_2 = 0$

Here, the energy functional is the sum of a linear and a quadratic term. In this case, either $(P^k)_{k\in\mathbb{N}}$ converges to a critical point of the problem (in practice a local minimizer), or it has two different accumulation points $P_{\mathrm{odd}}$ and $P_{\mathrm{even}}$, none of them being a critical point, and the iterates oscillates between the two, in the sense that $P^{2k+1} \to P_{\mathrm{odd}}$ and $P^{2k} \to P_{\mathrm{even}}$ when $k \to \infty$ [41, 124]. This is due to the fact

that we have

$$P^{2k+1} = \operatorname{argmin}\Big\{\widetilde{E}_{c_1,0}(P^{2k}, P),\ P \in \mathcal{M}_1\Big\},$$
$$P^{2k+2} = \operatorname{argmin}\Big\{\widetilde{E}_{c_1,0}(P, P^{2k+1}),\ P \in \mathcal{M}_1\Big\},$$

with

$$\widetilde{E}_{c_1,0}(P, P') \coloneqq \frac{1}{2}\mathrm{Tr}(hP) + \frac{1}{2}\mathrm{Tr}(hP') + \frac{c_1}{2}\rho_P^* J \rho'_P,$$

so that $(P^{2k}, P^{2k+1})$ converges to a minimizer of $\widetilde{E}_{c_1,0}$ on $\mathcal{M}_1 \times \mathcal{M}_1$. When $c_1$ is small, the simple SCF algorithm converges: for $c_1 = 0$, the matrix $h$ has a nondegenerate lowest eigenvalue and the algorithm converges in one iteration. When the value of $c_1$ increases, we observe numerically a bifurcation at a critical value $c_{1,*} \simeq 0.28$ after which the algorithm oscillates between two states.

**The case $c_2 = 1$**

Here the energy is not quadratic, and the previous theory does not apply. In Figure 2.5, we vary $c_1$ and plot the value of $\rho_1$ for the last 40 out of 1,500 SCF iterations. For this case, we still observe that the algorithm converges for $c_1$ small enough ($0 \leqslant c_1 < c_{1,*} \simeq 1.38$), and oscillates between two states for $c_1$ slightly larger than $c_{1,*}$. However, in contrast with the $c_2 = 0$ case, this is followed by a cascade of cycles of increasing periods, transitioning to a chaotic region, following the "period-doubling route to chaos" observed for other types of chaotic systems such as the logistic map [190].

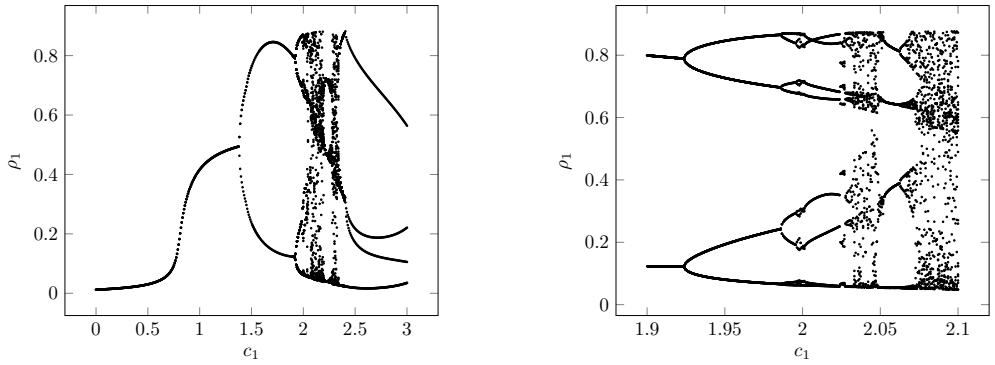



FIGURE 2.5 – Chaotic behaviour of the simple SCF map for the energy functional $E_{c_1,1}$ defined by (2.4.3) and $N = 1$ as a function of $c_1$. On the left is a global view of the bifurcation and the right is a zoom on part of the interval on which we observe a chaotic behaviour.

### 2.4.4 Local convergence for a 1D nonlinear Schrödinger equation

In this section, we present a simple 1D numerical experiment to validate on a more physically relevant system the sharpness of the convergence rates we derived in the previous section. As the goal here is to analyse the behaviour of the simplest representative of each class of algorithms when physical parameters (such as the gap) vary, we chose unpreconditioned algorithms. We consider a discretized 1D Gross–Pitaevskii model ($N = 1$) on the torus, and its (nonphysical) fermionic counterpart for $N = 2$. At the continuous level, the minimization set is

$$\big\{\gamma \in \mathcal{L}(L^2_{\mathrm{per}}),\ \gamma^2 = \gamma = \gamma^*,\ \mathrm{Tr}(\gamma) = N\big\},$$

where $\mathcal{L}(L^2_{\mathrm{per}})$ is the space of bounded operators on $L^2_{\mathrm{per}} \coloneqq \{u \in L^2_{\mathrm{loc}}(\mathbb{R}) \mid u(\cdot - 1) = u(\cdot)\}$, and the energy functional is defined as

$$\mathcal{E}_\alpha(\gamma) = \mathrm{Tr}_{L^2_{\mathrm{per}}}\left(-\frac{1}{2}\Delta\gamma\right) + \int_0^1 V\rho_\gamma + \frac{\alpha}{2}\int_0^1 \rho_\gamma^2,$$

where $\rho_\gamma$ is the density of the density matrix $\gamma$, $\alpha \in \mathbb{R}_+$ and $V$ is an asymmetric double-well external potential chosen equal to

$$V(x) \coloneqq -C\Big(\exp\Big(-c\cos(\pi(x-0.20))^2\Big) + 2\exp\Big(-c\cos(\pi(x+0.25))^2\Big)\Big), \tag{2.4.4}$$

with $c = 30$ and $C = 20$ (Figure 2.6).

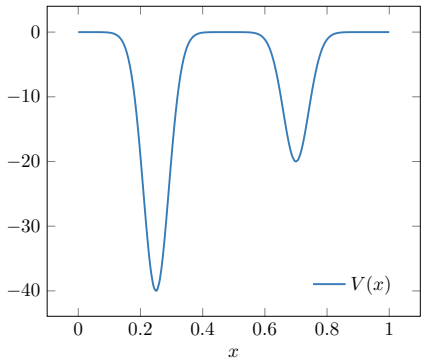



FIGURE 2.6 – $V$ for $c = 30$ and $C = 20$.

The Euler–Lagrange equations of this minimization problem are

$$\gamma_* = \sum_{i=1}^{N} (\phi_i, \cdot)_{L^2_{\mathrm{per}}} \phi_i, \quad \rho_* = \sum_{i=1}^{N} |\phi_i|^2, \quad -\frac{1}{2}\Delta\phi_i + V\phi_i + \alpha\rho_*\phi_i = \varepsilon_i\phi_i, \quad \int_0^1 \phi_i\phi_j = \delta_{ij}. \tag{2.4.5}$$

We discretize this model using the finite difference method with a uniform grid of step size $\delta = 1/N_{\mathrm{b}}$, which leads to the finite-dimensional model

$$\inf\{E_\alpha(P),\ P \in \mathcal{M}_N\}, \tag{2.4.6}$$

where

$$\forall\ P \in \mathcal{H}, \quad E_\alpha(P) \coloneqq \mathrm{Tr}(hP) + \frac{\alpha}{2}\delta\sum_{i=1}^{N_{\mathrm{b}}}\left(\frac{P_{ii}}{\delta}\right)^2, \tag{2.4.7}$$

the nonzero entries of the matrix $h \in \mathbb{R}^{N_{\mathrm{b}}\times N_{\mathrm{b}}}$ being given by

$$\forall\ 1 \leqslant i \leqslant N_{\mathrm{b}}, \quad h_{ii} = \frac{1}{\delta^2} + V(i\delta), \quad h_{i,i+1} = h_{i,i-1} = -\frac{1}{2\delta^2}.$$

where we identify the sites 0 and $N_{\mathrm{b}}$ on the one hand and $N_{\mathrm{b}}+1$ and 1 on the other. With this discretization, the discrete density is then given by $\rho(i\delta) \approx \rho_i \coloneqq P_{ii}/\delta$. We compare the fixed-step gradient descent and damped SCF algorithms (Algorithm 2.1 and Algorithm 2.2) on this problem for various values of $\alpha$, using as starting point the ground-state for $\alpha = 0$. The functional $E$ is smooth. To check Assumption 2.2 we notice that $E$ is a convex functional of $P$, so that $\nabla^2 E(P) \geqslant 0$. Therefore, at any local minimizer satisfying the strong *Aufbau* principle, $\boldsymbol{\Omega}_* + \boldsymbol{K}_* \geqslant \boldsymbol{\Omega}_* \geqslant \eta > 0$ and therefore Assumption 2.2 is satisfied. In the case where the *Aufbau* principle is not satisfied, Assumption 2.2 is not *a priori* always satisfied, so we check it *a posteriori* by computing the lowest eigenvalue of $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$.

We prove in the Appendix the following lemma, which collects some mathematical properties of the discretized model under consideration. The proof of this lemma, given in the appendix, strongly relies on the properties of our particular model (one-dimensional difference equation with periodic boundary conditions and a specific nonlinearity).

**Lemma 2.2** (Mathematical properties of (2.4.6))**.** *Let $\alpha \in \mathbb{R}_+$.*

1. *For $N = 1$, the optimization problem (2.4.6) has a unique minimizer $P_*$. In addition, $P_*$ can be written as $P_* = \phi_*\phi_*^*$, with $\phi_* \in \mathbb{R}^{N_{\mathrm{b}}}$, $\phi_*^*\phi_* = 1$, and $\phi_*$ positive component wise, and $P_*$ satisfies the strong* Aufbau *principle.*

2. *For* $2 \leqslant N \leqslant N_\mathrm{b}$*, with* $N_\mathrm{b} \neq 2N$ *if* $N_\mathrm{b} \in 4\mathbb{N}^*$*, the relaxed constrained optimization problem*

$$\inf\{E_\alpha(P),\ P \in \mathrm{CH}(\mathcal{M}_N)\}, \tag{2.4.8}$$

*where* $\mathrm{CH}(\mathcal{M}_N) = \{P \in \mathcal{H},\ P = P^*,\ 0 \leqslant P \leqslant 1, \mathrm{Tr}(P) = N\}$ *is the convex hull of* $\mathcal{M}_N$*, has a unique minimizer* $P_*$*. Either* $P_* \in \mathcal{M}_N$*, in which case* $P_*$ *is the unique minimizer of* (2.4.6) *and satisfies the* Aufbau *principle, or* $P_* \notin \mathcal{M}_N$*, in which case the eigenvalues* $\varepsilon_1 \leqslant \cdots \leqslant \varepsilon_{N_\mathrm{b}}$ *of the mean field Hamiltonian matrix* $H_* = \nabla E_\alpha(P_*)$ *satisfy* $\varepsilon_{N-1} < \varepsilon_N = \varepsilon_{N+1} < \varepsilon_{N+2}$ *and none of the local minimizers to* (2.4.6) *satisfies the* Aufbau *principle.*

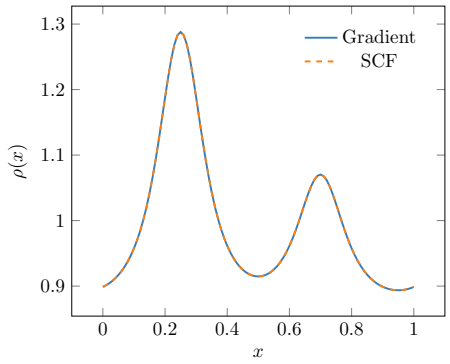
Note that the unique minimizer $P_*$ to the relaxed constraint problem (2.4.8) can be computed using the optimal damping algorithm (ODA) introduced in [40]. As shown in the proof of Lemma 2.2, the only case when the minimizer $P_*$ is not unique is the very particular case when $N_\mathrm{b} \in 4\mathbb{N}^*$, $N_\mathrm{b} = 2N$, and all the entries $[V_\mathrm{eff}]_i := V(i\delta) + \alpha\delta^{-1}[P_*]_{ii}$ of the effective potential are equal. In the rest of this section, we consider the cases $N = 1$ and $N = 2$.

**The case** $N = 1$

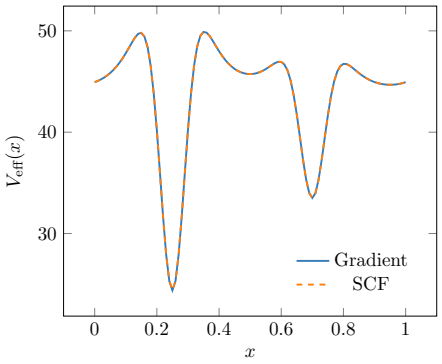
It follows from Lemma 2.2, Theorem 2.2 and Theorem 2.3 that Algorithm 2.1 and Algorithm 2.2 locally converge to the unique minimizer $P_*$ as long as $\beta$ is chosen small enough. The resulting densities, effective potentials and convergence behaviour of both algorithms are plotted in Figure 2.7 for $N_\mathrm{b} = 100$. The SCF algorithm converges faster in terms of number of iterations, as a smaller $\beta$, hence more steps, are required for the gradient to converge. This is expected from the large spectral radius of the matrix $h$ in the absence of preconditioning.

For the gradient algorithm, the convergence rate is consistent with the spectral radius of the Jacobian matrix $1 - \beta J_\mathrm{grad}$. For the damped SCF algorithm with the ground-state of the core Hamiltonian as starting point, surprisingly, we observe an asymptotic convergence rate slightly faster than that expected from the spectral radius of the Jacobian matrix $1 - \beta J_\mathrm{SCF}$. Using a random perturbation of the ground-state of the core Hamiltonian as starting point again gives a convergence rate consistent with the spectral radius.

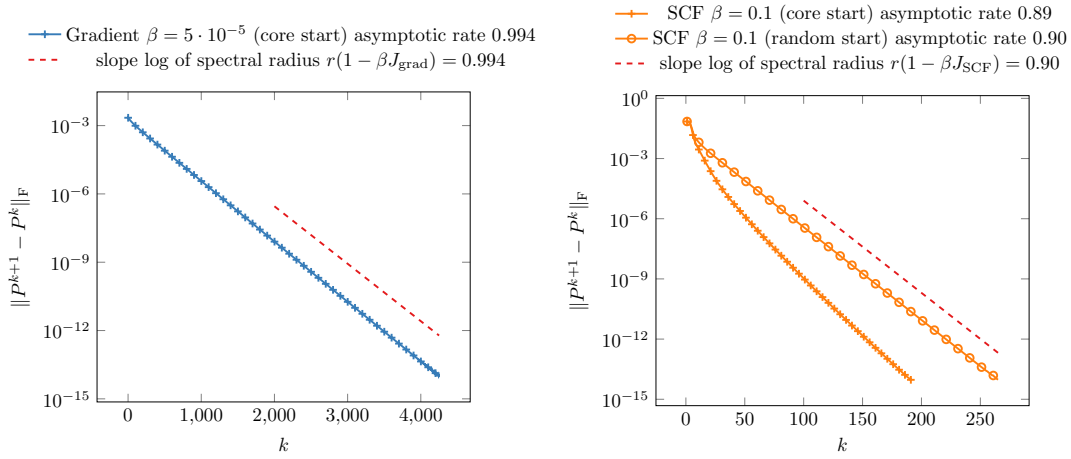
The explanation of this effect is to be found in the repartition of the error among the eigenvectors of $J_\mathrm{SCF}$. Since both $\boldsymbol{\Omega}_*$ and $\boldsymbol{K}_*$ are positive semidefinite operators, the eigenvalues of $J_\mathrm{SCF} = 1 + \boldsymbol{\Omega}_*^{-1}\boldsymbol{K}_*$ are greater than 1, and the convergence for $\beta$ small is limited by the modes associated with eigenvalues of $J_\mathrm{SCF}$ close to 1. These eigenvalues correspond to high eigenvalues of $\boldsymbol{\Omega}_*$, and therefore to highly oscillatory modes. When the initial guess is chosen as the ground-state of the core Hamiltonian, these modes are only weakly excited and do not contribute to the observed convergence rate before convergence is achieved. When the initial guess is randomly perturbed, this effect is not present and the convergence rate is consistent with the spectral radius. For the gradient algorithm, the rate-limiting modes are associated with small eigenvalues of $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$, which are not oscillatory, and this effect is not present either.




(a) Density $\rho$ at convergence.




(b) Effective potential $V_{\text{eff}} = V + \alpha\rho$ at convergence.

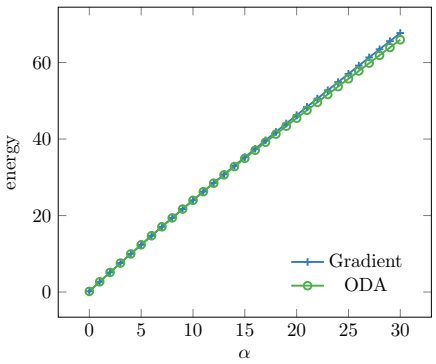



(c) Error decay for SCF and gradient descent algorithms (every few mark only is plotted for the sake of visibility). Marked lines are the evolution of the error $\left\|P^{k+1} - P^k\right\|_{\text{F}}$ and dashed lines represents the slope computed with the spectral radius of the Jacobian matrix, computed by finite differences.

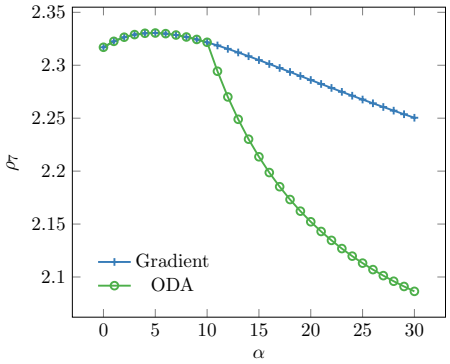
FIGURE 2.7 – Convergence of Algorithm 2.1 and Algorithm 2.2 for $N = 1$, $\alpha = 50$ and $N_{\text{b}} = 100$.

**The case $N = 2$**

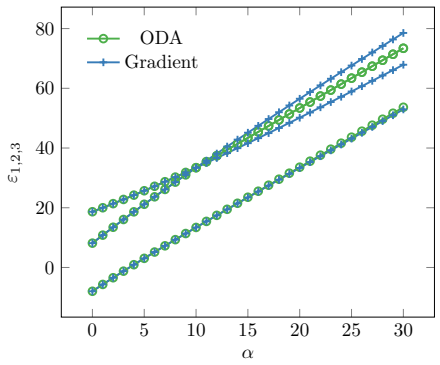
Since the second-smallest eigenvalue of the matrix $h$ is strictly lower than the third one, for $\alpha$ small enough, the unique minimizer $P_*$ of (2.4.8) is on $\mathcal{M}_2$ and satisfies the strong *Aufbau* principle, and both the gradient descent and SCF algorithm locally converge to $P_*$. For larger values of $\alpha$, the two alternatives of Lemma 2.2 appear. We plot the energy, the density at an arbitrarily chosen point and the eigenvalues of the solutions $P^{\mathrm{grad}}$ and $P^{\mathrm{ODA}}$ obtained by the gradient and the ODA algorithm as a function of $\alpha$ in Figure 2.8, evidencing a bifurcation for a critical value of $\alpha_{\mathrm{c}} \simeq 10$, after which these two solutions are different.



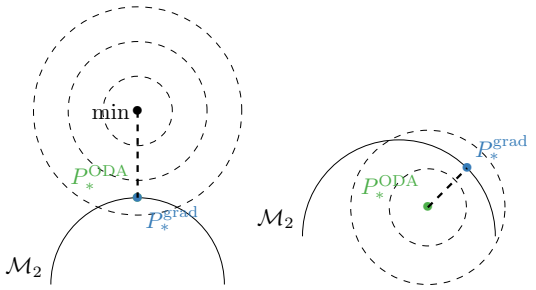

(a) Bifurcation on the energy.




(b) Bifurcation on the value of $\rho_7$ ($7^{\mathrm{th}}$ element of the discrete density for $N_{\mathrm{b}} = 40$).




(c) Bifurcation on the first three eigenvalues of the Hamiltonians $H(P_*^{\mathrm{ODA}})$ and $H(P_*^{\mathrm{grad}})$. For ODA, after the collision, the second and third eigenvalues are degenerate.




(d) ODA minimum before (left) and after (right) the bifurcation. Dashed lines are the energy level sets around the minimum.

FIGURE 2.8 – Bifurcation on the energy, the density and the eigenvalues as a function of $\alpha$ for $N = 2$, $N_{\mathrm{b}} = 40$.

For $\alpha$ lower than the critical value $\alpha_{\mathrm{c}} \simeq 10$, $P_*^{\mathrm{ODA}}$ is on the manifold $\mathcal{M}_2$ and satisfies the strong *Aufbau* principle. The algorithms all converge to this solution: $P_*^{\mathrm{grad}} = P_*^{\mathrm{SCF}} = P_*^{\mathrm{ODA}} = P_*$. However for $\alpha > \alpha_{\mathrm{c}}$ in the range tested, $P_*^{\mathrm{ODA}} \notin \mathcal{M}_2$. A geometrical interpretation of the bifurcation is sketched on Figure 2.8(d): the level sets of the function $E_\alpha$ are degenerate ellipsoids. Below $\alpha_{\mathrm{c}}$, the intersection of the nonempty closed convex set $\mathrm{CH}(\mathcal{M}_2)$ with the level set of $E_\alpha$ of lowest energy belongs to $\mathcal{M}_2$, while this is no longer the case beyond $\alpha_{\mathrm{c}}$.

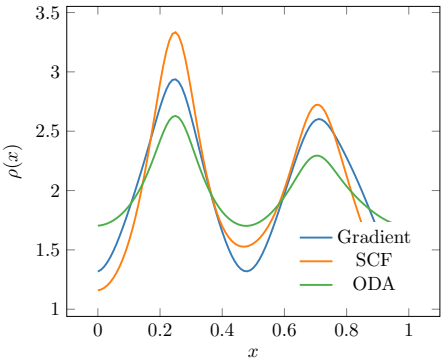
For $\alpha > \alpha_{\mathrm{c}}$, the solutions obtained by the ODA, gradient and SCF algorithm differ, as shown in Figure 2.9:

- the lowest second and third eigenvalues of $H(P_*^{\mathrm{ODA}})$ are degenerate ($\varepsilon_2 = \varepsilon_3$) and $P_*^{\mathrm{ODA}} \notin \mathcal{M}_2$. More precisely, we have

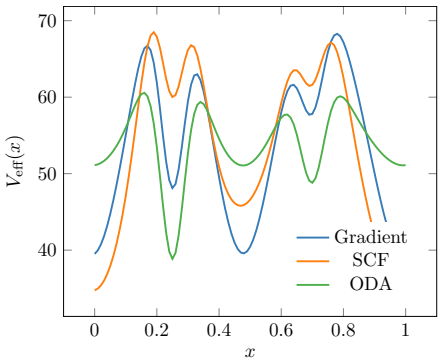
$$P_*^{\mathrm{ODA}} = \phi_1\phi_1^* + (1-f)\phi_2\phi_2^* + f\phi_3\phi_3^* \quad \text{with} \quad H(P_*^{\mathrm{ODA}})\phi_i = \varepsilon_i\phi_i, \quad \phi_i^*\phi_j = \delta_{ij}$$

  with a fractional occupation $0 < f < 1$;

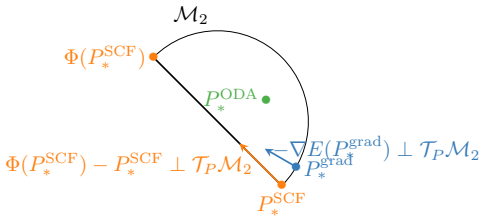
- Algorithm 2.1 and Algorithm 2.2 converge to two different limits $P_*^{\text{grad}}$ and $P_*^{\text{SCF}}$, none of them satisfying the *Aufbau* principle:
  - $P_*^{\text{grad}}$ is a local minimizer of $E_\alpha$ on $\mathcal{M}_2$, which does not satisfy the *Aufbau* principle. More precisely, $P_*^{\text{grad}}$ is the orthogonal projector on the space generated by the eigenvectors associated with the lowest first and third eigenvalues of $H(P_*^{\text{grad}})$;
  - $P_*^{\text{SCF}}$ satisfies $\mathbf{\Pi}_{P_*^{\text{SCF}}}\left(\Phi(P_*^{\text{SCF}}) - P_*^{\text{SCF}}\right) = 0$, but $\Phi(P_*^{\text{SCF}}) - P_*^{\text{SCF}} \neq 0$ and $[H(P_*^{\text{SCF}}), P_*^{\text{SCF}}] \neq 0$. The iterates are trapped as the search direction is orthogonal to the tangent space (Figure 2.9(c)). The limit point $P_*^{\text{SCF}}$ is a spurious stationary state of the SCF iteration which is not physically relevant, not being a critical point of $E_\alpha$.




(a) Densities of $P_*^{\text{grad}}$, $P_*^{\text{SCF}}$ and $P_*^{\text{ODA}}$.




(b) Effective potential $V_{\text{eff}} = V + \alpha\rho_P$ for $P = P_*^{\text{grad}}$, $P_*^{\text{SCF}}$ and $P_*^{\text{ODA}}$.




(c) Geometrical interpretation of the limiting points of the gradient descent, SCF et ODA algorithms.

| | Gradient | SCF | ODA |
|---|---|---|---|
| $\varepsilon_1$ | 52.9 | 51.3 | 53.6 |
| $\varepsilon_2$ | 67.9 | 67.8 | 73.4 |
| $\varepsilon_3$ | 78.5 | 79.6 | 73.4 |

(d) Lowest three eigenvalues of $H(P)$ for $P = P_*^{\text{grad}}$, $P = P_*^{\text{SCF}}$ and $P = P_*^{\text{ODA}}$.

FIGURE 2.9 – Results obtained with the gradient descent, damped SCF and ODA algorithm for $N = 2$, $\alpha = 30$ and $N_{\text{b}} = 100$: the limiting points $P_*^{\text{grad}}$ and $P_*^{\text{SCF}}$ lay by construction on the manifold $\mathcal{M}_2$, while $P_*^{\text{ODA}}$ does not (it only belongs to its convex hull $\text{CH}(\mathcal{M}_2)$).

### 2.4.5 Kohn–Sham density functional theory

We now investigate a more realistic computation: the electronic structure of a silicon crystal using Kohn–Sham density functional theory (KS-DFT). We used the `DFTK.jl` code [101], which solves the equations of KS-DFT in a plane-wave basis under a pseudopotential approximation. All computations below use the local density approximation (LDA) of the exchange-correlation energy [116, 144], Goedecker–Teter–Hutter (GTH) pseudopotentials [79], a cut-off energy of 30 hartree, and Γ-only Brillouin zone sampling for simplicity, although the same behaviour was observed with different exchange-correlation functionals and fine Brillouin zone discretizations. In all cases, the initial guess for the algorithms is a superposition of atom-centered densities. The DFTK code as well as the script used to produce these results are available at `https://dftk.org/`.

We consider the case of silicon in its standard face-centered cubic phase. With the chosen pseudopotentials and without spin polarization, silicon has four occupied orbitals: $N = 4$. We examine the convergence of algorithms as a function of the lattice constant $a$ (the size of the computational domain). In the equilibrium state of silicon ($a \approx 10.26$ bohrs), there is a gap of about 0.08 hartree between the

occupied and virtual states. As the lattice constant $a$ is increased, this gap decreases, until it closes at $a \approx 11.408$ bohrs. We examine the convergence of self-consistent iterations, using both fixed-step damped density mixing with four values of the mixing parameter $\beta$ ($\beta = 1$ – no damping –, $\beta = 0.5$, $\beta = 0.2$, $\beta = 0.1$), as well as the self-consistent iteration accelerated with Anderson acceleration (also known as Pulay's DIIS method [54, 167, 168]). We plot the convergence of the density residual $\|\rho_{\Phi(P_n)} - \rho_{P_n}\|_2$ as a function of the iterations for three values of $a$, with decreasing gaps.

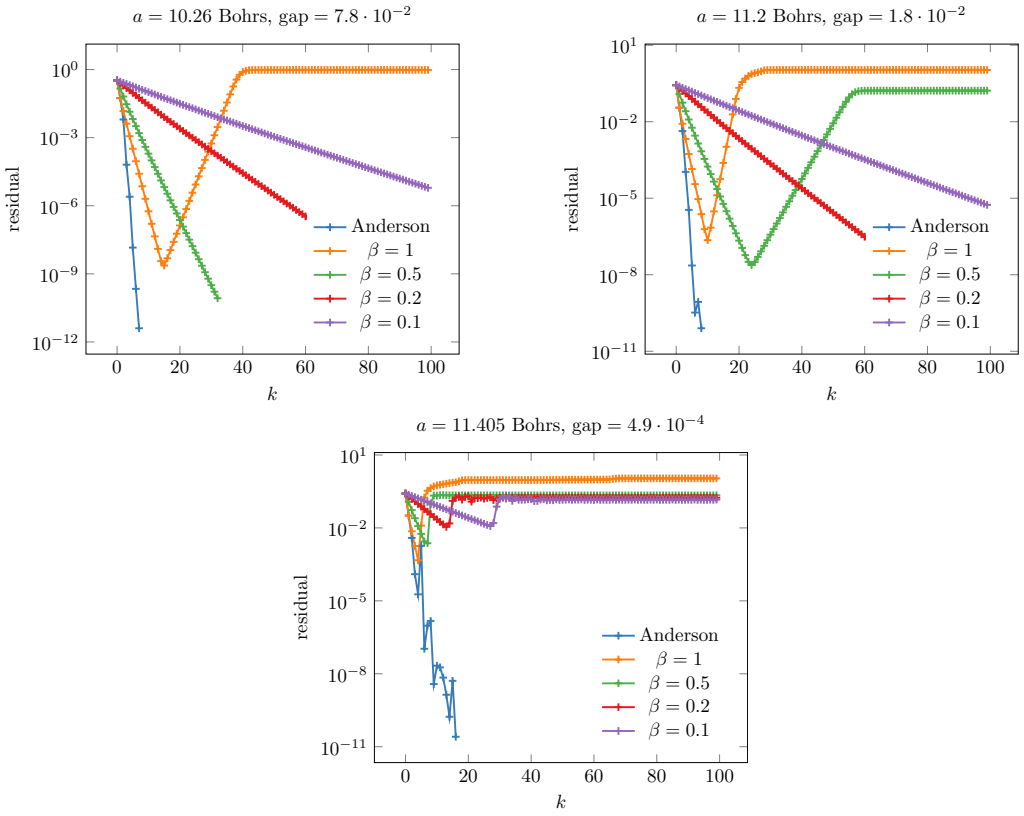



FIGURE 2.10 – Convergence curves of the density residual as a function of the number of iterations $k$ for silicon with different lattice constants $a$.

In the first case, with $a = 10.26$ bohrs, the simple (undamped) SCF method appears to be converging for almost 20 iterations, but then diverges, until the density residual stabilizes at a positive value, as predicted in [41] for the Hartree–Fock model. The damped methods appear to converge. For $a = 11.2$ bohrs, the damping method with $\beta = 0.5$ does not converge. When the lattice constant is further increased to 11.405 bohrs, with a small gap of $4.9 \times 10^{-4}$ hartree, the fixed-step damped SCF iterations do not converge for the tested values of the damping parameter ($\beta = 1, 0.5, 0.2, 0.1$).

When it occurs, the transient behaviour of apparent-convergence before eventual divergence is unusually long. For instance for $\beta = 1$ at $a = 10.26$ bohrs, the method appears to be converging for almost 20 iterations, up to a reduction in residual of a factor $10^{-8}$. In fact, it is consistent with an initial error of the order of machine precision (about $10^{-16}$ here) being amplified at a constant rate. The cause of this effect appears to be that the divergent modes break the natural inversion symmetry of the crystal in this particular case: we have checked that the divergence occurs much sooner if we break this symmetry by perturbing the positions of the atoms around their symmetric positions (at 9 iterations by perturbing the position of one atom by 10%). In practice, in the symmetric case, one way to overcome this issue is to ensure during the algorithm that, at each step, we have a symmetric solution. Note that this phenomenon is reminiscent of that observed in Figure 2.7, where all the modes were not fully excited, making the convergence faster than expected.

It is remarkable that the convergent methods (and even the divergent ones before their divergence) appear to have the exact same slope as with $a = 10.26$ bohrs. This is consistent with our result: assuming the main effect of increasing the lattice constant is to decrease the gap, while keeping the lowest eigenvalues of $\boldsymbol{\Omega}_*^{-1}\boldsymbol{K}_*$ constant, then for $\beta$ small enough the convergence is limited by the lowest, not the highest,

eigenvalues of this operator.

In all these cases, Anderson acceleration was able to converge to a solution, even in presence of a very small gap, albeit in an irregular fashion. We attribute this to the well-known fact that, in the linear regime, Anderson acceleration is equivalent to the GMRES algorithm. Since the GMRES algorithm is a Krylov method, it is robust to the presence of a large eigenvalue, and achieves convergence even though the underlying iteration is strongly divergent. This shows a limitation of our theoretical convergence rates, which do not capture the reduced sensitivity of accelerated methods to a small gap.

## 2.5 Conclusion

In this chapter, we examined the convergence of two simple representatives in the class of direct minimization and SCF algorithms. We showed that both algorithms converge locally when the damping parameter is chosen small enough. We derived their convergence rates; we showed that the damped SCF algorithm is sensitive to the gap, while the gradient method is sensitive to the spectral radius of the Hamiltonian. We confirmed these results with numerical experiments. The goal here was not to propose efficient algorithms, but to analyse the behaviour of the simplest representatives of each class. However, accelerated algorithms are generally found to follow the trend suggested by our theoretical results, although we showed that the Anderson-accelerated SCF algorithm was able to converge quickly even in the presence of a single very small gap.

In practice, should the SCF or direct minimization class of algorithms be preferred? The answer depends not only on the convergence rate studied in this chapter, but also on the cost of each step, and the robustness of the algorithm. We examine two prototypical situations.

In quantum chemistry using Gaussian basis sets to solve the Hartree–Fock model or Kohn–Sham density functional theory using hybrid functionals, the rate-limiting step is often the computation of the Fock matrix $H(P)$. In this case, an iteration of a gradient descent and a damped SCF algorithm are of roughly equal cost. In most cases, solutions for isolated molecules satisfy the Aufbau principle, and the damped SCF algorithm, suitably robustified (for instance using the ODA algorithm) and accelerated (for instance with the DIIS algorithm), converges reliably and efficiently towards a solution. Direct minimization algorithms are then only useful in the cases where local or semilocal functionals are used [177] and the Aufbau principle is violated, or when SCF algorithms tend to converge to saddle points (for instance for computations involving spin).

In condensed-matter physics using plane-wave basis sets to solve the Kohn–Sham density functional theory with local or semilocal functionals, the matrices $P$ and $H$ are not stored explicitly. Solving the linear eigenproblem is then done using iterative block eigensolvers, which can be understood as specialized direct minimization algorithms in the case of a linear energy functional $E(P) = \operatorname{Tr}(H_0 P)$. In this case, direct minimization algorithms effectively merge the two loops of the SCF and linear eigensolver, and should therefore be more efficient. Another interest of direct minimization algorithms is their robustness, as the choice of a step size can be made in order to minimize the energy, unlike the damped SCF algorithm where choosing an appropriate damping parameter is often done empirically.

Despite this, direct minimization algorithms are rarely used in condensed-matter physics. The main reason seems to be that challenging problems are often metallic in character, and require the introduction of a finite temperature. Direct minimization algorithms then need to optimize over the occupations as well as the orbitals, a significantly more complex task [28, 75, 145]. A thorough comparison of the performance and robustness of direct minimization and self-consistent approach for these systems would be an interesting topic of inquiry. A number of implementation "tricks" commonly used to accelerate the convergence of iterative eigensolvers (for instance, using a block size larger than the number of eigenvectors sought) might also play a large role in performance comparison for the two classes of algorithms: understanding how to generalize these to direct minimization would be interesting.

We discussed in Remark 2.4 preconditioning for both direct minimization and SCF algorithms. The concept of preconditioning for Riemannian optimization problems seems not to have been explored much in the mathematical literature, except in some specific models and preconditioners (see for instance [13, 208] for the Gross–Pitaevskii model), and a deeper analysis of this would be interesting. In particular, this

is necessary to extend the convergence theory presented in this chapter to infinite-dimensional settings.

## Appendix: proof of Lemma 2.2

For any $\alpha \in \mathbb{R}_+$ the energy functional $E_\alpha$ is smooth and convex on the nonempty compact convex set $\mathrm{CH}(\mathcal{M}_N)$. The set of minimizers to (2.4.8) is therefore nonempty, compact and convex. Let $P_*$ and $P'_*$ be two minimizers of $E_\alpha$ and let $\rho_*$ and $\rho'_*$ be their densities: $\rho_{*,i} = \delta^{-1}(P_*)_{ii}$ and $\rho'_{*,i} = \delta^{-1}(P'_*)_{ii}$. For all $\theta \in [0,1]$, we have

$$I_\alpha = E_\alpha(\theta P_* + (1-\theta)P'_*) = I_\alpha + \frac{\alpha}{2\delta}\sum_{i=1}^{N_\mathrm{b}}\Big((\theta\rho_{*,i} + (1-\theta)\rho'_{*,i})^2 - (\theta\rho_{*,i}^2 + (1-\theta){\rho'_{*,i}}^2)\Big),$$

where $I_\alpha = E_\alpha(P_*) = E_\alpha(P'_*)$ is the minimum of (2.4.8). Since the function $\mathbb{R} \ni x \mapsto x^2 \in \mathbb{R}$ is strictly convex, we obtain that $\rho_* = \rho'_*$. Therefore, all the minimizers of (2.4.8) share the same density, hence the same mean-field Hamiltonian matrix $H_*$. If $P_*$ is a minimizer of (2.4.8), it satisfies the first-order optimality condition (Euler inequality)

$$\forall\ P \in \mathrm{CH}(\mathcal{M}_N), \quad \mathrm{Tr}(H_*(P - P_*)) \geqslant 0,$$

from which we infer by a classical argument that

$$P_* = \mathbf{1}_{(-\infty,\mu)}(H_*) + Q_* \quad \text{with} \quad 0 \leqslant Q_* \leqslant 1, \quad \mathrm{Ran}\,(Q_*) \subset \mathrm{Ker}\,(H_* - \mu), \quad \mathrm{Tr}(P_*) = N, \tag{2.5.1}$$

for some Fermi level $\mu \in \mathbb{R}$ (the Lagrange multiplier of the constraint $\mathrm{Tr}(P) = N$). Let $\varepsilon_1 \leqslant \cdots \leqslant \varepsilon_{N_\mathrm{b}}$ be the eigenvalues of $H_*$, counting multiplicities. If $\varepsilon_N < \varepsilon_{N+1}$, then we necessarily have $P_* = \mathbf{1}_{(-\infty,\varepsilon_N]}(H_*)$, so that (2.4.8) has a unique minimizer, $P_*$ is on $\mathcal{M}_N$ and therefore is also the unique minimizer of (2.4.6), and it satisfies the strong *Aufbau* principle.

Let us now consider the case when $\varepsilon_N = \varepsilon_{N+1} =: \mu$. Since the eigenvalue problem $H_*\psi = \mu\psi$ is a second-order difference equation

$$\frac{-\psi_{i+1} + 2\psi_i - \psi_{i-1}}{2\delta^2} + V_{\mathrm{eff},i}\psi_i = \mu\psi_i, \quad 1 \leqslant i \leqslant N_\mathrm{b}, \tag{2.5.2}$$

(here and in the sequel we use the convention that $\psi_0 = \psi_{N_\mathrm{b}}$ and $\psi_{N_\mathrm{b}+1} = \psi_1$) with $V_\mathrm{eff} = V + \alpha\rho_*$, the eigenspace $\mathrm{Ker}\,(H_* - \mu)$ is at most of dimension 2. We therefore have $\varepsilon_{N-1} < \varepsilon_N = \varepsilon_{N+1} < \varepsilon_{N+2}$.

Using the variational characterization of the ground-state eigenvalue, we have

$$\varepsilon_1 = \min_{\psi\in\mathbb{R}^{N_\mathrm{b}},\,\psi^*\psi=1} \psi^* H_*\psi \quad \text{with} \quad \psi^* H_*\psi = \sum_{i=1}^{N_\mathrm{b}}\left|\frac{\psi_{i+1} - \psi_i}{\delta}\right|^2 + \sum_{i=1}^{N_\mathrm{b}} V_{\mathrm{eff},i}|\psi_i|^2. \tag{2.5.3}$$

Since $||x| - |y|| \leqslant |x - y|$ for all $x, y \in \mathbb{R}$ with equality if and only if $x$ and $y$ have the same sign, we infer from (2.5.3), that all the entries of a ground-state eigenvector of $H_*$ have the same sign. In particular, two normalized ground-state eigenvectors of $H_*$ cannot be orthogonal. This implies that the ground-state eigenvalue of $H_*$ is simple, *i.e.* $\varepsilon_1 < \varepsilon_2$. The first statement of Lemma 2.2 straightforwardly follows from the results established so far.

To prove the second statement, assume that $N \geqslant 2$ and that (2.4.8) has two distinct minimizers $P_*$ and $P'_*$ sharing the same density. In view of (2.5.1), this can only occur if $\varepsilon_N = \varepsilon_{N+1} =: \mu$. Using an orthonormal basis $(\phi, \psi)$ of $\mathrm{Ker}\,(H_* - \mu)$ consisting of eigenvectors of $P_*$, we can assume without loss of generality that

$$\begin{aligned} P_* &= \mathbf{1}_{(-\infty,\mu)}(H_*) + (1-f)\phi\phi^* + f\psi\psi^*, \\ P'_* &= \mathbf{1}_{(-\infty,\mu)}(H_*) + (1-a)\phi\phi^* + a\psi\psi^* + b(\phi\psi^* + \psi\phi^*), \end{aligned}$$

with $0 \leqslant f \leqslant 1$, $0 \leqslant a \leqslant 1$ and $b^2 \leqslant a(1-a)$. Since $P_*$ and $P'_*$ have the same density, we have for all $1 \leqslant i \leqslant N_\mathrm{b}$, $(1-f)\phi_i^2 + f\psi_i^2 = (1-a)\phi_i^2 + a\psi_i^2 + 2b\phi_i\psi_i$ , that is $(a-f)\phi_i^2 - 2b\phi_i\psi_i - (a-f)\psi_i^2 = 0$.

If $a = f$, then $b \neq 0$ since $P_* \neq P'_*$ by assumption, so that $\phi_i \psi_i = 0$ for all $1 \leqslant i \leqslant N_\mathrm{b}$. From (2.5.2), we see that it is not possible to have $\psi_i = \psi_{i+1} = 0$ (otherwise, $\psi$ would be identically equal to zero), and the same holds true for $\phi$. Therefore, $N_\mathrm{b}$ must be even, and either all the odd entries of $\phi$ and all the even entries $\psi$ must vanish, or the other way round. We then infer from (2.5.2) that this implies that $\phi_{i+2} + \phi_i = \psi_{i+2} + \psi_i = 0$ for all $1 \leqslant i \leqslant N_\mathrm{b}$, and that all the entries of $V_\mathrm{eff}$ are equal to $\mu - \delta^{-2}$. This implies that $N_\mathrm{b} \in 4\mathbb{N}^*$ and that the states $\phi$ and $\psi$ are given by $\phi_{2i} = c(-1)^i$, $\phi_{2i+1} = 0$, $\psi_{2i} = 0$, $\psi_{2i+1} = c'(-1)^i$ for all $1 \leqslant i \leqslant N_\mathrm{b}$, where $c$ and $c'$ are normalization constants. By explicit diagonalization of the matrix $H_* = H(0) + (\mu - \delta^{-2}) I_{N_\mathrm{b}}$ one can check that the states $\phi$ and $\psi$ are therefore those spanning the two-dimensional space associated to the two-fold degenerate eigenvalues $\varepsilon_{N_\mathrm{b}/2} = \varepsilon_{1+N_\mathrm{b}/2}$ of $H_*$. This is only possible if $N = N_\mathrm{b}/2$. The case $\alpha = f$ can thus be excluded for $2 \leqslant N \leqslant N_\mathrm{b}$, with $N_\mathrm{b} \neq 2N$ if $N_\mathrm{b} \in 4\mathbb{N}^*$.

If $a \neq f$, we have for all $1 \leqslant i \leqslant N_\mathrm{b}$, $\phi_i^2 - 2\gamma\phi_i\psi_i - \psi_i^2 = 0$, for $\gamma = \frac{b}{a-f}$, and up to replacing $\psi$ with $-\psi$, we can assume without loss of generality that $\gamma \geqslant 0$. Denoting by $C_\pm \coloneqq \gamma \pm \sqrt{1+\gamma^2}$ the roots of the polynomial $x^2 - 2\gamma x + 1$, with $C_+C_- = -1$, we obtain that for each $1 \leqslant i \leqslant N_\mathrm{b}$, either $\phi_i = C_+\psi_i$ or $\phi_i = C_-\psi_i$. Using the discrete Schrödinger equation (2.5.2) satisfied by both $\phi$ and $\psi$, we see that if $\phi_i = C_+\psi_i$ and $\phi_{i+1} = C_+\psi_{i+1}$ for some $1 \leqslant i \leqslant N_\mathrm{b}$, then $\phi = C_+\psi$, and likewise if $C_+$ is replaced by $C_-$. This is impossible since $\phi$ and $\psi$ are orthonormal. Therefore, we must have $\phi_{2i} = C_+\psi_{2i}$ and $\phi_{2i+1} = C_-\psi_{2i+1}$ (or the other way around), and $N_\mathrm{b}$ must be even. Using again (2.5.2), this leads to $\phi_{i+2} + \phi_i = 0$ and $\psi_{i+2} + \psi_i = 0$ for all $1 \leqslant i \leqslant N_\mathrm{b}$ and therefore, as in the previous case, that $N_\mathrm{b} \in 4\mathbb{N}^*$, $N_\mathrm{b} = 2N$, that all the entries of $V_\mathrm{eff}$ are equal and that $\phi$ and $\psi$ span the two-dimensional space associated to the two-fold degenerate eigenvalues $\varepsilon_N = \varepsilon_{N+1}$ of $H_*$.

This proves that for $2 \leqslant N \leqslant N_\mathrm{b}$, with $N_\mathrm{b} \neq 2N$ if $N_\mathrm{b} \in 4\mathbb{N}^*$, (2.4.8) has a unique minimizer $P_*$. If $P_* \in \mathcal{M}_N$, it is of course also the unique minimizer of (2.4.6), and $P_*$ satisfies the *Aufbau* principle. Conversely, if $P'_* \in \mathcal{M}_N$ is a local minimizer of (2.4.6) satisfying the *Aufbau* principle, we have

$$\forall\ P \in \mathrm{CH}(\mathcal{M}_N), \quad \mathrm{Tr}(H(P'_*)(P - P'_*)) \geqslant 0,$$

which means that $P'_*$ is a solution to the Euler inequality for (2.4.8), and therefore a global minimizer of this convex problem. Since the minimizer $P_*$ of (2.4.8) is unique, we finally obtain that if $P_* \notin \mathcal{M}_N$, then none of the local minimizers of (2.4.6) satisfies the *Aufbau* principle.

Chapter 3

# Practical error bounds for properties in plane-wave electronic structure calculations

This chapter has been published in the article [GK2]:

*Eric Cancès, Geneviève Dusson, Gaspard Kemlin and Antoine Levitt. Practical error bounds for properties in plane-wave electronic structure calculations. SIAM Journal on Scientific Computing, 44(5):B1312-B1340 (2022).* https://arxiv.org/abs/2111.01470

**Abstract** We propose accurate computable error bounds for quantities of interest in plane-wave electronic structure calculations, in particular ground-state density matrices and energies, and interatomic forces. These bounds are based on an estimation of the error in terms of the residual of the solved equations, which is then efficiently approximated with computable terms. After providing coarse bounds based on an analysis of the inverse Jacobian, we improve on these bounds by solving a linear problem in a small dimension that involves a Schur complement. We numerically show how accurate these bounds are on a few representative materials, namely silicon, gallium arsenide and titanium dioxide.

## Contents

**3.1 Introduction** . . . 62
**3.2 Mathematical framework** . . . 64
3.2.1 General framework . . . 64
3.2.2 First-order geometry . . . 65
3.2.3 Second-order geometry . . . 65
3.2.4 Density matrix and orbitals . . . 66
3.2.5 Metrics on the tangent space . . . 68
3.2.6 Correspondence rules . . . 69
**3.3 The periodic Kohn–Sham problem** . . . 69
3.3.1 The continuous problem . . . 69
3.3.2 Discretization . . . 70
3.3.3 Forces . . . 71
3.3.4 Numerical setup . . . 71
**3.4 A first error bound using linearization** . . . 72
3.4.1 Linearization in the asymptotic regime . . . 72
3.4.2 A simple error bound based on operator norms . . . 73
3.4.3 Error bounds on QoIs and applications to interatomic forces . . . 75
**3.5 Improved error bounds based on frequencies splitting** . . . 77
3.5.1 Spectral decomposition of the error . . . 77
3.5.2 Improving the error estimation . . . 78
**3.6 Numerical examples with more complex systems** . . . 80
**3.7 Conclusion** . . . 82

## 3.1 Introduction

This chapter focuses on providing practical error estimates for numerical approximations of electronic structure calculations. Such computations are key in many domains, as they allow for the simulation of systems at the atomic and molecular scale. They are particularly useful in the fields of chemistry, materials science, condensed matter physics, or molecular biology. Among the many electronic structure models available, Kohn–Sham (KS) density functional theory (DFT) [116] with semilocal density functionals is one of the most used in practice, as it offers a good compromise between accuracy and computational cost. We will focus on this model in this chapter. Note that the mathematical formulation of this problem is similar to that of the Hartree–Fock [89] or Gross–Pitaevskii equations [165], so that what we present in the context of DFT can be easily extended to such contexts. We will focus on plane-wave discretizations within the pseudopotential approximation, which are most suited for the study of crystals; some (but not all) of our methodology can be applied in other contexts as well, including the aforementioned Hartree–Fock or Gross–Pitaevskii models, as well as molecules simulated using plane-wave DFT.

In this field, the first and most crucial problem is the determination of the electronic ground-state of the system under consideration. Mathematically speaking, this problem is a constrained minimization problem. Writing the first-order optimality conditions of this problem leads to an eigenvalue problem that is nonlinear in the eigenvectors. At the continuous level, the unknown is a subspace of dimension $N_{\mathrm{el}}$, the number of electrons in the system; this subspace can be conveniently described using either the orthogonal projector on it (density matrix formalism) or an orthonormal basis of it (orbital formalism). This problem is well-known in the literature and the interested reader is referred to [GK1] and the references therein for more information on how it is solved in practice.

Solving this problem numerically requires a number of approximations, so that only an approximation of the exact solution can be computed. Being able to estimate the error between the exact and the approximate solutions is crucial, as this information can be used to reduce the high computational cost of such methods by an optimization of the approximation parameters, and maybe more importantly, to add error bars to quantities of interest (QoI) calculated from the approximate solution. In our context, such QoI are typically the ground-state energy of the system and the forces on the atoms in the system, but may also include any information computable from the Kohn–Sham ground-state.

While such error bounds have been developed already some time ago for boundary value problems, *e.g.* in the context of finite element discretization, using in particular *a posteriori* error estimation [197], the development of bounds in the context of electronic structure is quite recent, and still incomplete. Computable and guaranteed error bounds for linear eigenvalue problems have been proposed in the last decade [35, 36, 37, 47, 137]; we refer to [153, Chapter 10] for a recent monograph on the subject. Specifically for electronic structure calculations, some of us proposed guaranteed error bounds for linear eigenvalue equations [100]. For fully self-consistent (nonlinear) eigenvalue equations, an error bound was proposed for a simplified 1D Gross–Pitaevskii equation in [70]; however the computational cost of evaluating the error bound in this contribution is quite high. So far, no error bound has been proposed for the error estimation of general QoI in electronic structure calculations, in particular for the interatomic forces. This is what we are trying to achieve in this contribution.

In this chapter, we use a general approach based on a linearization of the Kohn–Sham equations. It is instructive to compare our approach to those used in a general context. Assume we want to find $x \in \mathbb{R}^n$ such that $f(x) = 0$, for some nonlinear function $f : \mathbb{R}^n \to \mathbb{R}^n$ (the residual). Near a solution $x_*$, we have $f(x) \approx f'(x)(x - x_*)$, and therefore, if $f'(x)$ is invertible, we have the error-residual relationship

$$x - x_* \approx f'(x)^{-1} f(x). \tag{3.1.1}$$

This is the same approximation that leads to the Newton algorithm. Assume now that we want to compute a real-valued QoI $A(x_*)$, where $A : \mathbb{R}^n \to \mathbb{R}$ is a $C^1$ function (*e.g.* the energy, a component of the interatomic forces, of the density, ...); then we have the approximate equality with computable right-hand side:

$$A(x) - A(x_*) \approx \nabla A(x) \cdot \big(f'(x)^{-1} f(x)\big). \tag{3.1.2}$$

From here, we obtain the simple first estimate

$$|A(x) - A(x_*)| \leqslant |\nabla A(x)| \big\| f'(x)^{-1} \big\|_{\mathrm{op}} |f(x)|,$$

where $|\cdot|$ is any chosen norm on $\mathbb{R}^n$, and $\|\cdot\|_{\text{op}}$ is the induced operator norms on $\mathbb{R}^{n\times n}$ (note that $\nabla A(x) \in \mathbb{R}^n$ and $f'(x) \in \mathbb{R}^{n\times n}$). This approximate bound can be turned into a rigorous one using information on the second derivatives of $f$; see for instance [181].

In extending this approach to Kohn–Sham models, we encounter several difficulties:

- The structure of our problem is not easily formulated as above because of the presence of constraints and degeneracies. We solve this using the geometrical framework of [GK1] to identify the appropriate analogue to the Jacobian $f'(x)$.
- The computation of the Jacobian and its inverse is prohibitively costly. We use iterative strategies to keep this cost manageable.
- Choosing the right norm is not obvious in this context. For problems involving partial differential equations, where $f$ includes partial derivatives, it is natural to consider Sobolev-type norms, with the aim of making $f'$ a bounded operator between the relevant function spaces. We explore different choices and their impacts on the error bounds.
- The operator norm inequalities

$$\left|\nabla A(x) \cdot \left(f'(x)^{-1} f(x)\right)\right| \leqslant |\nabla A(x)| \left\|f'(x)^{-1}\right\|_{\text{op}} |f(x)|$$

  are very often largely suboptimal, even with appropriate norms. We quantify this on representative examples.
- The structure of the residual $f(x)$ plays an important role. For instance, when $x$ results from a Galerkin approximation to a partial differential equation, $f(x)$ is orthogonal to the approximation space. In the context of plane-wave discretizations, this means the residual only contains high-frequency Fourier components. We demonstrate how this impacts the quality of the above bounds when $A$ represents the interatomic forces, in which case its derivative mostly acts on low-frequency components.

The main result of our work therefore lies in the derivation of an efficient, asymptotically accurate, way of approximating $\nabla A(x) \cdot \left(f'(x)^{-1} f(x)\right)$ using the specific structure of the residual $f(x)$ in a plane-wave discretization, where $A$ represents a component of the interatomic forces of the system. This approximation can then be used either to approach the actual error $A(x) - A(x_*)$ in (3.1.2) or to improve $A(x)$ by computing $A(x) - \nabla A(x) \cdot \left(f'(x)^{-1} f(x)\right)$, which is a better approximation of $A(x_*)$. These estimates are a new step towards robust and guaranteed *a posteriori* error estimates for Kohn–Sham models: this chapter reflects the process that lead to their derivation, by describing the issues raised when applying natural ideas and how we propose to solve these issues.

Throughout the chapter, we will provide numerical tests to illustrate our results. All these tests are performed with the DFTK software [101], a recent `Julia` package solving the Kohn–Sham equations in the pseudopotential approximation using a plane-wave basis, thus particularly well suited for periodic systems such as crystals [144]. We are mostly interested in three QoI: the ground-state energy, the ground-state density, and the interatomic forces, the latter being computed using the Hellmann–Feynman theorem. We will demonstrate the main points with simulations on a simple system (bulk silicon), then present results for more complicated systems.

We will be interested in this chapter only in quantifying the discretization error. However, the general framework we develop can be used also to treat other types of error (such as the ones resulting from the iterative solution of the underlying minimization problem). We only consider insulating systems at zero temperature, and do not consider the error due to finite Brillouin zone sampling [39]; extending the formalism to finite temperature and quantifying the sampling error, particularly for metals, is currently under investigation, see [99] for a first step in this direction.

The outline of this chapter is as follows. In Section 3.2, we present the mathematical framework related to the solution of the electronic structure minimization problem, describing in particular objects related to the tangent space of the constraint manifold. In Section 3.3, we present the Kohn–Sham model and the numerical framework in which our tests will be performed. In Section 3.4, we propose a first crude bound of the error between the exact and the numerically computed solution based on a linearization

argument as well as an operator norm inequality, both for the error on the ground-state density matrix and on the forces. In Section 3.5, we refine this bound by splitting between low and high frequencies, and using a Schur complement to refine the error bound on the low frequencies. Finally, in Section 3.6, we provide numerical simulations on more involved materials systems, namely on a gallium arsenide system (GaAs) and a titanium dioxide system ($TiO_2$), showing that the proposed bounds work well in those cases.

## 3.2 Mathematical framework

In this section, we present the models targeted by our study, as well as the elementary tools of differential geometry used to derive and compute the error bounds on the QoI.

### 3.2.1 General framework

The work we present here is valid for a large class of mean-field models including different instances of Kohn–Sham models, the Hartree–Fock model, and the time-independent Gross–Pitaevskii model and its various extensions. To study them in a unified way, we use a mathematical framework similar to the one in [GK1]. To keep the presentation simple, we will work in finite dimension and consider that the solutions of the problem can be very accurately approximated in a given finite-dimensional space of (high) dimension $\mathcal{N}$, which we identify with $\mathbb{C}^{\mathcal{N}}$. We denote by

$$\langle x, y \rangle \coloneqq \mathrm{Re}\,(x^* y)$$

the $\ell^2$ inner product of $\mathbb{C}^{\mathcal{N}}$, seen as a vector space over $\mathbb{R}$. We equip the $\mathbb{R}$-vector space of square Hermitian matrices

$$\mathcal{H} \coloneqq \mathbb{C}^{\mathcal{N}\times\mathcal{N}}_{\mathrm{herm}}$$

with the Frobenius inner product $\langle A, B\rangle_{\mathrm{F}} \coloneqq \mathrm{Re}\,(\mathrm{Tr}(A^*B))$. Note that although it is important in applications to allow for complex orbitals and density matrices, the space of Hermitian matrices is not a vector space over $\mathbb{C}$, and therefore we will always consider vector spaces over $\mathbb{R}$.

The density-matrix formulation of the mean-field model in this reference approximation space reads

$$\min\{E(P),\ P \in \mathcal{M}_{\mathcal{N}}\}, \quad \text{where} \quad \mathcal{M}_{\mathcal{N}} \coloneqq \left\{P \in \mathcal{H} \;\middle|\; P^2 = P,\ \mathrm{Tr}(P) = N_{\mathrm{el}}\right\} \tag{3.2.1}$$

is the manifold of rank-$N_{\mathrm{el}}$ orthogonal projectors (density matrices), and $E : \mathcal{H} \to \mathbb{R}$ is a $C^2$ nonlinear energy functional. The parameter $N_{\mathrm{el}}$ (with $1 \leqslant N_{\mathrm{el}} \leqslant \mathcal{N}$) is a fixed integer depending on the physical model, and not on its discretization in a finite-dimensional space. For mean-field electronic structure models, $N_{\mathrm{el}}$ is the number of electrons or electron pairs in the system (hence the notation $N_{\mathrm{el}}$); in the standard Gross–Pitaevskii model, $N_{\mathrm{el}} = 1$. The energy functional $E$ is of the form

$$E(P) \coloneqq \mathrm{Tr}(H_0 P) + E_{\mathrm{nl}}(P),$$

where $H_0$ is the linear part of the mean-field Hamiltonian, and $E_{\mathrm{nl}}$ a nonlinear contribution depending on the considered model (see Section 3.3 for the expressions in the Kohn–Sham model). For simplicity of presentation we will ignore spin in the formalism, but we will include it in the numerical simulations (see Remark 3.2). The set $\mathcal{M}_{\mathcal{N}}$ is diffeomorphic to the Grassmann manifold of $N_{\mathrm{el}}$-dimensional complex vector subspaces of $\mathbb{C}^{\mathcal{N}}$.

Problem (3.2.1) always has a minimizer since it consists in minimizing a continuous function on a compact set. This minimizer may or may not be unique, depending on the model and/or the physical system under study. We will not elaborate here on this uniqueness issue, and assume for the sake of simplicity that (3.2.1) has a unique minimizer, which we denote by $P_*$.

Besides the ground-state energy $E(P_*)$, we can compute from $P_*$ various other physical quantities of interest (QoI), for instance, the electronic density and the interatomic forces. We denote such a QoI by $A_* = A(P_*)$.

We consider the case when $\mathcal{N}$ is too large for problem (3.2.1) to be solved completely in the reference approximation space. To solve problem (3.2.1), we therefore consider a finite-dimensional subspace $\mathcal{X}$ of $\mathbb{C}^{\mathcal{N}}$ of dimension $N_{\rm b}$ and solve instead the variational approximation of (3.2.1) in $\mathcal{X}$, namely

$$\min\{E(P),\ P \in \mathcal{M}_{\mathcal{N}},\ \mathrm{Ran}\,(P) \subset \mathcal{X}\}. \tag{3.2.2}$$

Our goal is then to estimate the errors $\|A(P) - A_*\|$ on the QoI $A$, where $P$ is typically the minimizer of (3.2.2), given the variational space $\mathcal{X}$, and the norm is specific to the QoI. The latter can be a scalar, *e.g.* the ground-state energy, or a finite or infinite dimensional vector, *e.g.* the interatomic forces, or the ground-state density.

### 3.2.2 First-order geometry

The manifold $\mathcal{M}_{\mathcal{N}}$ is a smooth manifold. Its tangent space $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ at $P \in \mathcal{M}_{\mathcal{N}}$ is given by

$$\begin{aligned}\mathcal{T}_P\mathcal{M}_{\mathcal{N}} &= \{X \in \mathcal{H} \mid PX + XP = X,\ \mathrm{Tr}(X) = 0\}\\ &= \left\{X \in \mathcal{H} \;\middle|\; PXP = 0, P^\perp X P^\perp = 0\right\},\end{aligned}$$

where $P^\perp = 1 - P$ is the orthogonal projection on $\mathrm{Ran}\,(P)^\perp$ for the canonical inner product of $\mathbb{C}^{\mathcal{N}}$. The set $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ is the set of Hermitian matrices that are off-diagonal in the block decomposition induced by $P$ and $P^\perp$; more explicitly, if $P = U\begin{pmatrix} I_{N_{\rm el}} & 0 \\ 0 & 0\end{pmatrix}U^*$ for some unitary $U \in \mathrm{U}(\mathcal{N})$, then

$$\mathcal{T}_P\mathcal{M}_{\mathcal{N}} = \left\{X = U\begin{pmatrix} 0 & Y^* \\ Y & 0\end{pmatrix}U^*,\ Y \in \mathbb{C}^{(\mathcal{N}-N_{\rm el})\times N_{\rm el}}\right\}.$$

The orthogonal projection $\mathbf{\Pi}_P$ on $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ for the Frobenius inner product is given by

$$\forall\ X \in \mathcal{H}, \quad \mathbf{\Pi}_P(X) = PXP^\perp + P^\perp XP = [P,[P,X]] \in \mathcal{T}_P\mathcal{M}_{\mathcal{N}}, \tag{3.2.3}$$

where $[A,B] \coloneqq AB - BA$ is the commutator of $A$ and $B$. Linear operators acting on spaces of matrices are sometimes referred to as *super-operators* in the physics literature. Throughout this chapter, super-operators will be written in bold fonts.

The mean-field Hamiltonian is the gradient of the energy at a given point $P$ (again for the Frobenius inner product):

$$H(P) \coloneqq \nabla E(P) = H_0 + \nabla E_{\rm nl}(P).$$

To simplify the notation, we set

$$H_* = H(P_*) = \nabla E(P_*). \tag{3.2.4}$$

The first-order optimality condition for (3.2.1) is that $\nabla E(P_*)$ is orthogonal to the tangent space $\mathcal{T}_{P_*}\mathcal{M}_{\mathcal{N}}$, which can be written, using (3.2.3) and (3.2.4), as $\mathbf{\Pi}_{P_*}H(P_*) = [P_*,[P_*,H(P_*)]] = 0$. This corresponds to the residual

$$R(P) = \mathbf{\Pi}_P H(P) = [P,[P,H(P)]]$$

being zero at $P_*$. The residual function $R$ can be seen as a nonlinear map from $\mathcal{H}$ to itself, and its restriction to $\mathcal{M}_{\mathcal{N}}$ as a vector field on $\mathcal{M}_{\mathcal{N}}$ since for all $P \in \mathcal{M}_{\mathcal{N}}$, $R(P) \in \mathcal{T}_P\mathcal{M}_{\mathcal{N}}$.

### 3.2.3 Second-order geometry

We introduce the super-operators $\mathbf{\Omega}(P)$ and $\boldsymbol{K}(P)$, defined at $P \in \mathcal{M}_N$ and acting on $\mathcal{H}$. These operators were already introduced in [GK1, Section 2.2], but we recall here their definitions for completeness. To simplify the notation, we will set $\boldsymbol{K}_* \coloneqq \boldsymbol{K}(P_*)$, $\mathbf{\Omega}_* \coloneqq \mathbf{\Omega}(P_*)$.

The super-operator $\boldsymbol{K}_* \in \mathcal{L}(\mathcal{H})$ is the Hessian of the energy projected onto the tangent space to $\mathcal{M}_{\mathcal{N}}$ at $P_*$:

$$\boldsymbol{K}_* \coloneqq \mathbf{\Pi}_{P_*}\boldsymbol{\nabla}^2 E(P_*)\mathbf{\Pi}_{P_*} = \mathbf{\Pi}_{P_*}\boldsymbol{\nabla}^2 E_{\rm nl}(P_*)\mathbf{\Pi}_{P_*}. \tag{3.2.5}$$

By construction, $\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}$ is an invariant subspace of $\boldsymbol{K}_*$. Note that $\boldsymbol{K}_* = 0$ for linear eigenvalue problems, *i.e.* when $E_{\rm nl} = 0$.

The super-operator $\boldsymbol{\Omega}_* \in \mathcal{L}(\mathcal{H})$ is defined by

$$\forall\, X \in \mathcal{H}, \quad \boldsymbol{\Omega}_* X = -[P_*, [H_*, X]]. \tag{3.2.6}$$

The tangent space $\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}$ is also an invariant subspace of $\boldsymbol{\Omega}_*$.

It is shown in [GK1] that the energy of a density matrix $P = P_* + X + O\Big(\|X\|_{\rm F}^2\Big) \in \mathcal{M}_\mathcal{N}$ with $X \in \mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}$ is $E(P) = E(P_*) + \langle X, (\boldsymbol{\Omega}_* + \boldsymbol{K}_*)X\rangle_{\rm F} + o\Big(\|X\|_{\rm F}^2\Big)$. The restriction of the operator $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ to the invariant subspace $\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}$ can therefore be identified with the second-order derivative of $E$ on the manifold $\mathcal{M}_\mathcal{N}$. Since $P_*$ is a minimum, it follows that

$$\boldsymbol{\Omega}_* + \boldsymbol{K}_* \geqslant 0 \quad \text{on } \mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}.$$

Note that in general, the second-order derivative of a function defined on a smooth manifold is not an intrinsic object; it depends not only on the tangent structure of the manifold, but also on the chosen affine connection. However, at the critical point $P_*$ of $E$ on the manifold, the contributions to the second derivative due the connection vanish and the second-order derivative becomes intrinsic.

For our purposes, it will be convenient to define this second-order derivative also outside of $P_*$. Relying on (3.2.5)–(3.2.6), we define for any $P \in \mathcal{M}_\mathcal{N}$ the super-operators $\boldsymbol{\Omega}(P) \in \mathcal{L}(\mathcal{H})$ and $\boldsymbol{K}(P) \in \mathcal{L}(\mathcal{H})$ through

$$\forall\, X \in \mathcal{H}, \quad \boldsymbol{\Omega}(P)X = -[P, [H(P), X]] \quad \text{and} \quad \boldsymbol{K}(P)X = \boldsymbol{\Pi}_P \boldsymbol{\nabla}^2 E(P) \boldsymbol{\Pi}_P X. \tag{3.2.7}$$

Both $\boldsymbol{\Omega}(P)$ and $\boldsymbol{K}(P)$ admit $\mathcal{T}_P\mathcal{M}_\mathcal{N}$ as an invariant subspace and their restrictions to $\mathcal{T}_P\mathcal{M}_\mathcal{N}$ are Hermitian for the Frobenius inner product. The map $\mathcal{M}_\mathcal{N} \ni P \mapsto \boldsymbol{\Omega}(P) + \boldsymbol{K}(P) \in \mathcal{L}(\mathcal{H})$ is smooth and the restriction of $\boldsymbol{\Omega}(P) + \boldsymbol{K}(P)$ to $\mathcal{T}_P\mathcal{M}_\mathcal{N}$ provides a computable approximation of the second-order derivative of $E : \mathcal{M}_\mathcal{N} \to \mathbb{R}$ in a neighbourhood of $P_*$ (whatever the choice of the affine connection).

### 3.2.4 Density matrix and orbitals

The framework we have outlined above is particularly convenient for stating the second-order conditions, but much too expensive computationally as it requires the storage and manipulation of (low-rank) large matrices. In practice, it is more effective to work directly with orbitals, *i.e.* write for any $P \in \mathcal{M}_\mathcal{N}$

$$P = \Phi\Phi^* = \sum_{i=1}^{N_{\rm el}} |\phi_i\rangle\langle\phi_i| \tag{3.2.8}$$

where $\Phi = (\phi_1|\cdots|\phi_{N_{\rm el}})$ is a collection of $N_{\rm el}$ orbitals $\phi_i \in \mathbb{C}^\mathcal{N}$ satisfying $\Phi^*\Phi = I_{N_{\rm el}}$ and $\mathrm{Span}\,(\phi_1, \ldots, \phi_{N_{\rm el}}) = \mathrm{Ran}\,(P)$, and where we used Dirac's bra-ket notation: for $\phi, \psi \in \mathbb{C}^\mathcal{N}$, $\langle\phi, \psi\rangle = \phi^*\psi$ and $|\phi\rangle\langle\psi| = \phi\psi^*$. Problem (3.2.1) can be reformulated as

$$\min\big\{E(\Phi\Phi^*), \Phi \in \mathbb{C}^{\mathcal{N}\times N_{\rm el}}, \Phi^*\Phi = I_{N_{\rm el}}\big\}.$$

Note that the orbitals are only defined up to a unitary transform: if $U \in \mathrm{U}(N_{\rm el})$ is a unitary matrix, then $\widetilde{\Phi} \coloneqq \Phi U$ and $\Phi$ give rise to the same density matrix. This means that the minimizers of this minimization problem are never isolated, which creates technical difficulties that are not present in the density matrix formalism.

Let us fix a $\Phi = (\phi_1|\cdots|\phi_{N_{\rm el}}) \in \mathbb{C}^{\mathcal{N}\times N_{\rm el}}$ with $\Phi^*\Phi = I_{N_{\rm el}}$, and consider an element $X$ of the tangent plane $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_\mathcal{N}$. By completing $\Phi$ to an orthogonal basis and writing out $X$ in this basis, it is easy to see that the constraints $X^* = X$, $PXP = 0$, $P^\perp X P^\perp = 0$ imply that $X$ can be put in the form

$$X = \sum_{i=1}^{N_{\mathrm{el}}} |\phi_i\rangle\langle\xi_i| + |\xi_i\rangle\langle\phi_i| = \Phi\Xi^* + \Xi\Phi^* \tag{3.2.9}$$

where $\Xi = (\xi_1|\cdots|\xi_{N_{\mathrm{el}}}) \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}$ is a set of orbital variations satisfying $\Phi^*\Xi = 0$. Furthermore, under this condition, $\Xi$ is unique, so that (3.2.9) establishes a bijection between $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_{\mathcal{N}}$ and $\{\Xi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}} \mid \Phi^*\Xi = 0\}$. We will therefore treat equivalently elements of the tangent space $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ either in the density matrix representation $X$ or the orbital representation $\Xi$, writing

$$\Xi \simeq_\Phi X. \tag{3.2.10}$$

This orbital representation of $P$ by $\Phi$ is more economical computationally, only requiring the storage and manipulation of orbitals $\Phi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}$ satisfying $\Phi^*\Phi = I_{N_{\mathrm{el}}}$. Similarly, the manipulation of objects $X$ in the tangent plane $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_{\mathcal{N}}$ is more efficiently done through their orbital variations $\Xi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}$ satisfying $\Phi^*\Xi = 0$.

All operations on density matrices or their variations can indeed be carried out in this orbital representation. For instance, the computation of the energy can be performed efficiently in practice, as explained in Section 3.3, and the residual at $P = \Phi\Phi^*$ also has a nice representation in terms of orbitals:

$$R(\Phi\Phi^*) \simeq_\Phi H\Phi - \Phi(\Phi^* H\Phi) \quad \text{with } H \text{ evaluated at } \Phi\Phi^*,$$

which is easily recognized as similar to the residual of a linear eigenvalue problem.

Likewise, operators on $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_{\mathcal{N}}$ can be identified in this fashion. For instance,

$$\boldsymbol{\Omega}(\Phi\Phi^*)(\Phi\Xi^* + \Xi\Phi^*) \simeq_\Phi P^\perp(H\Xi - \Xi(\Phi^* H\Phi)) \quad \text{with } H \text{ evaluated at } \Phi\Phi^*. \tag{3.2.11}$$

The computation of $\boldsymbol{K}$ can be performed similarly:

$$\boldsymbol{K}(\Phi\Phi^*)(\Phi\Xi^* + \Xi\Phi^*) \simeq_\Phi P^\perp(\delta H\phi_i)_{i=1,\dots,N_{\mathrm{el}}} \quad \text{with } \delta H = \frac{\mathrm{d}H}{\mathrm{d}P}(\Phi\Xi^* + \Xi\Phi^*). \tag{3.2.12}$$

Finally, note that all the numerical results in this chapter are performed using the orbital formalism.

**Remark 3.1.** Note that the condition that $\Phi^*\Xi = 0$ is not necessary for $\Phi\Xi^* + \Xi\Phi^*$ to define an element of $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_{\mathcal{N}}$. However, without this gauge condition, $\Xi$ is not unique. This is simply a manifestation at the infinitesimal level of the noninjectivity of the map $\Phi \mapsto \Phi\Phi^*$ between $\{\Phi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}, \Phi^*\Phi = I_{N_{\mathrm{el}}}\}$ and $\mathcal{M}_{\mathcal{N}}$. Because of this, the derivative $\Xi \mapsto \Phi\Xi^* + \Xi\Phi^*$ is not injective between the tangent spaces $\{\Xi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}, \Phi^*\Xi + \Xi^*\Phi = 0\}$ and $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_{\mathcal{N}}$. In more concrete terms, in the example case where $\phi_1, \dots, \phi_{N_{\mathrm{el}}}$ are the first $N_{\mathrm{el}}$ basis vectors, any element $X$ is of the form $\begin{pmatrix} 0 & Z^* \\ Z & 0 \end{pmatrix}$ which can be written in the form (3.2.9) with $\Xi = \begin{pmatrix} 0 \\ Z \end{pmatrix}$. However, such an $X$ can also be written in the form (3.2.9) with $\Xi = \begin{pmatrix} A \\ Z \end{pmatrix}$ for any anti-hermitian matrix $A$. The gauge condition $\Phi^*\Xi = 0$ forces $A$ to be zero, making $\Xi$ unique. In more formal terms, the map $\Phi \mapsto \Phi\Phi^*$ induces a principal bundle structure on the base space $\mathcal{M}_{\mathcal{N}}$ (the Grassmann manifold) with total space $\{\Phi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}, \Phi^*\Phi = I_{N_{\mathrm{el}}}\}$ (the Stiefel manifold) and characteristic fiber $\mathrm{U}(N_{\mathrm{el}})$. This naturally splits the tangent space $\{\Xi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}, \Phi^*\Xi + \Xi^*\Phi = 0\}$ into the *vertical space* $\{\Phi A,\ A \text{ anti-hermitian}\}$, and a complementary *horizontal space*, which we take to be the orthogonal complement, $\{\Xi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}} \mid \Phi^*\Xi = 0\}$.

The orbital formalism can be used to give a more concrete interpretation of the first-order optimality condition $R(P_*) = 0$. Indeed, this condition can be rewritten as

$$P_* H_* P_*^\perp = 0, \quad P_*^\perp H_* P_* = 0,$$

from which it follows that $P_*$ and $H_* = H(P_*)$ can be jointly diagonalized in an orthonormal basis:

$$H_*\phi_{*n} = \lambda_{*n}\phi_{*n}, \quad \langle\phi_{*m}, \phi_{*n}\rangle = \delta_{mn}, \quad P_* = \sum_{n=1}^{N_{\mathrm{el}}} |\phi_{*n}\rangle\langle\phi_{*n}|. \tag{3.2.13}$$

In many applications, the orbitals $\phi_{*1}, \ldots, \phi_{*N_{\mathrm{el}}}$ spanning the range of $P_*$ (see (3.2.13)) are those corresponding to the lowest $N_{\mathrm{el}}$ eigenvalues of $H_*$. This is called the *Aufbau* principle in physics and chemistry. This principle is always satisfied in the (unrestricted) Hartree–Fock setting, and most of the times in the Kohn–Sham setting. Under the *Aufbau* principle, we can assume that the $\lambda_n$'s are ranked in nondecreasing order. The orbitals $\phi_i$, $1 \leqslant i \leqslant N_{\mathrm{el}}$, are called occupied, and the orbitals $\phi_a$, $N_{\mathrm{el}} \leqslant a \leqslant \mathcal{N}$, are called virtual (it is customary to label the occupied orbitals by indices $i, j, k, l$, and virtual orbitals by indices $a, b, c, d$). The operator $\boldsymbol{\Omega}_*$ can be written explicitly using the tensor basis $\phi_{*m} \otimes \phi_{*n}$. We have indeed

$$\boldsymbol{\Omega}_* = \sum_{i=1}^{N_{\mathrm{el}}} \sum_{a=N_{\mathrm{el}}+1}^{\mathcal{N}} (\lambda_a - \lambda_i)(|\phi_{*i} \otimes \phi_{*a}\rangle\langle\phi_{*i} \otimes \phi_{*a}| + |\phi_{*a} \otimes \phi_{*i}\rangle\langle\phi_{*a} \otimes \phi_{*i}|),$$

and it follows that the lowest eigenvalue of the restriction of $\boldsymbol{\Omega}_*$ to $\mathcal{T}_{P_*}\mathcal{M}_{\mathcal{N}}$ is $\lambda_{N_{\mathrm{el}}+1} - \lambda_{N_{\mathrm{el}}} \geqslant 0$. The operator $\boldsymbol{\Omega}_*$ is therefore positive on $\mathcal{T}_{P_*}\mathcal{M}_{\mathcal{N}}$, and coercive if there is an energy gap between the $N_{\mathrm{el}}{}^{\mathrm{th}}$ and $(N_{\mathrm{el}}+1)^{\mathrm{st}}$ eigenvalues of $H_*$ (see *e.g.* [GK1]).

### 3.2.5 Metrics on the tangent space

The isomorphism between $X = \Phi\Xi^* + \Xi\Phi^* \in \mathcal{T}_{\Phi\Phi^*}\mathcal{M}_{\mathcal{N}}$ and the set of orbital variations $\Xi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}$ with $\Phi^*\Xi = 0$ is unitary under the Frobenius inner product up to a factor of 2: $\|X\|_{\mathrm{F}}^2 = 2\|\Xi\|_{\mathrm{F}}^2$.

In practice, it is often advantageous to work using different inner products. This is in particular the case for partial differential equations involving unbounded operators, where using Sobolev-type metrics better respects the natural analytic structure of the problem and therefore allows for better bounds, compare *e.g.* the results of (5.34) and (5.35) on Figure 4 in [37]. To that end, consider a metric on $\mathbb{C}^{\mathcal{N}}$ given by

$$\langle \xi_1, \xi_2 \rangle_T = \langle \xi_1, T\xi_2 \rangle.$$

Here $T$ is a coercive Hermitian operator on $\mathbb{C}^{\mathcal{N}}$ representing the metric; for instance, taking $T$ to be a discretization of the operator $1 - \Delta$ we recover the classical Sobolev $\mathrm{H}^1$ norm. A basic problem is that the projection $P^\perp$ onto the orthogonal complement of $\mathrm{Ran}\,(P)$ does not necessarily commute with $T$. As a result, there are various nonequivalent ways to lift this metric to one on the tangent space $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_{\mathcal{N}}$. We select here the computationally simplest. The operator

$$M = P^\perp T^{1/2} P^\perp T^{1/2} P^\perp \tag{3.2.14}$$

is positive definite on the subspace $\mathrm{Ran}\,(P)^\perp$ of $\mathbb{C}^{\mathcal{N}}$, and induces a metric $\langle \xi_1, M\xi_2 \rangle$ on that space. The point of this formulation is to make it easy to compute $M^{1/2} = P^\perp T^{1/2} P^\perp$. Note that, since $P^\perp$ and $T$ do not commute, $M^{-1/2} \neq P^\perp T^{-1/2} P^\perp$. However, $P^\perp T^{-1/2} P^\perp M^{1/2}$ is well-conditioned, so that computing the action of $M^{-1/2}$ on a vector can be performed efficiently by an iterative algorithm involving repeated applications of the operators $T^{1/2}$ and $T^{-1/2}$. The same holds for $M^{-1}$. Furthermore, practical numerical results are typically not very sensitive to these issues, so that other (nonequivalent) reasonable alternatives to (3.2.14) yield similar results.

The metric on $\mathrm{Ran}\,(P)^\perp$ immediately induces a metric on $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_{\mathcal{N}}$ given by, in the orbital representation associated with $\Phi$,

$$\langle \Xi_1, \Xi_2 \rangle_{\boldsymbol{M}} = \mathrm{Re}\,(\mathrm{Tr}(\Xi_1^* M \Xi_2)) = \sum_{i=1}^{N_{\mathrm{el}}} \mathrm{Re}\,(\langle \xi_{1,i}, M\xi_{2,i} \rangle),$$

for $\Xi_1 = (\xi_{1,i})_{1\leqslant i\leqslant N_{\mathrm{el}}}, \Xi_2 = (\xi_{2,i})_{1\leqslant i\leqslant N_{\mathrm{el}}}$. This defines an operator $\boldsymbol{M}$ on $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_{\mathcal{N}}$ through the relationship $\boldsymbol{M}X \simeq_\Phi (M\xi_i)_{1\leqslant i\leqslant N_{\mathrm{el}}}$ when $X \simeq_\Phi (\xi_i)_{1\leqslant i\leqslant N_{\mathrm{el}}}$. Similarly to $M$, we can compute powers and inverses of $\boldsymbol{M}$ easily.

This formalism has the disadvantage that the same metric is used for every orbital variation. In practice this may not be sensible, as different orbitals can correspond to different energy ranges. Therefore

we slightly modify the above formalism by applying a different metric on each individual orbital variation, following standard practice used in preconditioners for plane-wave density functional theory [159]. Introducing a family $(T_1, \dots, T_{N_{\rm el}})$ of coercive Hermitian operators on $\mathbb{C}^{N_{\rm el}}$, we set

$$M_i \coloneqq P^\perp T_i^{1/2} P^\perp T_i^{1/2} P^\perp \quad \text{and} \quad \boldsymbol{M} X \simeq_\Phi (M_i \xi_i)_{1 \leqslant i \leqslant N_{\rm el}}. \tag{3.2.15}$$

### 3.2.6 Correspondence rules

As explained above, the density matrices will be preferred for the mathematical analysis while orbitals will be used in practice in the numerical simulations. We summarize below the correspondence between the density matrix and molecular orbital formulations, and the practical way the different operators we introduced are computed. For a given $P \in \mathcal{M}_\mathcal{N}$ and $(\phi_i)_{1 \leqslant i \leqslant N_{\rm el}}$ the associated set of occupied orbitals, there holds

$$\begin{aligned}
P \in \mathcal{M}_\mathcal{N} &\quad \leftrightarrow \quad \Phi = (\phi_1 | \cdots | \phi_{N_{\rm el}}) \in \mathbb{C}^{\mathcal{N} \times N_{\rm el}} \text{ s.t. } P = \Phi\Phi^* && \text{(state)},\\
X \in \mathcal{T}_P \mathcal{M}_\mathcal{N} &\quad \leftrightarrow \quad \Xi \in \mathbb{C}^{\mathcal{N} \times N_{\rm el}} \text{ s.t. } \Phi^*\Xi = 0 && \text{(perturbation)},\\
R(P) = [P,[P,H(P)]] &\quad \leftrightarrow \quad (r_1 | \cdots | r_{N_{\rm el}}) = P^\perp H(P)\Phi && \text{(residual)},\\
\boldsymbol{\Omega}(P)(X) &\quad \leftrightarrow \quad P^\perp(H\Xi - \Xi(\Phi^* H \Phi)) && \text{(see (3.2.11))},\\
\boldsymbol{K}(P)(X) &\quad \leftrightarrow \quad P^\perp(\delta H \phi_i)_{1 \leqslant i \leqslant N_{\rm el}} && \text{(see (3.2.12))},\\
\boldsymbol{M}^s X &\quad \leftrightarrow \quad (M_i^s \xi_i)_{1 \leqslant i \leqslant N_{\rm el}} \text{ for } s = -1, -1/2, 1/2, 1 && \text{(see (3.2.15))}.
\end{aligned}$$

## 3.3 The periodic Kohn–Sham problem

### 3.3.1 The continuous problem

We consider an $\mathcal{R}$-periodic system, $\mathcal{R}$ being a Bravais lattice with unit cell $\Gamma$ and reciprocal lattice $\mathcal{R}^*$ (the set of vectors $G$ such that $G \cdot R \in 2\pi\mathbb{Z}$ for all $R \in \mathcal{R}$). For the sake of simplicity, we present here the formalism for the (artificial) Kohn–Sham model for a finite system of $N_{\rm el}$ electrons on the unit cell $\Gamma$ with periodic boundary conditions. This is distinct from the more physical periodic Kohn–Sham problem for an infinite crystal with $N_{\rm el}$ electrons by unit cell, which is usually treated by using the supercell approach and Bloch theorem. Practical computations are performed for the latter model using Monkhorst–Pack Brillouin zone sampling [150] (see also [39] for a mathematical analysis of this method). The mathematical framework is very similar, with additional sums over $k$ points.

At the continuous level, a Kohn–Sham state is described by a density matrix $\gamma$, a rank-$N_{\rm el}$ orthogonal projector acting on the space $\mathrm{L}^2_\#$ of square integrable periodic functions. Ignoring constant terms modelling interactions between ions (*i.e.* atomic nuclear and frozen core electrons), the Kohn–Sham energy of $\gamma$ is given by $E^{\rm KS}(\gamma) = \mathrm{Tr}(h_0\gamma) + E^{\rm Hxc}(\rho_\gamma)$ (the superscript Hxc stands for Hartree-exchange-correlation), with

$$h_0 = -\frac{1}{2}\Delta + v_{\rm loc} + v_{\rm nloc}, \quad E^{\rm Hxc}(\rho) = \int_\Gamma \left( \frac{1}{2}\rho V_{\rm H}(\rho)(x) + e_{\rm xc}(\rho(x)) \right) \mathrm{d}x.$$

In the above expressions, $\rho_\gamma$ is the density associated with the trace-class operator $\gamma$ (formally $\rho_\gamma(x) = \gamma(x,x)$ where $\gamma(x,x')$ is the integral kernel of $\gamma$), $e_{\rm xc} : \mathbb{R}_+ \to \mathbb{R}$ a given exchange-correlation energy, and $V_{\rm H}(\rho)$ the Hartree potential, defined as the unique periodic solution with zero mean of the Poisson equation $-\Delta V_{\rm H}(\rho) = 4\pi\big(\rho - \fint_\Gamma \rho\big)$. In the pseudopotential approximation that we use in our numerical results, $v_{\rm loc}$ is a local potential given by

$$\forall\, x \in \mathbb{R}^3, \quad v_{\rm loc}(x) \coloneqq \sum_{R \in \mathcal{R}} \sum_{j=1}^{N_{\rm at}} v_{\rm loc}^j(x - (X_j + R)), \tag{3.3.1}$$

and $v_{\mathrm{nloc}}$ a nonlocal potential in Kleinmann–Bylander [114] form given by

$$v_{\mathrm{nloc}} \coloneqq \sum_{R\in\mathcal{R}} \sum_{j=1}^{N_{\mathrm{at}}} \sum_{a,b=1}^{n_{\mathrm{proj},j}} \left|p_a^j(\cdot - (X_j+R))\right\rangle C_{ab}^j \left\langle p_b^j(\cdot - (X_j+R))\right|, \tag{3.3.2}$$

where $N_{\mathrm{at}}$ is the number of atoms in $\Gamma$, the $X_j$'s are the positions of the atoms inside the unit cell $\Gamma$, $v_{\mathrm{loc}}^j : \mathbb{R}^3 \to \mathbb{R}$ is a local radial potential, $n_{\mathrm{proj},j}$ denotes the number of projectors for atom $j$, and the $p_{ab}^j : \mathbb{R}^3 \to \mathbb{C}$ are given smooth functions. We use in particular the Goedecker–Teter–Hutter (GTH) pseudopotentials [79, 91] whose functional forms for the $v_{\mathrm{loc}}^j$ and $p_{ab}^j$ are analytic ($v_{\mathrm{loc}}^j$ is a radial Gaussian function multiplied by a radial polynomial, and $p_{ab}^j$ is a radial Gaussian function multiplied by a solid spherical harmonics).

The Kohn–Sham Hamiltonian associated to a density matrix $\gamma$ is given by

$$h_\gamma = h_0 + V_{\mathrm{H}}(\rho_\gamma) + e'_{\mathrm{xc}}(\rho_\gamma),$$

where $V_{\mathrm{H}}(\rho_\gamma)$ and $e'_{\mathrm{xc}}(\rho_\gamma)$ are interpreted as local (multiplication) operators. Similarly, we have

$$D^2_\gamma(E^{\mathrm{Hxc}}(\rho_\gamma)) \cdot Q = V_{\mathrm{H}}(\rho_Q) + e''_{\mathrm{xc}}(\rho_\gamma)\rho_Q.$$

**Remark 3.2** (Spin)**.** The expressions above are given for a system of "spinless electrons" to accommodate the simple geometrical formalism of Section 3.2. Real systems (and the numerical simulations we perform in the following sections) include spin; in this case, the energy is $E^{\mathrm{KS}}(\gamma) = 2\mathrm{Tr}(h_0\gamma) + E^{\mathrm{Hxc}}(\rho_\gamma)$, where $\rho_\gamma(x) = 2\gamma(x,x)$, $\nabla E^{\mathrm{KS}}(\gamma) = 2(h_0 + V_{\mathrm{H}}(\rho_\gamma) + e'_{\mathrm{xc}}(\rho_\gamma))$ and $D^2(E^{\mathrm{KS}}(\gamma)) \cdot Q = 4V_{\mathrm{H}}(\rho_Q) + 4e''_{\mathrm{xc}}(\rho_\gamma)\rho_Q$.

### 3.3.2 Discretization

For each vector $G$ of the reciprocal lattice $\mathcal{R}^*$, we denote by $e_G$ the Fourier mode with wave-vector $G$:

$$\forall\, x \in \mathbb{R}^3, \quad e_G(x) \coloneqq \frac{1}{\sqrt{|\Gamma|}} \exp(\mathrm{i}G\cdot x)$$

where $|\Gamma|$ is the Lebesgue measure of the unit cell $\Gamma$. The family $(e_G)_{G\in\mathcal{R}^*}$ is an orthonormal basis of $\mathrm{L}^2_\#$, the space of locally square integrable $\mathcal{R}$-periodic functions (and an orthogonal basis of the $\mathcal{R}$-periodic Sobolev space $\mathrm{H}^s_\#$, endowed with its usual inner product, for any $s \in \mathbb{R}$). In the so-called plane-wave discretization methods, the Kohn–Sham model is discretized using the finite-dimensional approximation spaces

$$\mathcal{X}_{E_{\mathrm{cut}}} \coloneqq \mathrm{Span}\left\{ e_G,\ G \in \mathcal{R}^* \;\middle|\; \frac{1}{2}|G|^2 \leqslant E_{\mathrm{cut}} \right\},$$

where $E_{\mathrm{cut}} > 0$ is a given energy cut-off chosen by the user.

The connection with the formalism introduced in Section 3.2 is the following:

- we choose a large reference energy cut-off $E_{\mathrm{cut,ref}}$ and set

$$\mathcal{N} \coloneqq \dim(\mathcal{X}_{E_{\mathrm{cut,ref}}}) = \#\left\{ G \in \mathcal{R}^* \mid \frac{1}{2}|G|^2 \leqslant E_{\mathrm{cut,ref}} \right\};$$

- we identify $\mathcal{X}_{E_{\mathrm{cut,ref}}}$ with $\mathbb{C}^{\mathcal{N}}$ by labelling the reciprocal lattice vectors from 1 to $\mathcal{N}$ in such a way that for all $1 \leqslant i < j \leqslant \mathcal{N}$, $|G_i| \leqslant |G_j|$;

- the set of rank-$N_{\mathrm{el}}$ orthogonal projectors $\gamma$ on $\mathrm{L}^2_\#$ such that $\mathrm{Ran}\,(\gamma) \subset \mathcal{X}_{E_{\mathrm{cut,ref}}}$ can then be identified with the manifold $\mathcal{M}_{\mathcal{N}}$ defined in (3.2.1) through the mapping

$$\gamma = \sum_{i,j=1}^{\mathcal{N}} P_{ij} |e_{G_i}\rangle\langle e_{G_j}|;$$

- the noninteracting Hamiltonian matrix $H_0 \in \mathbb{C}^{\mathcal{N}\times\mathcal{N}}$ has entries

$$[H_0]_{ij} = \langle e_{G_i} | h_0 | e_{G_j} \rangle_{\mathrm{L}^2_\#},$$

and the nonlinear component of the energy $E_{\mathrm{nl}} : \mathcal{H} \to \mathbb{R}$ is any $C^2$-extension of the function defined on $\mathcal{M}_\mathcal{N}$ by

$$E_{\mathrm{nl}}(P) = E^{\mathrm{Hxc}}(\rho_P) \quad \text{where} \quad \rho_P(x) = |\Gamma|^{-1/2} \sum_{i,j=1}^{\mathcal{N}} P_{ij} e_{G_j - G_i}(x).$$

The entries of the core Hamiltonian matrix can be computed explicitly:

$$[H_0]_{ij} = \frac{|G_i|^2}{2}\delta_{i,j} + [V_{\mathrm{loc}}]_{ij} + [V_{\mathrm{nloc}}]_{ij}$$

$$\text{with} \quad [V_{\mathrm{loc}}]_{ij} = \langle e_{G_i} | v_{\mathrm{loc}} | e_{G_j} \rangle_{\mathrm{L}^2_\#} \quad \text{and} \quad [V_{\mathrm{nloc}}]_{ij} = \langle e_{G_i} | v_{\mathrm{nloc}} | e_{G_j} \rangle_{\mathrm{L}^2_\#},$$

where the above inner products can be computed exactly through the Fourier transforms of the $v^j_{\mathrm{loc}}$ and $p^j_{ab}$ (known exactly for GTH pseudopotentials). Note also that the density $\rho_P$ can be expanded on a finite number of Fourier modes and can therefore be easily stored in memory. Since the Poisson equation is trivially solvable in the plane-wave basis, this enables the exact computation of the Hartree energy. The exchange-correlation energy however cannot be computed explicitly, and is approximated using numerical quadrature. In all the numerical results, we select the parameters of this numerical quadrature such that it does not affect too much the results, see Remark 3.5.

### 3.3.3 Forces

The total ground-state energy depends on the atomic positions $\mathfrak{X} = (X_j)_{1\leqslant j\leqslant N_{\mathrm{at}}}$ both explicitly (ion-ion interaction energy and ion-electron interaction potentials $V_{\mathrm{loc}}$ and $V_{\mathrm{nloc}}$) and through the fact that the solution $P_*$ depends on $\mathfrak{X}$:

$$\mathcal{E}(\mathfrak{X}) = E(\mathfrak{X}, P_*(\mathfrak{X})).$$

The force acting on atom $j$ is defined as $F_j(\mathfrak{X}) = -\nabla_{X_j}\mathcal{E}(\mathfrak{X})$. Because of the Hellman–Feynman theorem, the term involving the derivative of $P_*$ with respect to $X_j$ vanishes [144], and the final result is

$$F_j = -\mathrm{Tr}\big((\nabla_{X_j}(V_{\mathrm{loc}} + V_{\mathrm{nloc}}))P_*\big). \tag{3.3.3}$$

This involves the partial derivatives of the matrix elements of $V_{\mathrm{loc}} + V_{\mathrm{nloc}}$ with respect to the atomic positions, which can be computed analytically from (3.3.1) and (3.3.2).

### 3.3.4 Numerical setup

For all the computations and examples on silicon, we use the DFTK software [101] within the LDA approximation, with Teter 93 exchange-correlation functional [79] and a $2 \times 2 \times 2$ $k$-point grid, and a reference solution computed with $E_{\mathrm{cut,ref}} = 125$ Ha, to which we compare results obtained with smaller values of $E_{\mathrm{cut}}$. We checked that $E_{\mathrm{cut,ref}} = 125$ Ha was a high enough energy cut-off to have fully converged results, up to the accuracy we need to test our numerical methods. We use the usual periodic lattice for the FCC phase of silicon, with lattice constant $a = 10.26$ bohrs, close to the equilibrium configuration. All results are expressed in atomic units: energies are in hartree and forces are in hartree/bohr. Note that the discretization grid of the Brillouin zone is not fine enough to have fully converged results, but is still sufficient to illustrate our points. Note also that the same results are observed for semilocal functionals, such as PBE-GGA [160]. Other functionals, such as meta-GGA and hybrid functionals, are out of the scope of this chapter.

The two atoms of silicon inside a cell are placed at first at their equilibrium positions with fractional coordinates $(-\frac{1}{8}, -\frac{1}{8}, -\frac{1}{8})$ and $(\frac{1}{8}, \frac{1}{8}, \frac{1}{8})$, and then the second one is slightly displaced by $\frac{1}{20}(0.24, -0.33, 0.12)$ to get nonzero interatomic forces.

The discretized Kohn–Sham equations are solved by a standard SCF procedure. The main computational bottleneck is the partial diagonalization of the mean-field Hamiltonian at each SCF step. This is done using an iterative eigenvalue solver, which only requires applying mean-field Hamiltonian matrices to a set of $N_{\mathrm{el}}$ trial orbitals and simple operations on vectors. In a plane-wave basis set of size $N_{\mathrm{b}}$, the former operation can be done efficiently through the use of the fast Fourier transform for a total cost of $O(N_{\mathrm{el}} N_{\mathrm{b}}(\log N_{\mathrm{b}} + \sum_j n_{\mathrm{proj},j}))$. We refer to [144] for more details. The application of the super-operators $\boldsymbol{\Omega}$ and $\boldsymbol{K}$ to a set of $N_{\mathrm{el}}$ orbital variations (see (3.2.11) and (3.2.12)) involves additional linear algebra operations, for an additional cost of $O({N_{\mathrm{el}}}^2(N_{\mathrm{b}} + N_{\mathrm{el}}))$.

In this setting, the reference values for the energy is $E_* = -7.838$ Ha and the interatomic forces are, in hartree/bohr,

$$F_* = \begin{bmatrix} -0.0656 & 0.0656 \\ 0.0619 & -0.0619 \\ -0.0352 & 0.0352 \end{bmatrix},$$

where the first column are the forces acting on the first atom in each direction, and the second column are the forces acting on the second atom.

## 3.4 A first error bound using linearization

Now that the mathematical and numerical frameworks are laid down, we turn to the estimation of the error between the reference solution computed with a large energy cut-off $E_{\mathrm{cut,ref}}$ and approximations thereof. We first start by deriving a linearization estimate and illustrating numerically its applicability. We then propose a very coarse bound on the error on the density matrix and the forces, based on the (expensive) evaluation of an operator norm. We will show in the next section how to improve this bound.

### 3.4.1 Linearization in the asymptotic regime

We assume that $P_*$ is a nondegenerate local minimizer of $E$ in the sense that there exists $\eta > 0$ such that $\boldsymbol{\Omega}_* + \boldsymbol{K}_* \geqslant \eta$ on the tangent space $\mathcal{T}_{P_*}\mathcal{M}_{\mathcal{N}}$. This implies in particular that $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ is invertible on the invariant subspace $\mathcal{T}_{P_*}\mathcal{M}_{\mathcal{N}}$.

Recall that for any trial density matrix $P \in \mathcal{M}_{\mathcal{N}}$, the residual of the problem is

$$R(P) = \boldsymbol{\Pi}_P H(P) = [P, [P, H(P)]] \in \mathcal{T}_P\mathcal{M}_{\mathcal{N}},$$

so that $R$ defines a smooth vector field on $\mathcal{M}_{\mathcal{N}}$ (a section of the tangent bundle $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$) which vanishes at $P_*$. For $P \in \mathcal{M}_{\mathcal{N}}$ in the vicinity of $P_*$, we have

$$P - P_* = \boldsymbol{\Pi}_{P_*}(P - P_*) + O\Big(\|P - P_*\|_{\mathrm{F}}^2\Big) = \boldsymbol{\Pi}_P(P - P_*) + O\Big(\|P - P_*\|_{\mathrm{F}}^2\Big). \tag{3.4.1}$$

It follows from the definitions (3.2.5)–(3.2.6) of $\boldsymbol{\Omega}_*$ and $\boldsymbol{K}_*$ that $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ is the Jacobian of the map $P \mapsto R(P)$ at $P_*$. Therefore, the optimality condition $R(P_*) = 0$ and the above expansions yield, for all $P \in \mathcal{M}_{\mathcal{N}}$ close enough to $P_*$,

$$\begin{aligned} R(P) &= (\boldsymbol{\Omega}_* + \boldsymbol{K}_*)\boldsymbol{\Pi}_{P_*}(P - P_*) + O\Big(\|P - P_*\|_{\mathrm{F}}^2\Big) \\ &= (\boldsymbol{\Omega}(P) + \boldsymbol{K}(P))\boldsymbol{\Pi}_P(P - P_*) + O\Big(\|P - P_*\|_{\mathrm{F}}^2\Big). \end{aligned} \tag{3.4.2}$$

By continuity, $\boldsymbol{\Omega}(P) + \boldsymbol{K}(P) \geqslant \frac{\eta}{2}$ on the tangent space $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ for $P \in \mathcal{M}_{\mathcal{N}}$ close enough to $P_*$, so that the restriction of the super-operator $\boldsymbol{\Omega}(P) + \boldsymbol{K}(P)$ to the invariant subspace $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ is self-adjoint and invertible. Using again (3.4.1) and the fact that $R(P) \in \mathcal{T}_P\mathcal{M}_{\mathcal{N}}$, we obtain, after inversion of $\boldsymbol{\Omega}(P) + \boldsymbol{K}(P)$ on the tangent space,

$$\boxed{P - P_* = ((\boldsymbol{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_{\mathcal{N}}})^{-1} R(P) + O\Big(\|P - P_*\|_{\mathrm{F}}^2\Big).} \tag{3.4.3}$$

This error-residual equation is the analogue in our case of the linearization (3.1.1), which identifies the super-operator $\boldsymbol{\Omega}(P) + \boldsymbol{K}(P)$ as the fundamental object in our study.

Based on this expansion, we can formulate the Newton algorithm to solve the equation $R(P_*) = 0$:

$$P^{k+1} = \mathfrak{R}_{P^k}\Big(P^k - (\boldsymbol{\Omega}^k + \boldsymbol{K}^k)^{-1} R(P^k)\Big),$$

where $\boldsymbol{\Omega}^k \coloneqq \boldsymbol{\Omega}(P^k)|_{\mathcal{T}_{P_k}\mathcal{M}_\mathcal{N}}$ and $\boldsymbol{K}^k \coloneqq \boldsymbol{K}(P^k)|_{\mathcal{T}_{P_k}\mathcal{M}_\mathcal{N}}$ and $\mathfrak{R}$ is a suitable retraction on $\mathcal{M}_\mathcal{N}$. A possible retraction is given in [GK1]. This Newton algorithm is expensive in practice, as it requires to solve iteratively a linear system; the cost of a Newton step is comparable to that of a full self-consistent field cycle. It is however a useful theoretical tool, and a starting point for further analysis and approximations.

To check the validity of the linearization (3.4.3), we focus on three quantities of interest: the ground-state energy, the ground-density density, and the interatomic forces acting on the two atoms in $\Gamma$. The reference values $E_*$, $\rho_*$ and $F_*$ of these QoIs are those obtained with the very large energy cut-off $E_{\text{cut,ref}} = 125$ Ha, defining a "fine grid" in real space *via* the discrete Fourier transform. For $E_{\text{cut}} < E_{\text{cut,ref}}$ defining a "coarse grid" in real space, we compute two approximations of the three QoIs:

1. $E_{\text{SCF}}$, $\rho_{\text{SCF}}$ and $F_{\text{SCF}}$ denote the approximations obtained from the variational solution of the Kohn–Sham problem on the coarse grid;
2. $E_{\text{Newton}}$, $\rho_{\text{Newton}}$ and $F_{\text{Newton}}$ denote the ones computed from the Kohn–Sham state obtained by one Newton step on the fine grid, starting from the variational solution of the Kohn–Sham problem on the coarse grid: as the SCF is converged on the coarse grid, we perform the Newton step on the fine grid in order to improve the approximation of $P_*$. That is, if $P$ is the variational solution on the coarse grid, $P - ((\boldsymbol{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}})^{-1} R(P)$ is a much better approximation of $P_*$ (for the metrics adapted to the chosen three QoIs).

The errors between these approximations and the reference values are plotted in Figure 3.1 as functions of $E_{\text{cut}}$. The errors on the ground-state density are measured with the $\mathrm{L}^2_\#$ metric, while the errors on the forces are measured with the Euclidean metric on $\mathbb{R}^{3\times 2}$.

For the simple case of a silicon crystal at the LDA level of theory, the linearization works very well, even for very small values of $E_{\text{cut}}$'s of the order of 5 Ha. Indeed the Kohn–Sham ground-state obtained by variational approximation on a coarse grid is significantly improved by one Newton step: the errors on the QoIs obtained with the latter are orders of magnitude smaller than the ones obtained with the former.

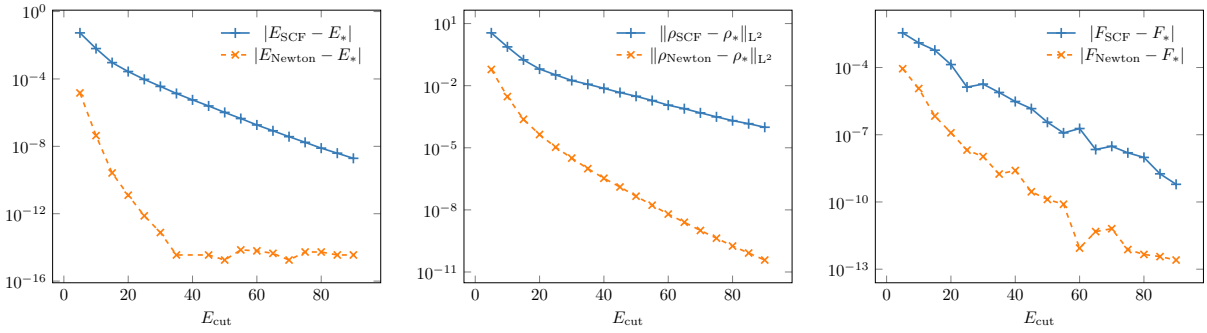



FIGURE 3.1 – Errors for the ground-state energy (hartree), ground-state density and interatomic forces (hartree/bohr) for Si as a function of $E_{\text{cut}}$, for both the variational solution of the Kohn–Sham problem on the coarse grid defined by $E_{\text{cut}}$ (solid line) and the post-processed solution obtained with one Newton step on the fine grid (dashed line). This shows that the linearization approximation is excellent, even for energy cut-offs as low as $E_{\text{cut}} = 5$ Ha.

### 3.4.2 A simple error bound based on operator norms

From (3.4.3) one can extract an error bound:

$$\begin{aligned} \|P - P_*\| &\approx \|\boldsymbol{\Pi}_P(P - P_*)\| \\ &\leqslant \big\|((\boldsymbol{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}})^{-1}\big\|_{\text{op}} \|R(P)\| \quad (+ \text{ h.o.t.}), \end{aligned} \tag{3.4.4}$$

where $\|\cdot\|_{\mathrm{op}}$ is the (super-)operator norm associated with the chosen norm $\|\cdot\|$ on $\mathcal{H}$. This bound is not guaranteed, but the results in Figure 3.1 suggest that it is very close to be guaranteed. Guaranteeing this bound could be done, provided that one could bound the higher-order terms rigorously [181]; this is an interesting prospect, but lies outside the scope of this chapter. To test the accuracy of this bound for a specific norm on $\mathcal{H}$, we would need to estimate the corresponding operator norm of the Hermitian operator $((\boldsymbol{\Omega}(P)+\boldsymbol{K}(P))_{\mathcal{T}_P\mathcal{M}_\mathcal{N}})^{-1}$ for all the $P$'s we are considering. In order to lower the computational burden, we consider instead the bound

$$\|P - P_*\| \leqslant \left\|((\boldsymbol{\Omega}_* + \boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}})^{-1}\right\|_{\mathrm{op}} \|R(P)\| \quad (+ \text{ h.o.t.}). \tag{3.4.5}$$

This enables us to compute the operator norm $\left\|((\boldsymbol{\Omega}_* + \boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}})^{-1}\right\|_{\mathrm{op}}$ only once, instead of computing it for every $P$. This is of course not accessible in practice, but we use it here for the sake of numerical experiment. Moreover, we can consider that the bounds (3.4.4) and (3.4.5) are almost equivalent since the results obtained in the previous section show that we are in the linear regime even for the lowest values of $E_{\mathrm{cut}}$ used in practice. The operator $(\boldsymbol{\Omega}_* + \boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}}$ is Hermitian for the Frobenius inner product and, thus, the operator norm $\left\|((\boldsymbol{\Omega}_* + \boldsymbol{K}_*)_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}})^{-1}\right\|_{\mathrm{op}}$ corresponding to the Frobenius norm on $\mathcal{H}$ is equal to the inverse of the smallest eigenvalue of $(\boldsymbol{\Omega}_* + \boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}}$. Standard iterative eigenvalue solvers for Hermitian operators can be used to compute this eigenvalue. We use here the LOBPCG algorithm [115].

We can see on Figure 3.2 (left panel) that when choosing the Frobenius norm on $\mathcal{H}$, the bound (3.4.5) leads to very crude error estimates: the error is overestimated by several orders of magnitude, and the bound becomes worse and worse as $E_{\mathrm{cut}}$ increases. This issue is well-known in the analysis of partial differential equations, where $\mathrm{L}^2$-type norms are not the natural ones to measure the error on the solution or the residual. Instead, for the Kohn–Sham equations and other second-order elliptic problems, it is more relevant to measure the error $P - P_*$ in $\mathrm{H}^1$-type Sobolev norms (energy norms) and the residual $R(P)$ in $\mathrm{H}^{-1}$-type Sobolev norms (dual norms). The linear operator linking the two quantities (here $(\boldsymbol{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}}$) is then expected to be a bounded isomorphism from the state error to the residual space for these norms. This suggests adapting the metrics on the tangent space $\mathcal{T}_P\mathcal{M}_\mathcal{N}$ in which we measure the error $P - P_*$ (or more precisely the leading term $\boldsymbol{\Pi}_P(P - P_*)$) on the one hand, and the residual $R(P)$ on the other hand. Similar considerations lead to the "kinetic energy preconditioning" used in practical computations [159]. Using the super-operator $\boldsymbol{M}$ on $\mathcal{T}_P\mathcal{M}_\mathcal{N}$ introduced in (3.2.15) with $T_i$ the diagonal operator on $\mathbb{C}^\mathcal{N}$ representing the operator $-\frac{1}{2}\Delta + t_i$ where $t_i = \frac{1}{2}\|\nabla\phi_i\|^2_{\mathrm{L}^2_\#}$ (kinetic energy of the $i^{\mathrm{th}}$ orbital), we obtain the bound

$$\left\|\boldsymbol{M}^{1/2}\boldsymbol{\Pi}_P(P - P_*)\right\|_{\mathrm{F}} \leqslant \left\|\boldsymbol{M}^{1/2}((\boldsymbol{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}})^{-1}\boldsymbol{M}^{1/2}\right\|_{\mathrm{op}} \left\|\boldsymbol{M}^{-1/2}R(P)\right\|_{\mathrm{F}}. \tag{3.4.6}$$

Here also, we lower the computational burden by replacing the first term in the RHS

$$\left\|\boldsymbol{M}^{1/2}((\boldsymbol{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}})^{-1}\boldsymbol{M}^{1/2}\right\|_{\mathrm{op}}$$

by the asymptotically equal quantity

$$\left\|\boldsymbol{M}_*^{1/2}((\boldsymbol{\Omega}_* + \boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}})^{-1}\boldsymbol{M}_*^{1/2}\right\|_{\mathrm{op}}.$$

The results are shown in Figure 3.2 (central panel). This time, the curves are almost parallel: the gap does not widen as $E_{\mathrm{cut}}$ increases. However, the bound is still an overestimate by more than one order of magnitude. This is due to the fact that

$$\left\|\boldsymbol{M}_*^{1/2}((\boldsymbol{\Omega}_* + \boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}})^{-1}\boldsymbol{M}_*^{1/2}\right\|_{\mathrm{op}} \approx 14.85$$

for this system, while the residual $R(P)$ is supported only on high-frequency Fourier modes, on which the operator $\boldsymbol{M}^{1/2}((\boldsymbol{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}})^{-1}\boldsymbol{M}^{1/2}$ is close to identity. The latter statement is supported by Proposition 3.1 in the appendix (see also the result [37, Proposition 5.10] concerning the linear setting). Thus, $\left\|\boldsymbol{M}^{-1/2}R(P)\right\|_{\mathrm{F}}$ is a good approximation of $\left\|\boldsymbol{M}^{1/2}\boldsymbol{\Pi}_P(P - P_*)\right\|_{\mathrm{F}}$, as shown on Figure 3.2 (central panel).

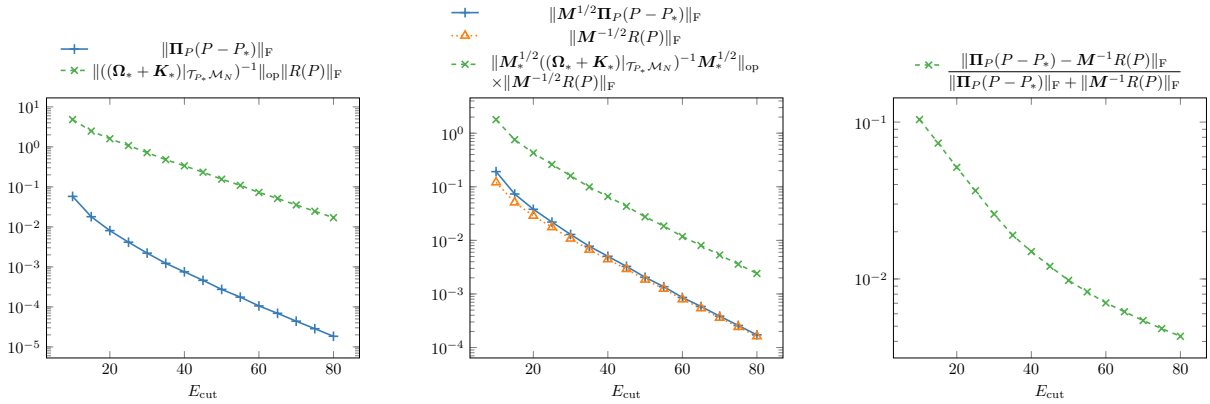



FIGURE 3.2 – Error bounds for Si based on (3.4.5) and (3.4.6). Left: $\mathrm{L}^2$-norm; Center: $\mathrm{H}^1$-type norm; Right: relative error between $\mathbf{\Pi}_P(P-P*)$ and $\boldsymbol{M}^{-1}R(P)$. It holds $\left\|((\mathbf{\Omega}_*+\boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_{\mathcal{N}}})^{-1}\right\|_{\mathrm{op}} \approx 11.23$ and $\left\|\boldsymbol{M}_*^{1/2}((\mathbf{\Omega}_*+\boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_{\mathcal{N}}})^{-1}\boldsymbol{M}_*^{1/2}\right\|_{\mathrm{op}} \approx 14.85$.

### 3.4.3 Error bounds on QoIs and applications to interatomic forces

Consider now a quantity of interest characterized by the smooth observable $A : \mathcal{M}_{\mathcal{N}} \to \mathcal{G}$, where $\mathcal{G}$ is a normed vector space (in particular, $\mathcal{G} = \mathbb{R}$ for real QoIs such as the ground-state energy, $\mathcal{G} = \mathbb{R}^{3N_{\mathrm{at}}}$ for interatomic forces, and, *e.g.*, $\mathcal{G} = \mathrm{L}^2_{\#}$ for the ground-state densities). For such a QoI, there holds for $P \in \mathcal{M}_{\mathcal{N}}$ in the vicinity of $P_*$,

$$A(P) - A_* = \mathrm{d}A(P) \cdot (\mathbf{\Pi}_P(P-P_*)) + \text{h.o.t.}, \tag{3.4.7}$$

where $\mathrm{d}A(P) \in \mathcal{L}(\mathcal{T}_P\mathcal{M}_{\mathcal{N}}; \mathcal{G})$ is the derivative of $A$ at $P$. We thus obtain the bound

$$\|A(P) - A_*\|_{\mathcal{G}} \leqslant \|\mathrm{d}A(P)\|_{\mathcal{T}_P\mathcal{M}_{\mathcal{N}}\to\mathcal{G}} \|\mathbf{\Pi}_P(P-P_*)\|_{\mathcal{T}_P\mathcal{M}_{\mathcal{N}}} \quad (+\text{ h.o.t.}) \tag{3.4.8}$$

for given norms $\|\cdot\|_{\mathcal{G}}$ and $\|\cdot\|_{\mathcal{T}_P\mathcal{M}_{\mathcal{N}}}$ on $\mathcal{G}$ and $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ respectively, and associated operator norm $\|\cdot\|_{\mathcal{T}_P\mathcal{M}_{\mathcal{N}}\to\mathcal{G}}$ on $\mathcal{L}(\mathcal{T}_P\mathcal{M}_{\mathcal{N}}; \mathcal{G})$.

Let us start with the simple case of the component of the force on atom $j$ along the direction $\alpha$ due to the local part of the pseudopotential. Since this QoI is scalar, we have $\mathcal{G} = \mathbb{R}$. Using (3.3.3), we get

$$F^{\mathrm{loc}}_{j,\alpha}(P) = -\mathrm{Tr}\left(\frac{\partial V_{\mathrm{loc}}}{\partial X_{j,\alpha}} P\right).$$

Thus (3.4.8) becomes, using the Frobenius norm on $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$,

$$|F^{\mathrm{loc}}_{j,\alpha}(P) - F^{\mathrm{loc}}_{j,\alpha}(P_*)| \leqslant \left\|\mathbf{\Pi}_P \frac{\partial V_{\mathrm{loc}}}{\partial X_{j,\alpha}}\right\|_{\mathrm{F}} \|\mathbf{\Pi}_P(P-P_*)\|_{\mathrm{F}} \quad (+\text{ h.o.t.}), \tag{3.4.9}$$

with

$$\mathbf{\Pi}_P \frac{\partial V_{\mathrm{loc}}}{\partial X_{j,\alpha}} \simeq_\Phi (1 - \Phi\Phi^*)\left(\frac{\partial V_{\mathrm{loc}}}{\partial X_{j,\alpha}}\Phi\right). \tag{3.4.10}$$

We plot in Figure 3.3 (left panel) the bound (3.4.9). The latter is pessimistic by more than three orders of magnitude, and its relative accuracy gets worse and worse as the cut-off energy increases.

The bound (3.4.9) using the operator norm of $\mathrm{d}A(P)$ being very inaccurate, we tested another approach consisting in using directly (3.4.7) to evaluate the error on the QoI by applying the derivative $\mathrm{d}A(P)$ to a computable approximation of $\mathbf{\Pi}_P(P-P_*)$. Relying on the results in the previous section showing that $\boldsymbol{M}^{-1/2}R(P)$ is a good approximation of $\boldsymbol{M}^{1/2}\mathbf{\Pi}_P(P-P_*)$ in Frobenius norm, it is tempting to replace $\mathbf{\Pi}_P(P-P_*)$ by $\boldsymbol{M}^{-1}R(P)$ in (3.4.7) and approximate $F(P)-F_*$ by $\mathrm{d}F(P)\cdot(\boldsymbol{M}^{-1}R(P))$, this approximation being justified by Figure 3.2 (right panel). Indeed the continuous counterpart of the asymptotic equivalence between $\boldsymbol{M}^{-1/2}R(P)$ and $\boldsymbol{M}^{1/2}\mathbf{\Pi}_P(P-P_*)$ for the Frobenius ($\mathrm{L}^2$-type) norm is that the preconditioned residual and the error on the density matrix are asymptotically equivalent in

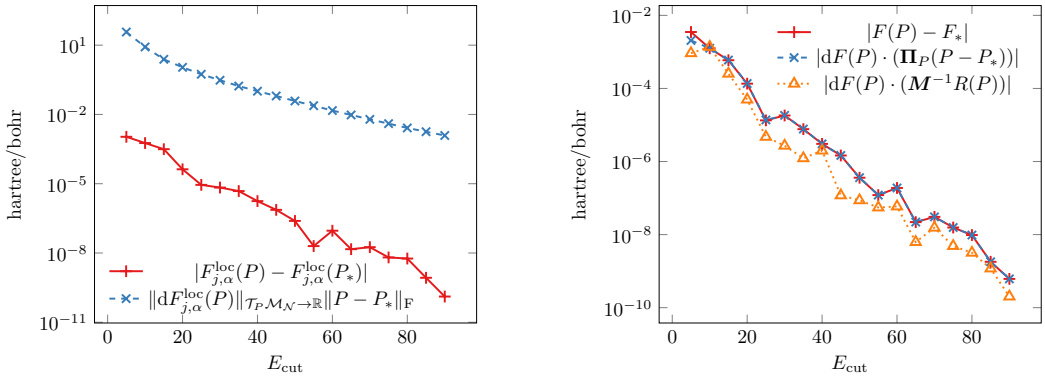



FIGURE 3.3 – Silicon. (Left panel) Inaccurate error bound (3.4.9) for the component of the force on atom $j = 1$ along direction $\alpha = (1, 0, 0)$ due to the local part of the pseudopotential. (Right panel) Approximation of $|F(P) - F_*|$ obtained by dropping the h.o.t. in the generic formula (3.4.7) and applying the derivative $\mathrm{d}F(P)$ either to the actual error $\mathbf{\Pi}_P(P-P_*)$ or the preconditioned residual $\boldsymbol{M}^{-1}R(P)$. The approximation $\mathrm{d}F(P)\cdot(\mathbf{\Pi}_P(P-P_*))$ matches asymptotically the error $F(P) - F_*$, validating again the rapid establishment of the linear regime. On the other hand, the approximation $\mathrm{d}F(P)\cdot(\boldsymbol{M}^{-1}R(P))$ does not match asymptotically.

$\mathrm{H}^1$-type norms, while the derivative of the interatomic forces observable is continuous on $\mathrm{H}^1$-type spaces. This idea is tested in Figure 3.3 (right panel). However, this leads to an underestimation of the error, although by a small factor. The reason is that even if $P - P_*$ and $M^{-1}R(P)$ do match asymptotically for the suitable norms, this is not the case for $\mathrm{d}F(P)\cdot(\mathbf{\Pi}_P(P - P_*))$ and $\mathrm{d}F(P)\cdot(\boldsymbol{M}^{-1}R(P))$ for reasons made clear in the next section.

**Remark 3.3.** In our simulations, the computation of $\mathrm{d}A(P)\cdot X$ for $X \in \mathcal{T}_P\mathcal{M}_\mathcal{N}$ is performed by forward-mode automatic differentiation using the `ForwardDiff.jl` `Julia` package [173].

We summarize the results of this section in Figure 3.4, displaying the combination of these bounds: the successive operator norms result in very inaccurate bounds (from six to eleven orders of magnitude) for the error on the forces.

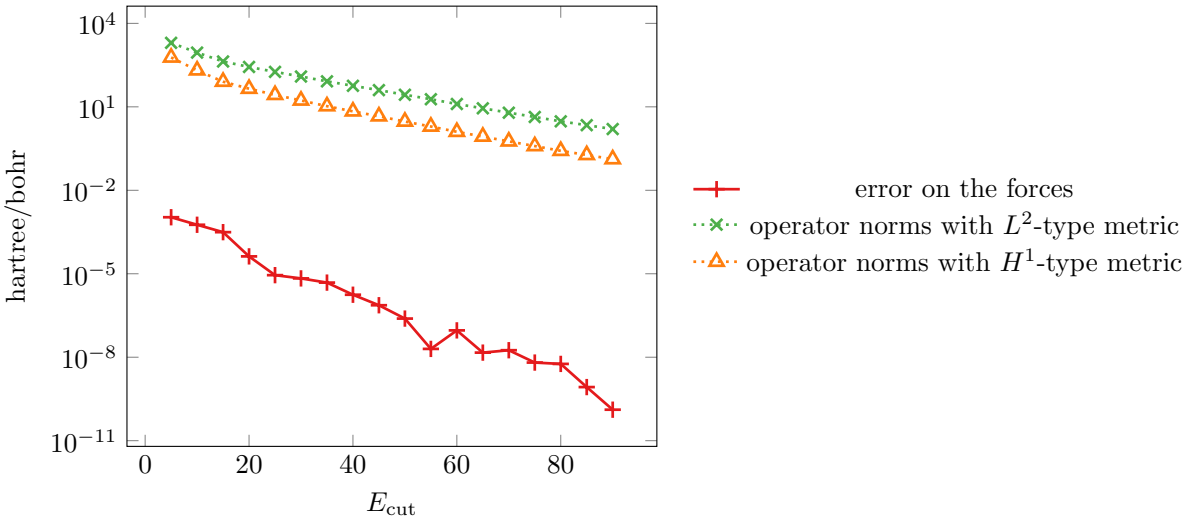



FIGURE 3.4 – Combination of the error estimate (3.4.8) on the interatomic forces with the error estimate on the error in $\mathrm{L}^2$-type norm (3.4.5) and $\mathrm{H}^1$-type norm (3.4.6). The inaccuracy of the bounds accumulates and results in extremely inaccurate bounds, from six to eleven orders of magnitude.

## 3.5 Improved error bounds based on frequencies splitting

### 3.5.1 Spectral decomposition of the error

In the previous section, we saw that even if $\mathbf{\Pi}_P(P-P_*)$ and $\boldsymbol{M}^{-1}R(P)$ are asymptotically equivalent in suitable norms, replacing the former by the latter in (3.4.7) when $A = F$ (interatomic forces) results in a large error, even in the asymptotic regime.

To analyse this issue, we use the decomposition

$$\mathcal{X}_{E_{\mathrm{cut,ref}}} = \mathcal{X}_{E_{\mathrm{cut}}} \oplus \mathcal{X}_{E_{\mathrm{cut}}}^{\perp}. \tag{3.5.1}$$

Since $\mathcal{X}_{E_{\mathrm{cut}}} = \mathrm{Span}\,(e_G,\ \frac{|G|^2}{2} \leqslant E_{\mathrm{cut}})$ and $\mathcal{X}_{E_{\mathrm{cut}}}^{\perp} = \mathrm{Span}\,(e_G,\ E_{\mathrm{cut}} < \frac{|G|^2}{2} \leqslant E_{\mathrm{cut,ref}})$, (3.5.1) corresponds to a low *vs* high frequency splitting. Using the identification of $\mathcal{X}_{E_{\mathrm{cut,ref}}} \equiv \mathbb{C}^{\mathcal{N}}$ introduced in Section 3.3.2, (3.5.1) boils down to decomposing $\mathbb{C}^{\mathcal{N}}$ as

$$\mathbb{C}^{\mathcal{N}} = \mathcal{X} \oplus \mathcal{X}^{\perp} \quad \text{with} \quad \mathcal{X} = \begin{pmatrix} \mathbb{C}^{N_{\mathrm{b}}} \\ 0_{\mathbb{C}^{\mathcal{N}-N_{\mathrm{b}}}} \end{pmatrix} \quad \text{and} \quad \mathcal{X}^{\perp} = \begin{pmatrix} 0_{\mathbb{C}^{N_{\mathrm{b}}}} \\ \mathbb{C}^{\mathcal{N}-N_{\mathrm{b}}} \end{pmatrix}.$$

Let $\Phi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}$ be such that $\Phi^*\Phi = I_{N_{\mathrm{el}}}$ and $P = \Phi\Phi^* \in \mathcal{M}_{\mathcal{N}}$. Combining the identification $\mathcal{X}_{E_{\mathrm{cut,ref}}} \equiv \mathbb{C}^{\mathcal{N}}$ described above with the relation (3.2.9) identifying a matrix $X$ of the tangent space $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ with a collection $\Xi = (\xi_1|\cdots|\xi_{N_{\mathrm{el}}}) \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}$ of orbital variations such that $\Phi^*\Xi = 0$, the decomposition (3.5.1) induces a decomposition of the tangent space $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ into two orthogonal subspaces $\mathbf{\Pi}_{E_{\mathrm{cut}}}\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ and $\mathbf{\Pi}_{E_{\mathrm{cut}}}^{\perp}\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ (for the Frobenius inner product):

$$\mathbf{\Pi}_{E_{\mathrm{cut}}}\left(\sum_{i=1}^{N_{\mathrm{el}}}|\phi_i\rangle\langle\xi_i| + |\xi_i\rangle\langle\phi_i|\right) \coloneqq \sum_{i=1}^{N_{\mathrm{el}}}|\phi_i\rangle\langle\Pi_{\mathcal{X}}\xi_i| + |\Pi_{\mathcal{X}}\xi_i\rangle\langle\phi_i|,$$
$$\mathbf{\Pi}_{E_{\mathrm{cut}}}^{\perp}\left(\sum_{i=1}^{N_{\mathrm{el}}}|\phi_i\rangle\langle\xi_i| + |\xi_i\rangle\langle\phi_i|\right) \coloneqq \sum_{i=1}^{N_{\mathrm{el}}}|\phi_i\rangle\langle\Pi_{\mathcal{X}}^{\perp}\xi_i| + |\Pi_{\mathcal{X}}^{\perp}\xi_i\rangle\langle\phi_i|,$$

where $\Pi_{\mathcal{X}}$ is the orthogonal projector on $\mathcal{X}$ (for the canonical inner product of $\mathbb{C}^{\mathcal{N}}$) and $\Pi_{\mathcal{X}}^{\perp} = 1 - \Pi_{\mathcal{X}}$. If $P$ solves the minimization problem (3.2.2), we infer from the first-order optimality conditions that the residual $R(P)$ is orthogonal to $\mathbf{\Pi}_{E_{\mathrm{cut}}}\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$, meaning that the vectors $r_i(P)$ such that

$$R(P) = \sum_{i=1}^{N_{\mathrm{el}}}|\phi_i\rangle\langle r_i(P)| + |r_i(P)\rangle\langle\phi_i|$$

belong to $\mathcal{X}^{\perp}$. Note that in practice, this is not exactly true for the full Kohn–Sham model because of the numerical quadrature errors involved in the treatment of the exchange-correlation terms.

Now $P-P_* \approx ((\mathbf{\Omega}(P)+\boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_{\mathcal{N}}})^{-1}R(P)$ contains two components: one in $\mathbf{\Pi}_{E_{\mathrm{cut}}}\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ and one in $\mathbf{\Pi}_{E_{\mathrm{cut}}}^{\perp}\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$. In the high-frequency subspace $\mathbf{\Pi}_{E_{\mathrm{cut}}}^{\perp}\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$, the leading term in $(\mathbf{\Omega}(P)+\boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_{\mathcal{N}}}$ comes from the contribution of the Laplacian arising in the Hamiltonian $h_0$, which is well approximated by the super-operator $\boldsymbol{M}$. This claim is supported by Proposition 3.1, in which we prove in a simplified setting that $((\mathbf{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_{\mathcal{N}}})^{-1}\mathbf{\Pi}_{E_{\mathrm{cut}}}^{\perp}$ is asymptotically equivalent to $\boldsymbol{M}^{-1}\mathbf{\Pi}_{E_{\mathrm{cut}}}^{\perp}$.

This is what we observe in Figure 3.5 (central and right panels): if $P$ is the solution to (3.2.2), the residual $R(P)$ is supported in $\mathbf{\Pi}_{E_{\mathrm{cut}}}^{\perp}\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ (up to numerical quadrature errors). In accordance with Proposition 3.1, the difference between the error $P - P_* \approx \mathbf{\Pi}_P(P - P_*)$ and the preconditioned residual $\boldsymbol{M}^{-1}R(P)$ (Figure 3.6) is smaller in Frobenius norm than the preconditioned residual itself. This explains our observations in Section 3.4.2 that $\|P - P_*\|_{\mathrm{F}}$ is well approximated by $\left\|\boldsymbol{M}^{-1}R(P)\right\|_{\mathrm{F}}$. However, this does not imply that $\mathrm{d}F(P)\cdot\mathbf{\Pi}_P(P - P_*)$ is well-approximated by $\mathrm{d}F(P)\cdot(\boldsymbol{M}^{-1}R(P))$. This is because the gradients $\nabla F_{j,\alpha}(P)$ are mostly supported on low frequencies, as illustrated in Figure 3.5 (left panel). Although the low-frequency contribution to the error $\mathbf{\Pi}_P(P - P_*)$ is of smaller magnitude than the high-frequency contribution, its contribution to $\mathrm{d}F(P)\cdot(\mathbf{\Pi}_P(P - P_*))$ is very significant. The fact that the low-frequency error is not captured at all by the purely high-frequency term $\boldsymbol{M}^{-1}R(P)$ is responsible for the poor approximation of the error $F(P) - F_*$ by $\mathrm{d}F(P)\cdot(\boldsymbol{M}^{-1}R(P))$.

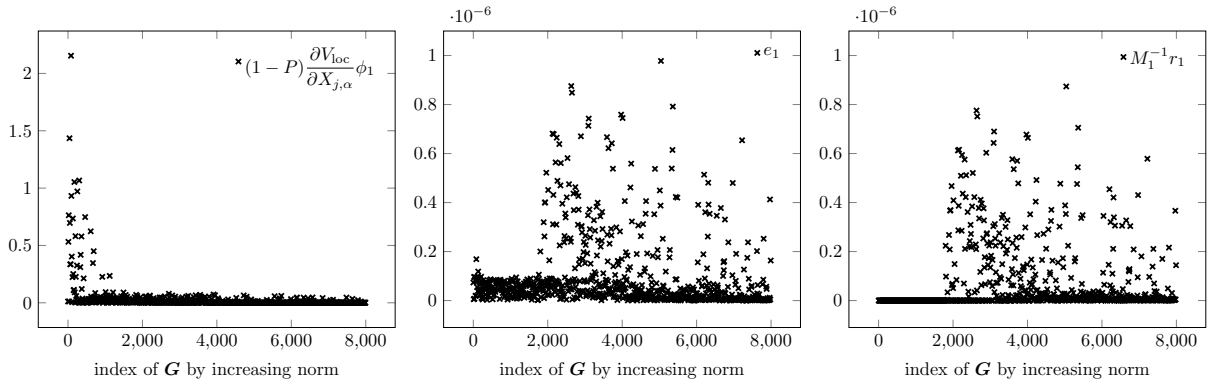



FIGURE 3.5 – Fourier coefficients moduli in the orbital representation $\boldsymbol{\Pi}_P(P-P_*) \simeq_\Phi (e_i)_{1\leqslant i\leqslant N}$ and $\boldsymbol{M}^{-1}R(P) \simeq_\Phi (M_i^{-1}r_i)_{1\leqslant i\leqslant N}$. (Left) Test function $(1-P)\frac{\partial V_{\text{loc}}}{\partial X_{j,\alpha}}\phi_1$ (see (3.4.10)).

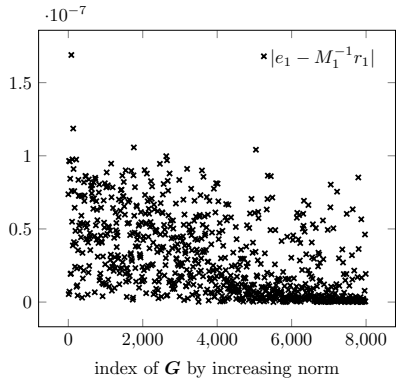



FIGURE 3.6 – Fourier coefficients of the difference between the error $e_1$ and the preconditioned residual $M_1^{-1}r_1$, where $\boldsymbol{\Pi}_P(P-P_*) \simeq_\Phi (e_i)_{1\leqslant i\leqslant N}$ and $\boldsymbol{M}^{-1}R(P) \simeq_\Phi (M_i^{-1}r_i)_{1\leqslant i\leqslant N}$. Low frequencies contribute greatly.

Now that we have understood the reason why it is not possible to approximate the error $F(P)-F_*$ on the interatomic forces by the computable term $\mathrm{d}F(P)\cdot(\boldsymbol{M}^{-1}R(P))$, we propose in the next section a way to evaluate this error, based on the linearization (3.4.3) and the frequencies splitting we just introduced.

### 3.5.2 Improving the error estimation

We now decompose tangent vectors and operators according to the splitting $\boldsymbol{\Pi}_{E_{\text{cut}}}\mathcal{T}_P\mathcal{M}_\mathcal{N}$ and $\boldsymbol{\Pi}^\perp_{E_{\text{cut}}}\mathcal{T}_P\mathcal{M}_\mathcal{N}$, which we respectively label by 1 and 2 for simplicity. In this way, the error-residual relationship can be written in concise form with obvious notation as

$$\begin{bmatrix}(\boldsymbol{\Omega}+\boldsymbol{K})_{11} & (\boldsymbol{\Omega}+\boldsymbol{K})_{12}\\ (\boldsymbol{\Omega}+\boldsymbol{K})_{21} & (\boldsymbol{\Omega}+\boldsymbol{K})_{22}\end{bmatrix}\begin{bmatrix}P_1-P_{*1}\\ P_2-P_{*2}\end{bmatrix} = \begin{bmatrix}R_1\\ R_2\end{bmatrix}.$$

Recall that $(\boldsymbol{\Omega}(P)+\boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}}$ is only invertible at high cost as it has a priori nonzero values on the four components of the operator arising from the low frequencies/high frequencies splitting of the operator. The computational cost for the inversion is equivalent to performing a Newton step on the reference grid. But we can make approximations to invert it only on the coarse grid $\mathcal{X}_{E_{\text{cut}}}$ and approximate the low frequency error components. In the same spirit as for the perturbation theory based post-processing method introduced in [38, 69] and the Feshbach–Schur method analysed in [71], we make the following approximations:

$$(\boldsymbol{\Omega}+\boldsymbol{K})_{21} \approx 0 \quad \text{and} \quad (\boldsymbol{\Omega}+\boldsymbol{K})_{22} \approx \boldsymbol{M}_{22},$$

which yields

$$\begin{bmatrix}(\boldsymbol{\Omega}+\boldsymbol{K})_{11} & (\boldsymbol{\Omega}+\boldsymbol{K})_{12}\\ 0 & \boldsymbol{M}_{22}\end{bmatrix}\begin{bmatrix}P_1-P_{*1}\\ P_2-P_{*2}\end{bmatrix} = \begin{bmatrix}R_1\\ R_2\end{bmatrix}$$

and therefore

$$P_2 - P_{*2} \approx \boldsymbol{M}_{22}^{-1} R_2, \tag{3.5.2}$$

$$P_1 - P_{*1} \approx (\boldsymbol{\Omega} + \boldsymbol{K})_{11}^{-1} (R_1 - (\boldsymbol{\Omega} + \boldsymbol{K})_{12} \boldsymbol{M}_{22}^{-1} R_2). \tag{3.5.3}$$

This requires only a single inexpensive computation on the fine grid. The main bottleneck is then to solve a linear system with operator $(\boldsymbol{\Omega} + \boldsymbol{K})_{11}$, which is as expensive as a full Newton step on the coarse grid $\mathcal{X}_{E_{\text{cut}}}$. Since $R_1 = 0$ when $P$ is the optimal Galerkin solution on $\mathcal{X}_{E_{\text{cut}}}$, we can understand the previous attempt to replace $P - P_*$ by $\boldsymbol{M}^{-1} R(P)$ as (3.5.2). Not neglecting $(\boldsymbol{\Omega} + \boldsymbol{K})_{12}$ in (3.5.3) gives rise to a correction on the coarse space also. We denote by $R_{\text{Schur}}(P)$ the new residual

$$R_{\text{Schur}}(P) = \begin{bmatrix} (\boldsymbol{\Omega} + \boldsymbol{K})_{11}^{-1} (R_1 - (\boldsymbol{\Omega} + \boldsymbol{K})_{12} \boldsymbol{M}_{22}^{-1} R_2) \\ \boldsymbol{M}_{22}^{-1} R_2 \end{bmatrix}.$$

To illustrate the validity of these approximations, we plotted in Figure 3.7 the components of $r_{\text{Schur}}$, the orbital representation of $R_{\text{Schur}}$. We see that this time, the error is well approximated by (3.5.3) in the low-frequency space.

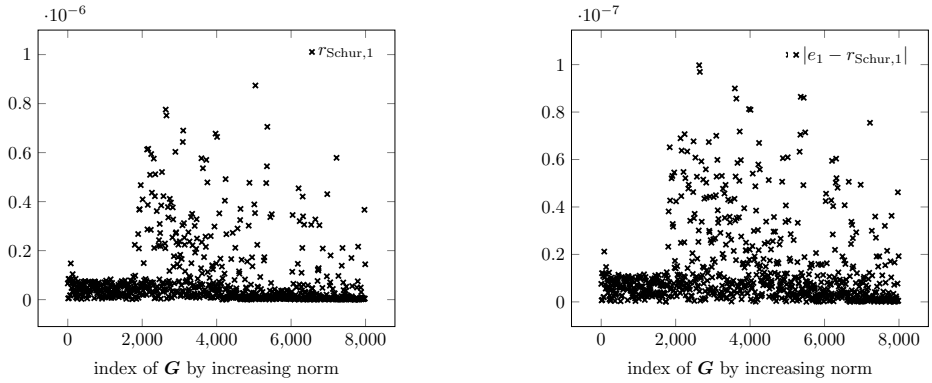



FIGURE 3.7 – Fourier coefficients of the new residual $r_{\text{Schur},1}$ and its comparison to the error $e_1$, where $\boldsymbol{\Pi}_P(P - P_*) \simeq_\Phi (e_i)_{1 \leqslant i \leqslant N}$ and $R_{\text{Schur}}(P) \simeq_\Phi (r_{\text{Schur},i})_{1 \leqslant i \leqslant N}$. (Left) Components of the modified residual $r_{\text{Schur},1}$. (Right) Difference between the error and the new residual: low frequencies are better approximated (compare with Figure 3.6).

In Figure 3.8, we plot the new estimate $\mathrm{d}F(P) \cdot (R_{\text{Schur}}(P))$ of the error $F(P) - F_*$ as well as the differences

$$\begin{aligned} F_{\text{err}} - F_* &\coloneqq F(P) - \mathrm{d}F(P) \cdot (\boldsymbol{\Pi}_P(P - P_*)) - F_*, \\ F_{\text{res}} - F_* &\coloneqq F(P) - \mathrm{d}F(P) \cdot (\boldsymbol{M}^{-1} R(P)) - F_*, \\ F_{\text{Schur}} - F_* &\coloneqq F(P) - \mathrm{d}F(P) \cdot (R_{\text{Schur}}(P)) - F_*, \end{aligned}$$

in order to have a better estimation of the improvement on the estimation of the error. With the Schur complement method, the new estimate better matches the error than the crude one simply using the residual: the accuracy of the estimation is approximately improved by one order of magnitude.

**Remark 3.4.** The quantity $\mathrm{d}F(P) \cdot (R_{\text{Schur}}(P))$ does not yield a guaranteed estimator of the error on the forces as it is obtained after several approximations and is only valid in the asymptotic regime. However, it can be computed for a cost comparable to the one of performing a SCF step on the same grid and can be used for two main purposes:

- as an error bound, as the error $F(P) - F_*$ is reasonably well approximated by $\mathrm{d}F(P) \cdot (R_{\text{Schur}}(P))$;
- as a more precise approximation of the QoI, as the forces $F_j(P)$ on atom $j$ obtained by a variational approximation on a coarse grid are improved by the post-processing $F_j(P) \mapsto F_j(P) - \mathrm{d}F_j(P) \cdot (R_{\text{Schur}}(P))$.

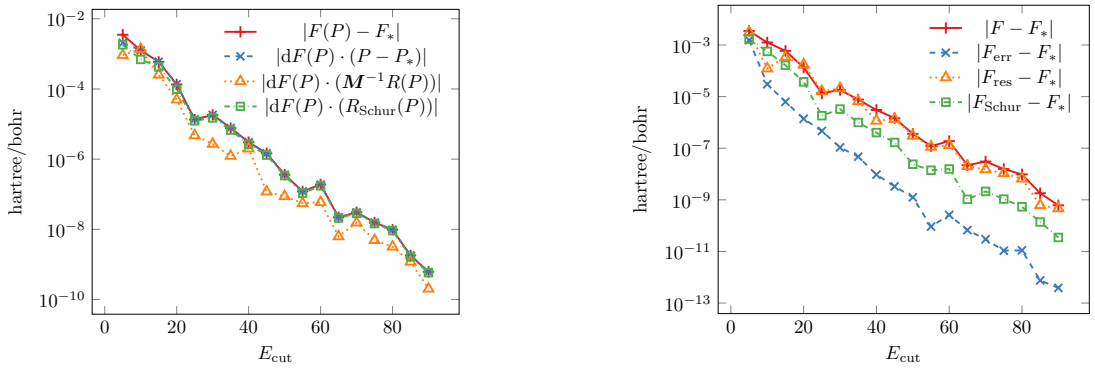



FIGURE 3.8 – (Left) Estimation of the error $F(P) - F_*$ with $\mathrm{d}F(P) \cdot X$ where $X$ is either the exact error $\boldsymbol{\Pi}_P(P - P_*)$, the preconditioned residual $\boldsymbol{M}^{-1}R(P)$ or the modified residual $R_{\text{Schur}}(P)$. (Right) Enhancement of the estimation of the forces by replacing $F(P)$ with $F(P) - \mathrm{d}F(P) \cdot X$ where $X$ is either the exact error $\boldsymbol{\Pi}_P(P - P_*)$, the preconditioned residual $\boldsymbol{M}^{-1}R(P)$ or the modified residual $R_{\text{Schur}}(P)$.

## 3.6 Numerical examples with more complex systems

We perform the same simulations as for silicon, but for more complex systems, namely GaAs and $TiO_2$. The calculations are still performed within the LDA approximation with GTH pseudopotentials and Teter 93 exchange-correlation functional, with a $2 \times 2 \times 2$ $k$-point grid to discretize the Brillouin zone, and the reference solutions are obtained for $E_{\text{cut,ref}} = 125$ Ha. We describe here the numerical setting for both systems.

**GaAs** We use the usual periodic lattice for the FCC phase of GaAs, with lattice constant 10.68 bohrs, close to but not exactly at the equilibrium configuration in order to get nonzero forces. The Ga atom is placed at fractional coordinates $(\frac{1}{8}, \frac{1}{8}, \frac{1}{8})$ and the As atom at fractional coordinates $(-\frac{1}{8}, -\frac{1}{8}, -\frac{1}{8})$. The Ga atom is then displaced by $\frac{1}{15}(0.24, -0.33, 0.12)$ to get nonzero forces. In this setting, the reference values for the energy is $E_* = -8.572$ Ha and the interatomic forces are, in hartree/bohr,

$$F_* = \begin{bmatrix} -0.0448 & 0.0448 \\ 0.0722 & -0.0722 \\ -0.0251 & 0.0251 \end{bmatrix},$$

where the first column are the forces acting on the Ga atom in each direction, and the second column are the forces acting on the As atom.

**$TiO_2$** We use the MP-2657 configuration in the primitive cell from the Materials Project [162]. We apply the small displacement $\frac{1}{5}(0.22, -0.28, 0.35)$ to the equilibrium position of the first Ti atom to get nonzero forces. In this setting, the reference values for the energy is $E_* = -71.589$ Ha and the interatomic forces are, in hartree/bohr,

$$F_* = \begin{bmatrix} -2.88 & 0.641 & 3.80 & 0.753 & -1.57 & -0.745 \\ 3.10 & -0.919 & -3.09 & -1.45 & 0.800 & 1.56 \\ 0.136 & 0.403 & -0.368 & -0.786 & 0.251 & 0.364 \end{bmatrix},$$

where the first two columns are the forces acting on the two Ti atoms in each direction, and the other columns are the forces acting on the four O atoms.

We plot in Figure 3.9 the energy, density and forces obtained after a Newton step on the fine grid starting from the variational solution on the coarse grid given by $E_{\text{cut}}$, for GaAs and $TiO_2$. The fast establishment of the asymptotic regime is confirmed for the two new systems as, even for small $E_{\text{cut}}$'s, the so-obtained QoIs are orders of magnitude more accurate than the ones obtained by the variational solution on the coarse grid.

We plot in Figure 3.10 the estimation of the actual error $F(P) - F_*$ with $\mathrm{d}F(P) \cdot X$ where $X$ is either $\boldsymbol{\Pi}_P(P - P_*)$, $R(P)$ or $R_{\text{Schur}}(P)$. In Figure 3.11, we plot the improvement of the estimation of the forces $F(P) - \mathrm{d}F(P) \cdot X$ where $X$ is either $\boldsymbol{\Pi}_P(P - P_*)$, $R(P)$ or $R_{\text{Schur}}(P)$. Just as for silicon, the estimation

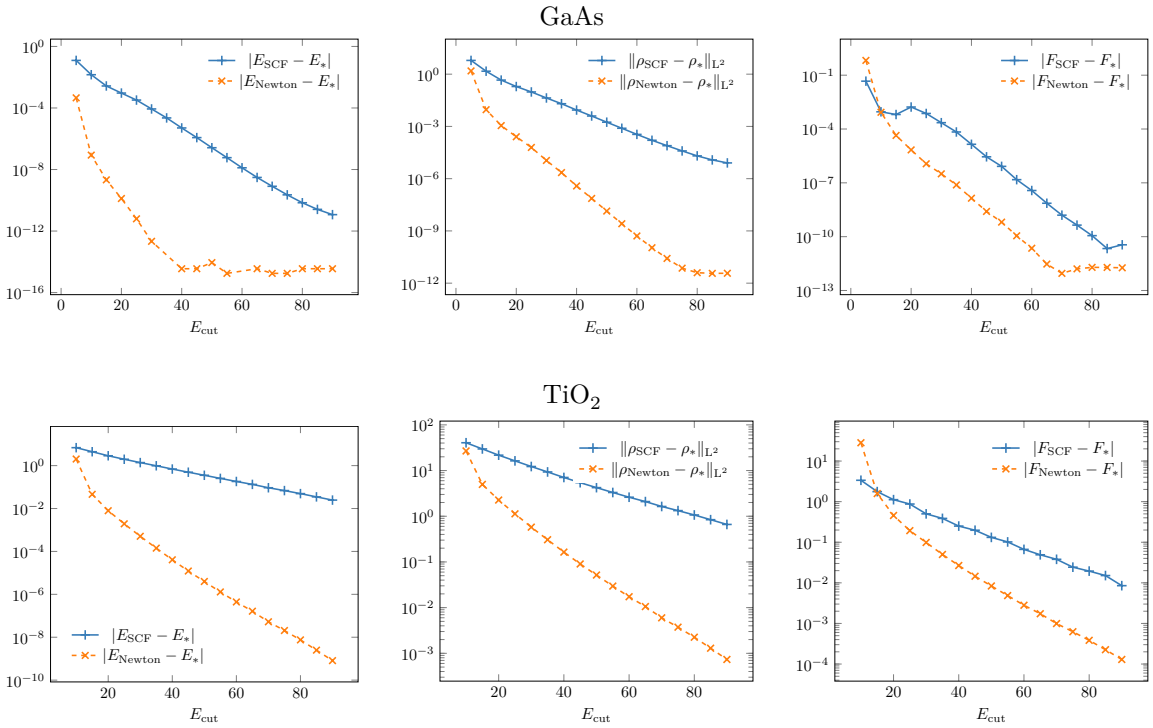



FIGURE 3.9 – Errors of some QoI as functions of $E_{\mathrm{cut}}$ (reference solution is obtained with $E_{\mathrm{cut,ref}} = 125$ Ha) for GaAS and $TiO_2$. Solid lines: errors obtained with the variational solution in the space $\mathcal{X}_{E_{\mathrm{cut}}}$. Dashed lines: errors obtained with one Newton step on the reference grid, starting from the variational solution in the space $\mathcal{X}_{E_{\mathrm{cut}}}$. Left panel: energy (hartree), central panel: discrete $L^2$ norm of the density, right panel: interatomic forces (hartree/bohr). To be compared with Figure 3.1.

is well improved with the modified residual $R_{\mathrm{Schur}}$. Note that in the GaAs case, there is a plateau for high $E_{\mathrm{cut}}$'s. This phenomenon is explained in the remark below.

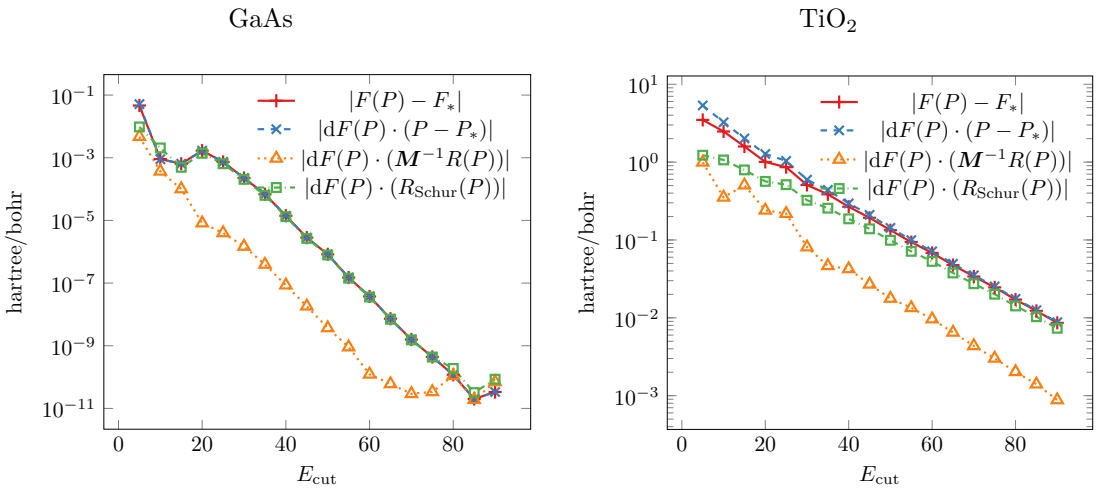



FIGURE 3.10 – Estimation of the error $F(P) - F_*$ with $\mathrm{d}F(P)\cdot X$ where $X$ is either the exact error $\mathbf{\Pi}_P(P - P_*)$, the preconditioned residual $\boldsymbol{M}^{-1}R(P)$ or the modified residual $R_{\mathrm{Schur}}(P)$. To be compared with Figure 3.8 (Left).

**Remark 3.5.** The plateau observed Figure 3.10 and Figure 3.11 for GaAs comes from the numerical quadrature scheme used to compute the exchange-correlation energy and the corresponding matrix elements. In fact, we also observed such plateaus for silicon and $TiO_2$ with the default quadrature scheme of DFTK, but these disappeared by using 8 times as many numerical quadrature points. With this more accurate numerical quadrature scheme, the plateau for GaAs is lower but still visible. It disappears when further increasing the number of quadrature points, at the price of longer computations.

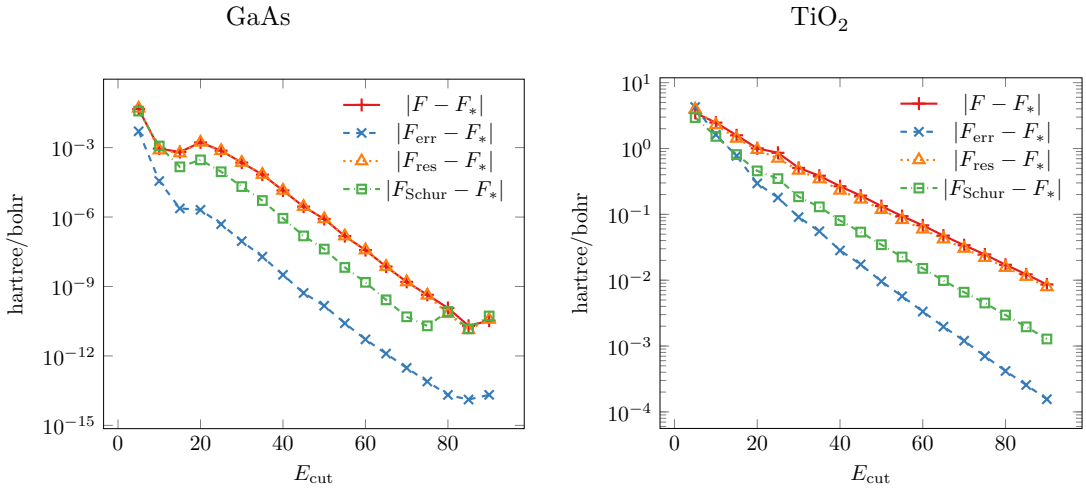



FIGURE 3.11 – Enhancement of the estimation of the forces by replacing $F(P)$ with $F(P) - \mathrm{d}F(P) \cdot X$ where $X$ is either the exact error $\mathbf{\Pi}_P(P - P_*)$, the preconditioned residual $\boldsymbol{M}^{-1}R(P)$ or the modified residual $R_{\mathrm{Schur}}(P)$. To be compared with Figure 3.8 (Right).

## 3.7 Conclusion

In this work, we have investigated methods to estimate the error on interatomic forces resulting from plane-wave discretizations of the Kohn–Sham equations. On the systems we investigated, we find the following:

- Linearizing the equations around a solution is a good approximation, even for energy cut-offs as small as 5 hartree (Figure 3.1).
- The naive approach based on the computation of operator norms proves to be extremely inefficient, overestimating the error by several orders of magnitude. This is the case even when using appropriate $H^1$-type norms (Figure 3.4). The reason is that the discretization error is mostly made up of high frequency components, whereas quantities of interest are mostly supported on low frequencies, resulting in very suboptimal Cauchy–Schwarz inequalities (Figure 3.5).
- Replacing directly the error by the preconditioned residual yields reasonable estimates of the errors, but they are not systematic upper bounds (Figure 3.3).
- A Schur approach based on a low/high frequency splitting systematically improves the solution and gives reliable estimates of the error (Figure 3.8), at the price of more computational work.

Our results validate on realistic test cases and for properties such as interatomic forces the frequency splitting approach already introduced in [38, 69, 71]. Thanks to the modular nature of DFTK and the use of automatic differentiation, the implementation of our estimates is relatively simple and convenient. It is publicly available at `https://github.com/gkemlin/paper-forces-estimator`. The algorithm proceeds in two steps: i) the computation of the residual on the fine grid and ii) a linear system solve involving the Jacobian on the coarse grid. The computational cost of step i) is negligible compared to that of step ii), which is roughly that of a full self-consistent computation on the coarse grid. Therefore, for roughly twice the cost of a standard computation, one obtains an accurate approximation of the discretization error on the interatomic forces (or, equivalently, a better estimate of the latter).

The scope of this work is limited to gapped systems at zero temperature and to the study of the discretization error. Interesting perspectives for future work include the application of this methodology to the error resulting from an incomplete self-consistent cycle, and to finite-temperature models, including metals (see [99] for an extension of the linearized equations to the finite-temperature case).

# Appendix: Mathematical justification

The purpose of this appendix is to explain mathematically in a simplified setting the observation in Section 3.4.2 that $\left\|\boldsymbol{M}^{-1/2}\boldsymbol{\Pi}_P R(P)\right\|_{\mathrm{F}}$ was a good approximation of $\left\|\boldsymbol{M}^{1/2}\boldsymbol{\Pi}_P(P-P_*)\right\|_{\mathrm{F}}$. For this purpose, we work in a slightly different framework than the one we used in the rest of the chapter, and consider the infinite-dimensional version of Problem (3.2.1) associated with the periodic Gross–Pitaevskii model in dimension $d \leqslant 3$, which reads as

$$E_* \coloneqq \min\{E(P),\ P \in \mathcal{M}_\infty\}, \tag{3.7.1}$$

with $\mathcal{M}_\infty \coloneqq \{P \in \mathcal{S}(\mathrm{L}^2_\#)\,|\,P^2 = P,\ \mathrm{Tr}(P) = 1,\ \mathrm{Ran}\,(P) \subset \mathrm{H}^1_\#\}$ and $E(P) \coloneqq \mathrm{Tr}((-\Delta + V)P) + \frac{1}{2}\int_\Gamma \rho_P^2$. Here $\mathcal{S}(\mathrm{L}^2_\#)$ denotes the space of self-adjoint operators on $\mathrm{L}^2_\#$, $V$ a given function of $\mathrm{L}^\infty_\#$, and $\rho_P$ the density of $P$. The condition $\mathrm{Ran}\,(P) \subset \mathrm{H}^1_\#$ ensures that both the linear and nonlinear terms in the energy functional $E(P)$ are well-defined and finite. It is convenient to rewrite (3.7.1) in the orbital framework. Any state $P \in \mathcal{M}_\infty$ is rank-1 and such that $\mathrm{Ran}\,(P) \subset \mathrm{H}^1_\#$. It can therefore be represented by a function $\phi \in \mathrm{H}^1_\#$ such that $\|\phi\|_{\mathrm{L}^2_\#} = 1$ through the relation $P = |\phi\rangle\langle\phi|$ (using Dirac's notation). The orbital formulation of problem (3.7.1) reads

$$E_* \coloneqq \min\{\mathcal{E}^{\mathrm{GP}}(\phi),\ \phi \in \mathrm{H}^1_\#,\ \|\phi\|_{\mathrm{L}^2_\#} = 1\}, \tag{3.7.2}$$

with $\mathcal{E}^{\mathrm{GP}}(\phi) \coloneqq \int_\Gamma |\nabla\phi|^2 + \int_\Gamma V|\phi|^2 + \frac{1}{2}\int_\Gamma |\phi|^4$. It is well-known (see *e.g.* the Appendix of [30]) that the minimizer of (3.7.1) is unique, and that the set of solutions of (3.7.2) is $(e^{i\alpha}\phi_*)_{\alpha\in\mathbb{R}}$, where $(\lambda_*, \phi_*) \in \mathbb{R} \times \mathrm{H}^1_\#$ is the unique solution to

$$\begin{cases} -\Delta\phi_* + V\phi_* + \phi_*^3 = \lambda_*\phi_*, \\ \|\phi_*\|_{\mathrm{L}^2_\#} = 1, \quad \phi_* > 0 \text{ on } \mathbb{R}^d. \end{cases} \tag{3.7.3}$$

We consider the variational approximation of (3.7.1) in the finite dimensional space

$$\mathcal{X}_N \coloneqq \mathrm{Span}\,(e_G,\ |G|^2/2 \leqslant N)$$

corresponding to a plane-wave discretization with energy cut-off $E_{\mathrm{cut}} = N$. We denote by $\Pi_N$ the $\mathrm{L}^2_\#$-orthogonal projector on $\mathcal{X}_N$ and by $\Pi_N^\perp \coloneqq 1 - \Pi_N$. For $N$ large enough, the approximate ground-state $P_N$ is unique and can be represented by a unique function $\phi_N$ real-valued and positive on $\mathbb{R}^3$ (see [30]), and it holds

$$\begin{cases} -\Delta\phi_N + \Pi_N\big(V\phi_N - \phi_N^3\big) = \lambda_N\phi_N, \\ \|\phi_N\|_{\mathrm{L}^2_\#} = 1, \end{cases}$$

for some uniquely defined $\lambda_N \in \mathbb{R}$. In addition, we have $\phi_* \in \mathrm{H}^2_\#$ and

$$\|\phi_N - \phi_*\|_{\mathrm{H}^2_\#} \underset{N\to\infty}{\longrightarrow} 0 \quad \text{and} \quad |\lambda_N - \lambda_*| \underset{N\to\infty}{\longrightarrow} 0. \tag{3.7.4}$$

Using similar notation as the one used in the rest of the chapter, we introduce the following quantities:

- $\Pi^\perp_{\phi_N}$ is the orthogonal projector (for the $\mathrm{L}^2_\#$ inner product) onto $\phi_N^\perp$;
- $A_N$ is the self-adjoint operator on $\phi_N^\perp$ defined by

$$A_N \coloneqq (\Omega_N + K_N) \tag{3.7.5}$$

  where $\Omega_N$ and $K_N$ represent, in the orbital framework, the super-operators $\boldsymbol{\Omega}(P_N)|_{\mathcal{T}_{P_N}\mathcal{M}_\infty}$ and $\boldsymbol{K}(P_N)|_{\mathcal{T}_{P_N}\mathcal{M}_\infty}$. We have

$$\forall\ \psi_N \in \phi_N^\perp, \quad \Omega_N\psi_N = \Pi^\perp_{\phi_N}\big(-\Delta + V + \phi_N^2 - \lambda_N\big)\psi_N, \tag{3.7.6}$$

$$\forall\ \psi_N \in \phi_N^\perp, \quad K_N\psi_N = \Pi^\perp_{\phi_N}\big(2\phi_N^2\psi_N\big); \tag{3.7.7}$$

- $M_N^{1/2}$ is the restriction of the operator $\Pi^\perp_{\phi_N}(1-\Delta)^{1/2}\Pi^\perp_{\phi_N}$ to the invariant subspace $\phi_N^\perp$.

We then have the following result, which justifies in this case the claim made in Section 3.4.2 that $\boldsymbol{M}^{-1/2}\boldsymbol{M}^{1/2}((\boldsymbol{\Omega}(P)+\boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}})^{-1}\boldsymbol{M}^{1/2}$ is close to identity on the subspace of high-frequency Fourier modes. It also justifies that $\boldsymbol{\Pi}_P(P-P_*) \approx \boldsymbol{M}^{-1}R(P)$ as $\boldsymbol{M}^{-1/2}$ is a uniformly bounded operator and $\boldsymbol{M}^{-1/2}R(P)$ is high-frequency:

$$\begin{aligned}\boldsymbol{\Pi}_P(P-P_*) &\approx \boldsymbol{M}^{-1/2}\boldsymbol{M}^{1/2}((\boldsymbol{\Omega}(P)+\boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}})^{-1}\boldsymbol{M}^{1/2}\boldsymbol{M}^{-1/2}R(P)\\ &\approx \boldsymbol{M}^{-1/2}\boldsymbol{M}^{-1/2}R(P) = \boldsymbol{M}^{-1}R(P).\end{aligned}$$

**Proposition 3.1.** *We have*

$$\lim_{N\to\infty}\left\|M_N^{1/2}(\Omega_N+K_N)^{-1}M_N^{1/2} - I_{\mathcal{X}_N^\perp}\right\|_{\mathcal{X}_N^\perp\to \mathrm{L}^2_\#} = 0.$$

*Proof.* Let $W_N \coloneqq V + 3\phi_N^2 - \lambda_N - 1$ and $W_* \coloneqq V + 3\phi_*^2 - \lambda_* - 1$. In view of (3.7.4), $W_N$ converges to $W_*$ in $\mathrm{L}^\infty_\#$ when $N$ goes to infinity. It also follows from [30, Lemma 1] that the self-adjoint operator

$$\begin{aligned}\widetilde{A}_* &\coloneqq -\Delta + V + 3\phi_*^2 - \lambda_* = (1-\Delta) + W_*\\ &= (1-\Delta)^{1/2}\Big(1 + (1-\Delta)^{-1/2}W_*(1-\Delta)^{-1/2}\Big)(1-\Delta)^{1/2}\end{aligned}$$

is coercive, hence, by the Lax–Milgram lemma, defines a continuous isomorphism from $\mathrm{H}^1_\#$ to $\mathrm{H}^{-1}_\#$. We denote by $\widetilde{A}_*^{-1}$ its inverse, seen as a bounded operator from $\mathrm{H}^{-1}_\#$ to $\mathrm{H}^1_\#$, so that $B_* \coloneqq (1-\Delta)^{1/2}\widetilde{A}_*^{-1}(1-\Delta)^{1/2}$ defines a bounded operator on $\mathrm{L}^2_\#$.

Using the convergence results (3.7.4) and standard perturbation theory it follows that for $N$ large enough, the operator $B_N \coloneqq (1-\Delta)^{1/2}\widetilde{A}_N^{-1}(1-\Delta)^{1/2}$, where $\widetilde{A}_N \coloneqq (1-\Delta)+W_N$, is bounded on $\mathrm{L}^2_\#$ uniformly in $N$, and that we have

$$B_N = \Big(1 + (1-\Delta)^{-1/2}W_N(1-\Delta)^{-1/2}\Big)^{-1} = 1 - B_N(1-\Delta)^{-1/2}W_N(1-\Delta)^{-1/2}. \tag{3.7.8}$$

We now compute the action of the operator $M_N^{1/2}A_N^{-1}M_N^{1/2} : \phi_N^\perp \to \phi_N^\perp$, relating it to $\widetilde{A}_N^{-1}$ and $B_N$, with $A_N$ defined in (3.7.5). Let $\xi_N \in X_N^\perp$. As $\phi_N \in X_N$, we have $X_N^\perp \subset \phi_N^\perp$ so that $\xi_N \in X_N^\perp \subset \phi_N^\perp$, and $M_N^{1/2}\xi_N = (1-\Delta)^{1/2}\xi_N \in X_N^\perp \subset \phi_N^\perp$ , where we used that $X_N$ and $X_N^\perp$ are invariant subspaces of the operator $(1-\Delta)^{1/2}$. Let $v_N \coloneqq A_N^{-1}M_N^{1/2}\xi_N = A_N^{-1}(1-\Delta)^{1/2}\xi_N \in \phi_N^\perp$ . Using (3.7.6) and (3.7.7), we get

$$\Pi_{\phi_N}^\perp\big(-\Delta + V + 3\phi_N^2 - \lambda_N\big)v_N = (1-\Delta)^{1/2}\xi_N, \quad \textit{i.e.} \quad \Pi_{\phi_N}^\perp\widetilde{A}_N v_N = (1-\Delta)^{1/2}\xi_N,$$

and therefore,

$$\widetilde{A}_N v_N = (1-\Delta)^{1/2}\xi_N + \alpha_N\phi_N,$$

where $\alpha_N = -\dfrac{\big\langle\phi_N,\widetilde{A}_N^{-1}(1-\Delta)^{1/2}\xi_N\big\rangle_{\mathrm{L}^2_\#}}{\big\langle\phi_N,\widetilde{A}_N^{-1}\phi_N\big\rangle_{\mathrm{L}^2_\#}} = -\dfrac{\big\langle(1-\Delta)^{-1/2}\phi_N,B_N\xi_N\big\rangle_{\mathrm{L}^2_\#}}{\big\langle\phi_N,\widetilde{A}_N^{-1}\phi_N\big\rangle_{\mathrm{L}^2_\#}} \in \mathbb{R}$ is characterized by the constraint $v_N \in \phi_N^\perp$. We thus obtain

$$v_N = \widetilde{A}_N^{-1}(1-\Delta)^{1/2}\xi_N - \frac{\big\langle(1-\Delta)^{-1/2}\phi_N,B_N\xi_N\big\rangle_{\mathrm{L}^2_\#}}{\big\langle\phi_N,\widetilde{A}_N^{-1}\phi_N\big\rangle_{\mathrm{L}^2_\#}}\widetilde{A}_N^{-1}\phi_N,$$

and therefore, as $v_N \in \phi_N^\perp$,

$$
\begin{aligned}
&M_N^{1/2} A_N^{-1} M_N^{1/2} \xi_N - \xi_N \\
&\quad = M_N^{1/2} v_N - \xi_N \\
&\quad = \Pi_{\phi_N}^\perp (1-\Delta)^{1/2} v_N - \xi_N \\
&\quad = (1-\Delta)^{1/2} v_N - \Big\langle \phi_N, (1-\Delta)^{1/2} v_N \Big\rangle_{\mathrm{L}^2_\#} \phi_N - \xi_N \\
&\quad = (B_N - 1)\xi_N - \langle \phi_N, B_N \xi_N \rangle_{\mathrm{L}^2_\#} \phi_N \\
&\qquad - \frac{\big\langle (1-\Delta)^{-1/2}\phi_N, B_N \xi_N \big\rangle_{\mathrm{L}^2_\#}}{\Big\langle \phi_N, \widetilde{A}_N^{-1} \phi_N \Big\rangle_{\mathrm{L}^2_\#}} \Big( (1-\Delta)^{1/2} \widetilde{A}_N^{-1} \phi_N - \Big\langle \phi_N, (1-\Delta)^{1/2} \widetilde{A}_N^{-1} \phi_N \Big\rangle_{\mathrm{L}^2_\#} \phi_N \Big) \\
&\quad = (B_N - 1)\xi_N - \langle \phi_N, B_N \xi_N \rangle_{\mathrm{L}^2_\#} \phi_N \\
&\qquad - \frac{\big\langle (1-\Delta)^{-1/2}\phi_N, B_N \xi_N \big\rangle_{\mathrm{L}^2_\#}}{\Big\langle \phi_N, \widetilde{A}_N^{-1} \phi_N \Big\rangle_{\mathrm{L}^2_\#}} \Big( B_N (1-\Delta)^{-1/2} \phi_N - \Big\langle \phi_N, B_N (1-\Delta)^{-1/2} \phi_N \Big\rangle_{\mathrm{L}^2_\#} \phi_N \Big) \\
&\quad = (B_N - 1)\xi_N - \langle \phi_N, (B_N - 1)\xi_N \rangle_{\mathrm{L}^2_\#} \phi_N \\
&\qquad - \frac{\big\langle (1-\Delta)^{-1/2}\phi_N, (B_N - 1) \xi_N \big\rangle_{\mathrm{L}^2_\#}}{\Big\langle \phi_N, \widetilde{A}_N^{-1} \phi_N \Big\rangle_{\mathrm{L}^2_\#}} \Big( B_N (1-\Delta)^{-1/2} \phi_N - \Big\langle \phi_N, B_N (1-\Delta)^{-1/2} \phi_N \Big\rangle_{\mathrm{L}^2_\#} \phi_N \Big),
\end{aligned}
$$

where we used the fact that $\xi_N \in \mathcal{X}_N^\perp$, while $\phi_N$ and $(1-\Delta)^{-1/2}\phi_N$ belong to $\mathcal{X}_N$. Using again (3.7.4) we obtain that for $N$ large enough,

$$
\begin{aligned}
\Big\| M_N^{1/2} A_N^{-1} M_N^{1/2} - I_{\mathcal{X}_N^\perp} \Big\|_{\mathcal{X}_N^\perp \to \mathrm{L}^2_\#} &= \sup_{\xi_N \in \mathcal{X}_N^\perp} \frac{\Big\| M_N^{1/2} A_N^{-1} M_N^{1/2} \xi_N - \xi_N \Big\|_{\mathrm{L}^2_\#}}{\|\xi_N\|_{\mathrm{L}^2_\#}} \\
&\leqslant C_* \big\| (B_N - 1)\Pi_N^\perp \big\|_{\mathrm{L}^2_\# \to \mathrm{L}^2_\#},
\end{aligned}
$$

where

$$
C_* := 3 + \frac{\|\phi_*\|_{\mathrm{H}^{-1}_\#}}{\Big\langle \phi_*, \widetilde{A}_*^{-1} \phi_* \Big\rangle_{\mathrm{L}^2_\#}} \times 2 \|B_*\|_{\mathrm{L}^2_\# \to \mathrm{L}^2_\#} \|\phi_*\|_{\mathrm{H}^{-1}_\#}.
$$

Finally, using (3.7.8), we have

$$
\begin{aligned}
\big\| (B_N - 1)\Pi_N^\perp \big\|_{\mathrm{L}^2_\# \to \mathrm{L}^2_\#} &\leqslant \Big\| B_N (1-\Delta)^{-1/2} W_N \Big\|_{\mathrm{L}^2_\# \to \mathrm{L}^2_\#} \Big\| (1-\Delta)^{-1/2} \Pi_N^\perp \Big\|_{\mathrm{L}^2_\# \to \mathrm{L}^2_\#} \\
&\leqslant \|B_N\|_{\mathrm{L}^2_\# \to \mathrm{L}^2_\#} \|W_N\|_{\mathrm{L}^\infty_\#} (1+2N)^{-1/2} \underset{N \to 0}{\longrightarrow} 0,
\end{aligned}
$$

since $\|B_N\|_{\mathrm{L}^2_\# \to \mathrm{L}^2_\#}$ and $\|W_N\|_{\mathrm{L}^\infty_\#}$ are uniformly bounded in $N$. This concludes the proof. □

Chapter 4

# Numerical stability and efficiency of response property calculations in density functional theory

This chapter coincides with [GKp1]:

*Eric Cancès, Michael F. Herbst, Gaspard Kemlin, Antoine Levitt and Benjamin Stamm. Numerical stability and efficiency of response property calculations in density functional theory. Submitted.* https://arxiv.org/abs/2210.04512.

**Abstract** Response calculations in density functional theory aim at computing the change in ground-state density induced by an external perturbation. At finite temperature these are usually performed by computing variations of orbitals, which involve the iterative solution of potentially badly-conditioned linear systems, the Sternheimer equations. Since many sets of variations of orbitals yield the same variation of density matrix this involves a choice of gauge. Taking a numerical analysis point of view we present the various gauge choices proposed in the literature in a common framework and study their stability. Beyond existing methods we propose a new approach, based on a Schur complement using extra orbitals from the self-consistent-field calculations, to improve the stability and efficiency of the iterative solution of Sternheimer equations. We show the success of this strategy on nontrivial examples of practical interest, such as Heusler transition metal alloy compounds, where savings of around 40% in the number of required cost-determining Hamiltonian applications have been achieved.

## Contents

**4.1 Introduction** . . . 88
**4.2 Mathematical framework** . . . 90
4.2.1 Periodic Kohn–Sham equations . . . 90
4.2.2 Density functional perturbation theory . . . 92
4.2.3 Plane-wave discretization and numerical resolution . . . 93
**4.3 Computing the response** . . . 93
4.3.1 Practical implementation . . . 93
4.3.2 Occupied-occupied contributions . . . 95
4.3.3 Computation of unoccupied-occupied contributions employing a Schur complement . . . 96
**4.4 Numerical tests** . . . 98
4.4.1 Insulators and semiconductors . . . 98
4.4.2 Metals . . . 99
4.4.3 Comparison to shifted Sternheimer approaches . . . 101
**4.5 Conclusion** . . . 102

## 4.1 Introduction

Kohn–Sham (KS) density functional theory (DFT) [104, 116] is the most popular approximation to the electronic many-body problem in quantum chemistry and materials science. It offers a favourable compromise between accuracy and computational efficiency for a vast majority of molecular systems and materials. In this work, we focus on KS-DFT approaches aiming at computing electronic ground-state (GS) properties. Having solved the minimization problem underlying DFT directly yields the ground-state density and corresponding energy. However, many quantities of interest, such as interatomic forces, (hyper)polarizabilities, magnetic susceptibilities, phonons spectra, or transport coefficients, correspond physically to the response of GS quantities to a change in external parameters (*e.g.* nuclear positions, electromagnetic fields). As such their mathematical expressions involve derivatives of the obtained GS solution with respect to these parameters. For example interatomic forces are *first-order* derivatives of the GS energy with respect to the atomic positions, and can actually be obtained without computing the derivatives of the GS density, thanks to the Hellmann–Feynman theorem [96]. On the other hand the computation of any property corresponding to *second- or higher-order* derivatives of the GS energy does require the computation of derivatives of the density. More precisely, it follows from Wigner's $(2n+1)$ theorem that $n^{\text{th}}$-order derivatives of the GS density are required to compute properties corresponding to $(2n)^{\text{th}}$- and $(2n+1)^{\text{st}}$-derivatives of the KS energy functional. More recent applications, such as the design of machine-learned exchange-correlation energy functionals, also require the computation of derivatives of the ground-state with respect to parameters, such as the ones defining the exchange-correlation functional [109, 113, 129].

Efficient numerical methods for evaluating these derivatives are therefore needed. The application of generic perturbation theory to the special case of DFT is known as density functional perturbation theory (DFPT) [15, 81, 82, 85]. See also [155] for applications to quantum chemistry, [14] for applications to phonon calculations, and [45] for a mathematical analysis of DFPT within the reduced Hartree–Fock (rHF) approximation (also called the Hartree approximation in the physics literature). Although the practical implementation of first- and higher-order derivatives computed by DFPT in electronic structure calculation software can be greatly simplified by Automatic Differentiation techniques [86], the efficiency of the resulting code crucially depends on the efficiency of a key building block: the computation of the linear response $\delta\rho$ of the GS density to an infinitesimal variation $\delta V$ of the total Kohn–Sham potential.

For reasons that will be detailed below, the numerical evaluation of the linear map $\delta V \mapsto \delta\rho$ is not straightforward, especially for periodic metallic systems. Indeed, DFT calculations for metallic systems usually require the introduction of a smearing temperature $T$, a numerical parameter which has nothing to do with the physical temperature (in practice, its value is often higher than the melting temperature of the crystal). For the sake of simplicity, let us first consider the case of a periodic simulation cell $\Omega$ containing an even number $N_{\text{el}}$ of electrons in a spin-unpolarized state (see Remark 4.1 for details on how this formalism allows for the computation of properties of perfect crystals). The Kohn–Sham GS at finite temperature $T > 0$ is then described by an $\mathrm{L}^2(\Omega)$-orthonormal set of orbitals $(\phi_n)_{n\in\mathbb{N}^*}$ with energies $(\varepsilon_n)_{n\in\mathbb{N}^*}$, which are the eigenmodes of the Kohn–Sham Hamiltonian $H$ associated with the GS density:

$$H\phi_n = \varepsilon_n\phi_n, \qquad \int_\Omega \phi_m^*(\boldsymbol{r})\phi_n(\boldsymbol{r})\mathrm{d}\boldsymbol{r} = \delta_{mn}, \qquad \varepsilon_1 \leqslant \varepsilon_2 \leqslant \varepsilon_3 \leqslant \cdots,$$

together with periodic boundary conditions. The GS density in turn reads

$$\rho(\boldsymbol{r}) = \sum_{n=1}^{+\infty} f_n|\phi_n(\boldsymbol{r})|^2 \quad \text{with} \quad f_n \coloneqq f\left(\frac{\varepsilon_n - \varepsilon_{\text{F}}}{T}\right), \tag{4.1.1}$$

where $f$ is a smooth occupation function converging to 2 at $-\infty$ and to zero at $+\infty$, *e.g.* the Fermi–Dirac smearing function $f(x) = \frac{2}{1+\mathrm{e}^x}$ (see Figure 4.1). The Fermi level $\varepsilon_{\text{F}}$ is the Lagrange multiplier of the neutrality charge constraint: it is the unique real number such that

$$\int_\Omega \rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} = \sum_{n=1}^{+\infty} f_n = \sum_{n=1}^{+\infty} f\left(\frac{\varepsilon_n - \varepsilon_{\text{F}}}{T}\right) = N_{\text{el}}.$$

It follows from perturbation theory that the linear response $\delta\rho$ of the density to an infinitesimal variation $\delta V$ of the *total* Kohn–Sham potential is given by

$$\delta\rho = \chi_0\delta V,$$

where $\chi_0$ is the independent-particle susceptibility operator (also called noninteracting density response function). Equivalently, this operator describes the linear response of a system of *noninteracting* electrons of density $\rho$ subject to an infinitesimal perturbation $\delta V$. It holds (see Section 4.3)

$$\delta\rho(\boldsymbol{r}) \coloneqq (\chi_0 \delta V)(\boldsymbol{r}) = \sum_{n=1}^{+\infty} \sum_{m=1}^{+\infty} \frac{f_n - f_m}{\varepsilon_n - \varepsilon_m} \phi_n^*(\boldsymbol{r}) \phi_m(\boldsymbol{r}) (\delta V_{mn} - \delta\varepsilon_{\mathrm{F}} \delta_{mn}), \tag{4.1.2}$$

where $\delta V_{mn} \coloneqq \langle \phi_m, \delta V \phi_n \rangle$, $\delta\varepsilon_{\mathrm{F}}$ is the induced variation of the Fermi level $\varepsilon_{\mathrm{F}}$, and $\delta_{mn}$ is the Kronecker symbol. We also use the convention $(f_n - f_n)/(\varepsilon_n - \varepsilon_n) = \frac{1}{T} f'\left(\frac{\varepsilon_n - \varepsilon_{\mathrm{F}}}{T}\right)$.

In practice, these equations are discretized on a finite basis set, so that the sums in (4.1.1) and (4.1.2) become finite. Since the number of basis functions $N_{\mathrm{b}}$ is often very large compared to the number of electrons in the system, it is very expensive to compute the sums as such. However, in practice it is possible to restrict to the computation of a number $N \ll N_{\mathrm{b}}$ of orbitals. These orbitals are then computed using efficient iterative methods [159].




FIGURE 4.1 – The occupation numbers $f_n$ for $T = 0$ (left) and $T > 0$ (right).

For insulating systems, there is a (possibly) large band gap between $\varepsilon_{N_{\mathrm{p}}}$ and $\varepsilon_{N_{\mathrm{p}}+1}$ which remains nonzero in the thermodynamic limit of a growing simulation cell. As a result, the calculation can be done at zero temperature, such that the occupation function $f$ becomes a step function (see Figure 4.1). The jump from 2 to 0 in the occupations occurs exactly when the lowest $N_{\mathrm{p}} = N_{\mathrm{el}}/2$ energy levels $\varepsilon_1 \leqslant \cdots \leqslant \varepsilon_{N_{\mathrm{p}}}$ are occupied with an electron pair (two electrons of opposite spin). Thus, $f_n = 2$ for $1 \leqslant n \leqslant N_{\mathrm{p}}$ and $f_n = 0$ for $n > N_{\mathrm{p}}$. As a result, $N$ can be chosen equal to the number of electron pairs $N_{\mathrm{p}}$ without any approximation. In contrast, for metallic systems $\varepsilon_{N_{\mathrm{p}}} = \varepsilon_{N_{\mathrm{p}}+1} = \varepsilon_{\mathrm{F}}$ in the zero-temperature thermodynamic limit (more precisely there is a positive density of states at the Fermi level in the limit of an infinite simulation cell), causing the denominators in the right-hand side of formula (4.1.2) to formally blow up. Calculations on metallic systems are thus done at finite temperature $T > 0$, in which case every orbital has a fractional occupancy $f_n \in (0, 2)$. However, since from a classical semiclassical approximation $\varepsilon_n$ tends to infinity as $n^{2/3}$ as $n \to \infty$, and $f$ decays very quickly, one can safely assume that only a finite number $N$ of orbitals have nonnegligible occupancies. This allows one to avoid computing $\phi_n$ for $n > N$. Under this approximation, a formal differentiation of (4.1.1) gives

$$\delta\rho(\boldsymbol{r}) = \sum_{n=1}^{N} f_n (\phi_n^*(\boldsymbol{r}) \delta\phi_n(\boldsymbol{r}) + \delta\phi_n^*(\boldsymbol{r}) \phi_n(\boldsymbol{r})) + \delta f_n |\phi_n(\boldsymbol{r})|^2. \tag{4.1.3}$$

However, while the response $\delta\rho$ to a given $\delta V$ is well-defined by (4.1.2), the set $(\delta\phi_n, \delta f_n)_{1 \leqslant n \leqslant N}$ is not. Indeed, the Kohn–Sham energy functional being in fact a function of the density matrix $\gamma = \sum_{n=1}^{N} f_n |\phi_n\rangle\langle\phi_n|$, any transformation of $(\delta\phi_n, \delta f_n)_{1 \leqslant n \leqslant N}$ leaving invariant the first-order variation

$$\delta\gamma \coloneqq \sum_{n=1}^{N} \delta f_n |\phi_n\rangle\langle\phi_n| + \sum_{n=1}^{N} f_n (|\phi_n\rangle\langle\delta\phi_n| + |\delta\phi_n\rangle\langle\phi_n|) \tag{4.1.4}$$

of the density matrix is admissible. This gauge freedom can be used to stabilize linear response calculations or, in the contrary, may lead to numerical instabilities. Denote by $P$ the orthogonal projector onto $\mathrm{Span}\,(\phi_n)_{1 \leqslant n \leqslant N}$, the space spanned by the orbitals considered as (partially) occupied, and by $Q = 1 - P$ the orthogonal projector onto the space $\mathrm{Span}\,(\phi_n)_{n > N}$ spanned by the orbitals considered as unoccupied. Then, the linear response of any occupied orbital can be decomposed as $\delta\phi_n = \delta\phi_n^P + \delta\phi_n^Q$ where:

- $\delta\phi_n^P = P\delta\phi_n \in \mathrm{Ran}\,(P)$ can be directly computed *via* a sum-over-state formula (explicit decomposition on the basis of $(\phi_n)_{n \leqslant N}$). This contribution can be chosen to vanish in the zero-temperature limit, as in that case $P\,\delta\gamma\,P = 0$. At finite temperature, a gauge choice has to be made and several options have been proposed in the literature;

- $\delta\phi_n^Q = Q\delta\phi_n \in \mathrm{Ran}\,(Q)$ is the unique solution of the so-called Sternheimer equation [188]

$$Q(H - \varepsilon_n)Q\delta\phi_n^Q = -Q\delta V\phi_n, \tag{4.1.5}$$

  where $H$ is the Kohn–Sham Hamiltonian of the system. This equation is possibly very ill-conditioned for $n = N$ if $\varepsilon_{N+1} - \varepsilon_N$ is very small.

This chapter addresses these two issues. First, we review and analyse the different gauge choices for $\delta\phi_n^P$ proposed in the literature and introduce a new one. We bring all these various gauge choices together in a new common framework and analyse their performance in terms of numerical stability. Second, for the contribution $\delta\phi_n^Q$, we investigate how to improve the conditioning of the linear system (4.1.5), which is usually solved with iterative solvers and we propose a new approach. This new approach is based on the fact that, as a byproduct of the iterative computation of the ground state orbitals $(\phi_n)_{n\leqslant N}$, one usually obtains relatively good approximations of the following eigenvectors. This information is often discarded for response calculations; we use them in a Schur complement approach to improve the conditioning of the iterative solve of the Sternheimer equation. We quantify the improvement of the conditioning obtained by this new approach and illustrate its efficiency on several metallic systems, from aluminium to transition metal alloys. We observe a reduction of typically 40% of the number of Hamiltonian applications (the most costly step of the calculation). The numerical tests have been performed with the DFTK software [101], a recently developed plane-wave DFT package in `Julia` allowing for both easy implementation of novel algorithms and numerical simulations of challenging systems of physical interest. The improvements suggested in this work are now the default choice in DFTK to solve response problems.

This chapter is organized as follows. In Section 4.2, we review the periodic KS-DFT equations and the associated approximations. We also present the mathematical formulation of DFPT and we detail the links between the orbitals' response $\delta\phi_n$ and the ground-state density response $\delta\rho$ for a given external perturbation, as well as the derivation of the Sternheimer equation (4.1.5). In Section 4.3, we propose a common framework for different natural gauge choices. Then, with focus on the Sternheimer equation and the Schur complement, we present the improved resolution to obtain $\delta\phi_n^Q$. Finally, in Section 4.4, we perform numerical simulations on relevant physical systems. In the appendix, we propose a strategy for choosing the number of extra orbitals motivated by a rough convergence analysis of the Sternheimer equation.

## 4.2 Mathematical framework

### 4.2.1 Periodic Kohn–Sham equations

We consider here a simulation cell $\Omega = [0,1)\boldsymbol{a}_1 + [0,1)\boldsymbol{a}_2 + [0,1)\boldsymbol{a}_3$ with periodic boundary conditions, where $(\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3)$ is a nonnecessarily orthonormal basis of $\mathbb{R}^3$. We denote by $\mathcal{R} = \mathbb{Z}\boldsymbol{a}_1 + \mathbb{Z}\boldsymbol{a}_2 + \mathbb{Z}\boldsymbol{a}_3$ the periodic lattice in the position space and by $\mathcal{R}^* = \mathbb{Z}\boldsymbol{b}_1 + \mathbb{Z}\boldsymbol{b}_2 + \mathbb{Z}\boldsymbol{b}_3$ with $\boldsymbol{a}_i \cdot \boldsymbol{b}_j = 2\pi\delta_{ij}$ the reciprocal lattice. Let us denote by

$$\mathrm{L}^2_\#(\mathbb{R}^3, \mathbb{C}) \coloneqq \{u \in \mathrm{L}^2_{\mathrm{loc}}(\mathbb{R}^3, \mathbb{C}) \mid u \text{ is } \mathcal{R}\text{-periodic}\} \tag{4.2.1}$$

the Hilbert space of complex-valued $\mathcal{R}$-periodic locally square integrable functions on $\mathbb{R}^3$, endowed with its usual inner product $\langle\cdot,\cdot\rangle$ and by $\mathrm{H}^s_\#(\mathbb{R}^3, \mathbb{C})$ the $\mathcal{R}$-periodic Sobolev space of order $s \in \mathbb{R}$

$$\mathrm{H}^s_\#(\mathbb{R}^3, \mathbb{C}) \coloneqq \left\{ u = \sum_{\boldsymbol{G}\in\mathcal{R}^*} \widehat{u}_{\boldsymbol{G}} e_{\boldsymbol{G}},\ \sum_{\boldsymbol{G}\in\mathcal{R}^*} (1 + |\boldsymbol{G}|^2)^s |\widehat{u}_{\boldsymbol{G}}|^2 < \infty \right\}$$

where $e_{\boldsymbol{G}}(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{r}}/\sqrt{|\Omega|}$ is the Fourier mode with wave-vector $\boldsymbol{G}$.

In atomic units, the KS equations for a system of $N_{\mathrm{el}} = 2N_{\mathrm{p}}$ spin-unpolarized electrons at finite temperature read

$$H_\rho\phi_n = \varepsilon_n\phi_n, \quad \varepsilon_1 \leqslant \varepsilon_2 \leqslant \cdots, \quad \langle\phi_n, \phi_m\rangle = \delta_{nm}, \quad \rho(\boldsymbol{r}) = \sum_{n=1}^{+\infty} f_n|\phi_n(\boldsymbol{r})|^2, \quad \sum_{n=1}^{+\infty} f_n = N_{\mathrm{el}}, \tag{4.2.2}$$

where $H_\rho$ is the Kohn–Sham Hamiltonian. It is given by

$$H_\rho = -\frac{1}{2}\Delta + V + V_\rho^{\mathrm{Hxc}} \tag{4.2.3}$$

where $V$ is the potential generated by the nuclei (or the ionic cores if pseudopotentials are used) of the system, and $V_\rho^{\mathrm{Hxc}}(\boldsymbol{r}) = V_\rho^{\mathrm{H}}(\boldsymbol{r}) + V_\rho^{\mathrm{xc}}(\boldsymbol{r})$ is an $\mathcal{R}$-periodic real-valued function depending on $\rho$. The Hartree potential $V_\rho^{\mathrm{H}}$ is the unique zero-mean solution to the periodic Poisson equation $-\Delta V_\rho^{\mathrm{H}}(\boldsymbol{r}) = 4\pi\Big(\rho(\boldsymbol{r}) - \frac{1}{|\Omega|}\int_\Omega \rho\Big)$ and the function $V_\rho^{\mathrm{xc}}$ is the exchange-correlation potential. $H_\rho$ is a self-adjoint operator on $\mathrm{L}^2_{\#}(\mathbb{R}^3, \mathbb{C})$, bounded below and with compact resolvent. Its spectrum is therefore composed of a nondecreasing sequence of eigenvalues $(\varepsilon_n)_{n\in\mathbb{N}^*}$ converging to $+\infty$. Since $H_\rho$ depends on the electronic density $\rho$, which in turn depends on the eigenfunctions $\phi_n$, (4.2.2) is a nonlinear eigenproblem, usually solved with *self-consistent field* (SCF) algorithms. These algorithms are based on successive partial diagonalizations of the Hamiltonian $H_{\rho_n}$ built from the current iterate $\rho_n$. See [GK1, 43, 133] and references therein for a mathematical presentation of SCF algorithms.

In (4.2.2), the $\phi_n$'s are the Kohn–Sham orbitals, with energy $\varepsilon_n$ and occupation number $f_n$. At finite temperature $T > 0$, $f_n$ is a real number in the interval $[0, 2)$ and we have

$$f_n = f\left(\frac{\varepsilon_n - \varepsilon_{\mathrm{F}}}{T}\right), \tag{4.2.4}$$

where $f$ is a fixed analytic *smearing* function, which we choose here equal to twice the Fermi–Dirac function: $f(x) = 2/(1 + \mathrm{e}^x)$. The Fermi level $\varepsilon_{\mathrm{F}}$ is then uniquely defined by the charge constraint

$$\sum_{n=1}^{+\infty} f_n = N_{\mathrm{el}}. \tag{4.2.5}$$

When $T \to 0$, $f((\cdot - \varepsilon_{\mathrm{F}})/T) \to 2 \times \mathbf{1}_{\{\cdot < \varepsilon_{\mathrm{F}}\}}$ almost everywhere, and only the lowest $N_{\mathrm{p}} = N_{\mathrm{el}}/2$ energy levels for which $\varepsilon_n < \varepsilon_{\mathrm{F}}$ are occupied by two electrons of opposite spins (see Figure 4.1): $f_n = 2$ for $n \leqslant N_{\mathrm{p}}$ and $f_n = 0$ for $n > N_{\mathrm{p}}$.

**Remark 4.1** (The case of perfect crystals)**.** Using a finite simulation cell $\Omega$ with periodic boundary conditions is usually the best way to compute the bulk properties of a material in the condensed phase. Indeed, KS-DFT simulations are limited to, say $10^3 - 10^4$ electrons, on currently available standard computer architectures. Simulating *in vacuo* a small sample of the material containing, say $10^3$ atoms, would lead to completely wrong results, polluted by surface effects since about half of the atoms would lay on the sample surface. Periodic boundary conditions are a trick to get rid of surface effects, at the price of artificial interactions between the sample and its periodic image. In the case of a perfect crystal with Bravais lattice $\mathbb{L}$ and unit cell $\omega$, it is natural to choose a periodic simulation (super)cell $\Omega = L\omega$ consisting of $L^3$ unit cells (we then have $\mathcal{R} = L\mathbb{L}$). In the absence of spontaneous symmetry breaking, the KS ground-state density has the same $\mathbb{L}$-translational invariance as the nuclear potential. Using Bloch theory, the supercell eigenstates $\phi_n$ can then be relabelled as $\phi_n(\boldsymbol{r}) = e^{i\boldsymbol{k}\cdot\boldsymbol{r}} u_{j\boldsymbol{k}}(\boldsymbol{r})$, where $u_{j\boldsymbol{k}}$ now has cell periodicity, and equations (4.2.2)–(4.2.4) can be rewritten as

$$H_{\rho,\boldsymbol{k}} u_{j\boldsymbol{k}} = \varepsilon_{j\boldsymbol{k}} u_{j\boldsymbol{k}}, \quad \varepsilon_{1\boldsymbol{k}} \leqslant \varepsilon_{2\boldsymbol{k}} \leqslant \cdots, \quad \langle u_{j\boldsymbol{k}}, u_{j'\boldsymbol{k}} \rangle = \delta_{jj'}, \tag{4.2.6}$$

$$\rho(\boldsymbol{r}) = \frac{1}{L^3} \sum_{\boldsymbol{k}\in\mathcal{G}_L} \sum_{j=1}^{+\infty} f_{j\boldsymbol{k}} |u_{j\boldsymbol{k}}(\boldsymbol{r})|^2, \quad \frac{1}{L^3} \sum_{\boldsymbol{k}\in\mathcal{G}_L} \sum_{j=1}^{+\infty} f_{j\boldsymbol{k}} = N_{\mathrm{el}}, \quad f_{j\boldsymbol{k}} = f\left(\frac{\varepsilon_{j\boldsymbol{k}} - \varepsilon_{\mathrm{F}}}{T}\right) \tag{4.2.7}$$

$$H_{\rho,\boldsymbol{k}} = \frac{1}{2}(-\mathrm{i}\nabla + \boldsymbol{k})^2 + V + V_\rho^{\mathrm{Hxc}}, \tag{4.2.8}$$

where $\mathcal{G}_L = L^{-1}\mathbb{L}^* \cap \omega^*$. Here $\mathbb{L}^*$ is the dual lattice of $\mathbb{L}$ and $\omega = \mathbb{R}^3/\mathbb{L}^*$ the first Brillouin zone of the crystal. In the thermodynamic limit $L \to \infty$, we obtain the periodic Kohn–Sham equations at finite temperature

$$H_{\rho,\boldsymbol{k}} u_{j\boldsymbol{k}} = \varepsilon_{j\boldsymbol{k}} u_{j\boldsymbol{k}}, \quad \varepsilon_{1\boldsymbol{k}} \leqslant \varepsilon_{2\boldsymbol{k}} \leqslant \cdots, \quad \langle u_{j\boldsymbol{k}}, u_{j'\boldsymbol{k}} \rangle = \delta_{jj'}, \tag{4.2.9}$$

$$\rho(\boldsymbol{r}) = \fint_{\omega^*} \sum_{j=1}^{+\infty} f_{j\boldsymbol{k}} |u_{j\boldsymbol{k}}(\boldsymbol{r})|^2 \mathrm{d}\boldsymbol{k}, \quad \fint_{\omega^*} \sum_{j=1}^{+\infty} f_{j\boldsymbol{k}} \mathrm{d}\boldsymbol{k} = N_{\mathrm{el}}, \quad f_{j\boldsymbol{k}} = f\left(\frac{\varepsilon_{j\boldsymbol{k}} - \varepsilon_{\mathrm{F}}}{T}\right). \tag{4.2.10}$$

This is a massive reduction in complexity, as now only computations on the unit cell have to be performed. For metals, the integrand on the Brillouin zone is discontinuous at zero temperature, which makes standard quadrature methods fail. Introducing a smearing temperature $T > 0$ allows one to smooth out the integrand, see [39, 125] for a numerical analysis of the smearing technique. We also refer for instance to [171, Section XIII.16] for more details on Bloch theory, to [33, 48] for a proof of the thermodynamic limit for perfect crystals in the rHF setting for both insulators and metals, and to [80] for the numerical analysis for insulators.

### 4.2.2 Density functional perturbation theory

We detail in this section the mathematical framework of DFPT. We first rewrite the Kohn–Sham equations (4.2.2) as the fixed-point equation for the density $\rho$

$$F\big(V + V_\rho^{\mathrm{Hxc}}\big) = \rho, \tag{4.2.11}$$

where $F$ is the potential-to-density mapping defined by

$$F(V)(\boldsymbol{r}) = \sum_{n=1}^{+\infty} f\left(\frac{\varepsilon_n - \varepsilon_{\mathrm{F}}}{T}\right) |\phi_n(\boldsymbol{r})|^2 \tag{4.2.12}$$

with $(\varepsilon_n, \phi_n)_{n\in\mathbb{N}^*}$ an orthonormal basis of eigenmodes of $-\frac{1}{2}\Delta + V$ and $\varepsilon_{\mathrm{F}}$ defined by (4.2.5). The solution of (4.2.11) defines a mapping from $V$ to $\rho$: the purpose of DFPT is to compute its derivative. Let $\delta V_0$ be a local infinitesimal perturbation, in the sense that it can be represented by a multiplication operator by a periodic function $\boldsymbol{r} \mapsto \delta V_0(\boldsymbol{r})$. By taking the derivative of (4.2.11) with the chain rule, we obtain the implicit equation for $\delta\rho$:

$$\delta\rho = F'\big(V + V_\rho^{\mathrm{Hxc}}\big) \cdot (\delta V_0 + K_\rho^{\mathrm{Hxc}}\delta\rho), \tag{4.2.13}$$

where the Hartree-exchange-correlation kernel $K_\rho^{\mathrm{Hxc}}$ is the derivative of the map $\rho \mapsto V_\rho^{\mathrm{Hxc}}$ and $F'\big(V + V_\rho^{\mathrm{Hxc}}\big)$ is the derivative of $F$ computed at $V + V_\rho^{\mathrm{Hxc}}$. In the field of DFT calculations, the latter operator is known as the *independent-particle susceptibility* operator and is denoted by $\chi_0$. This yields the Dyson equation

$$\delta\rho = \chi_0(\delta V_0 + K_\rho^{\mathrm{Hxc}}\delta\rho) \quad \Leftrightarrow \quad \delta\rho = \big(1 - \chi_0 K_\rho^{\mathrm{Hxc}}\big)^{-1}\chi_0\delta V_0. \tag{4.2.14}$$

This equation is commonly solved by iterative methods, which require efficient and robust computations of $\chi_0\delta V$ for various right-hand sides $\delta V$'s. In the rest of this chapter, we forget about the solution of (4.2.14) and focus on the computation of the noninteracting response $\delta\rho \coloneqq \chi_0\delta V$ for a given $\delta V$.

The operator $\chi_0$ maps $\delta V$ to the first-order variation $\delta\rho$ of the ground-state density of a noninteracting system of electrons ($K^{\mathrm{Hxc}} = 0$). Denoting $A_{mn} \coloneqq \langle\phi_m, A\phi_n\rangle$ for a given operator $A$, it holds

$$\delta\rho(\boldsymbol{r}) = \sum_{n=1}^{+\infty}\sum_{m=1}^{+\infty} \frac{f_n - f_m}{\varepsilon_n - \varepsilon_m}\phi_n^*(\boldsymbol{r})\phi_m(\boldsymbol{r})(\delta V_{mn} - \delta\varepsilon_{\mathrm{F}}\delta_{mn}), \tag{4.2.15}$$

where $\delta_{mn}$ is the Kronecker delta, $\delta\varepsilon_{\mathrm{F}}$ is the induced variation in the Fermi level and we use the following convention

$$\frac{f_n - f_n}{\varepsilon_n - \varepsilon_n} = \frac{1}{T}f'\left(\frac{\varepsilon_n - \varepsilon_{\mathrm{F}}}{T}\right) =: f_n'. \tag{4.2.16}$$

Charge conservation leads to

$$\int_\Omega \delta\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} = 0 \quad \Rightarrow \quad \delta\varepsilon_{\mathrm{F}} = \frac{\sum_{n=1}^{+\infty} f_n'\delta\varepsilon_n}{\sum_{n=1}^{+\infty} f_n'}, \tag{4.2.17}$$

where $\delta\varepsilon_n \coloneqq \delta V_{nn}$. We refer to [14] for a physical discussion of this formula, and to [44, 98, 125], where it is proven rigorously using contour integrals.

**Remark 4.2.** Similar to the discussion above on the computation of perfect crystal employing Bloch theory, response computations of perfect crystals can be performed by decomposing $\delta V_0$ in its Bloch modes. This allows for the efficient computation of phonon spectral or dielectric functions for instance.

**Remark 4.3.** We restricted our discussion for simplicity to local potentials, but the formalism can easily be extended to nonlocal perturbations (such as the ones created by pseudopotentials in the Kleinman-Bylander form [114]).

### 4.2.3 Plane-wave discretization and numerical resolution

In this chapter we are interested in plane-wave DFT calculations of metallic systems. This corresponds to a specific Galerkin approximation of the Kohn–Sham model using as variational approximation space

$$X_{N_\mathrm{b}} \coloneqq \mathrm{Span}\left\{e_{\boldsymbol{G}},\ \boldsymbol{G}\in\mathcal{R}^*,\ \frac{1}{2}|\boldsymbol{G}|^2 \leqslant E_\mathrm{cut}\right\}, \tag{4.2.18}$$

where $N_\mathrm{b}$ denotes the dimension of the discretization space, linked to the cut-off energy $E_\mathrm{cut}$. Denoting by $\Pi_{N_\mathrm{b}}$ the orthogonal projection onto $X_{N_\mathrm{b}}$ for the $\mathrm{L}^2_\#$ inner product, we then solve the discrete problem: find $\phi_1,\dots,\phi_{N_\mathrm{b}}\in X_{N_\mathrm{b}}$ such that

$$\begin{cases} \Pi_{N_\mathrm{b}} H_\rho \Pi_{N_\mathrm{b}} \phi_n = \varepsilon_n \phi_n, \quad \varepsilon_1 \leqslant \cdots \leqslant \varepsilon_{N_\mathrm{b}}, \\ \rho = \sum_{n=1}^{N_\mathrm{b}} f_n |\phi_n|^2, \quad \sum_{n=1}^{N_\mathrm{b}} f_n = N_\mathrm{el}, \quad f_n = f\left(\frac{\varepsilon_n-\varepsilon_\mathrm{F}}{T}\right), \\ \langle \phi_n, \phi_m\rangle = \delta_{nm}, \quad n,m=1,\dots,N_\mathrm{b}, \end{cases} \tag{4.2.19}$$

where $H_\rho$ is the Kohn–Sham Hamiltonian (or one of its Bloch fibers). This discretization method for Kohn–Sham equations has been analysed for instance in [31].

We emphasize again the point that not all $N_\mathrm{b}$ eigenpairs need to be computed. At zero temperature, only the $N = N_\mathrm{el}/2$ lowest energy Kohn–Sham orbitals need to be fully converged as they are the only occupied ones. At finite temperature, the number of bands with meaningful occupation numbers is usually higher than the number of electrons, but the fast decay of the occupation numbers allows to avoid computing all $N_\mathrm{b}$ eigenpairs. Determining the number of bands to compute is not easy as, at finite temperature, we do not know *a priori* the number of bands that are significantly occupied. A standard heuristic is to fully converge 20% more bands than the number of electrons pairs during the SCF. For the response calculation we then select the number $N$ of bands that have occupation numbers above some numerical threshold. On top of these bands, it is common in DFT calculations to add additional bands that are not fully converged by the successive eigensolvers. The main advantages of introducing these bands are: (i) they enhance the diagonalization procedure by increasing the gap between converged and uncomputed bands and (ii) adding extra bands is not very expensive when the diagonalization is performed with block-based methods, such as the LOBPCG algorithm [115].

## 4.3 Computing the response

### 4.3.1 Practical implementation

Using (4.2.15) as it stands is not possible because of the large sums. One possibility is to represent $\delta\rho$ through a collection of occupied orbital variations $(\delta\phi_n)_{1\leqslant n\leqslant N}$ and occupation number variations $(\delta f_n)_{1\leqslant n\leqslant N}$. One then has to make appropriate ansatz and gauge choices on the links between $\delta\rho$ and its representation. Differentiating the formula $\rho(\boldsymbol{r}) = \sum_{n=1}^N f_n|\phi_n(\boldsymbol{r})|^2$, one gets

$$\delta\rho(\boldsymbol{r}) = \sum_{n=1}^N f_n(\phi_n^*(\boldsymbol{r})\delta\phi_n(\boldsymbol{r}) + \delta\phi_n^*(\boldsymbol{r})\phi_n(\boldsymbol{r})) + \delta f_n|\phi_n(\boldsymbol{r})|^2. \tag{4.3.1}$$

Then, for $n\leqslant N$, we expand $f_n\delta\phi_n$ into the basis $(\phi_m)_{m\in\mathbb{N}}$. Defining

$$\Gamma_{mn} \coloneqq \langle \phi_m, f_n\delta\phi_n\rangle, \tag{4.3.2}$$

yields

$$\forall\ 1\leqslant n\leqslant N, \quad f_n\delta\phi_n = \sum_{m=1}^N \Gamma_{mn}\phi_m + f_n\delta\phi_n^Q \tag{4.3.3}$$

where $\delta\phi_n^Q \coloneqq Q\delta\phi_n$ and $Q$ is the orthogonal projector onto $\mathrm{Span}\,(\phi_m)_{N<m}$, the space spanned by the unoccupied orbitals. Plugging (4.3.3) into (4.3.1), we obtain, using symmetry between $n$ and $m$,

$$\delta\rho(\boldsymbol{r}) = \sum_{n,m=1}^N \phi_n^*(\boldsymbol{r})\phi_m(\boldsymbol{r})\left(\Gamma_{mn} + \overline{\Gamma_{nm}}\right) + \sum_{n=1}^N \delta f_n|\phi_n(\boldsymbol{r})|^2 + \sum_{n=1}^N 2f_n\mathrm{Re}\left(\phi_n^*(\boldsymbol{r})\delta\phi_n^Q(\boldsymbol{r})\right). \tag{4.3.4}$$

A first gauge choice can be made here. Using again the charge conservation, we get

$$0 = \int_\Omega \delta\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} \quad \Rightarrow \quad 0 = \sum_{n=1}^{N} 2\mathrm{Re}\left(\Gamma_{nn}\right) + \delta f_n. \tag{4.3.5}$$

Given that we adapt $\delta f_n$ accordingly we can thus assume $\Gamma_{nn} = 0$ for any $1 \leqslant n \leqslant N$. We will make this gauge choice from this point, leaving the constraint $\sum_{n=1}^{N} \delta f_n = 0$ to restrict possible choices of $\delta f_n$.

We now derive conditions on $(\Gamma_{mn})_{1\leqslant n,m\leqslant N}$, $(\delta f_n)_{1\leqslant n\leqslant N}$ and $(\delta\phi_n^Q)_{1\leqslant n\leqslant N}$ so that the ansatz we made is a valid representation of $\delta\rho$, that is to say (4.3.4) coincides with (4.2.15). To this end, we rewrite (4.2.15) as

$$\delta\rho(\boldsymbol{r}) = \sum_{n,m=1}^{N} \frac{f_n - f_m}{\varepsilon_n - \varepsilon_m}\phi_n^*(\boldsymbol{r})\phi_m(\boldsymbol{r})(\delta V_{mn} - \delta\varepsilon_{\mathrm{F}}\delta_{mn}) + \sum_{n=1}^{N}\sum_{m=N+1}^{+\infty} 2\frac{f_n}{\varepsilon_n - \varepsilon_m}\mathrm{Re}\left(\phi_n^*(\boldsymbol{r})\phi_m(\boldsymbol{r})\delta V_{mn}\right), \tag{4.3.6}$$

where the terms $f_n$, $f_m$ for which $n, m > N+1$ have been neglected because of their small occupation numbers and we used the symmetry between $n$ and $m$ for the terms with $1 \leqslant n \leqslant N$, $m > N$. From a term by term comparison between (4.3.4) and (4.3.6), we infer first from the $n = m$ term and the gauge choice $\Gamma_{nn} = 0$ that $\delta f_n = f_n'(\delta V_{nn} - \delta\varepsilon_{\mathrm{F}}) = f_n'(\delta\varepsilon_n - \delta\varepsilon_{\mathrm{F}})$. Note that, thanks to the definition (4.2.17) of $\delta\varepsilon_{\mathrm{F}}$, charge conservation is indeed satisfied. Next, for the first sum to coincide between (4.3.4) and (4.3.6), we see that the $\Gamma_{mn}$'s have to satisfy

$$\forall\ 1 \leqslant n, m \leqslant N, m \neq n, \qquad \Gamma_{mn} + \overline{\Gamma_{nm}} = \frac{f_n - f_m}{\varepsilon_n - \varepsilon_m}\delta V_{mn} =: \Delta_{mn}. \tag{4.3.7}$$

Finally, since $\delta\phi_n^Q \in \mathrm{Span}\left(\phi_m\right)_{N<m}$, we deduce from the last sum in (4.3.4) and (4.3.6) that $\delta\phi_n^Q$ can be computed as the unique solution of the linear system

$$\forall\ 1 \leqslant n \leqslant N, \quad Q(H_\rho - \varepsilon_n)Q\delta\phi_n^Q = -Q\delta V\phi_n, \tag{4.3.8}$$

sometimes known in DFT as the Sternheimer equation [188]. Note that $\delta\phi_N^Q$ can be arbitrarily large, since $\varepsilon_{N+1} - \varepsilon_N$ may be arbitrarily small. However, this does not pose a problem in practice as $\delta\phi_N^Q$ is multiplied by $f_N$ (cf. (4.3.4)), which is very small.

To summarize, the response $\delta\rho = \chi_0\delta V$ can be computed as

$$\delta\rho(\boldsymbol{r}) = \sum_{n=1}^{N} 2f_n\mathrm{Re}\left(\phi_n^*(\boldsymbol{r})\delta\phi_n(\boldsymbol{r})\right) + \delta f_n|\phi_n(\boldsymbol{r})|^2. \tag{4.3.9}$$

Here, $\delta f_n = f_n'(\delta\varepsilon_n - \delta\varepsilon_{\mathrm{F}})$, and $\delta\phi_n$ is separated into two contributions:

$$\forall\ 1 \leqslant n \leqslant N, \quad \delta\phi_n = \delta\phi_n^P + \delta\phi_n^Q, \tag{4.3.10}$$

where $(\delta\phi_n^P, \delta\phi_n^Q) \in \mathrm{Ran}\left(P\right) \times \mathrm{Ran}\left(Q\right)$ with $P$ the orthogonal projector onto $\mathrm{Span}\left(\phi_m\right)_{1\leqslant m\leqslant N}$ and $Q = 1 - P$. These two contributions are computed as follows:

- $\delta\phi_n^P$ is computed *via* a sum-over-states $m \neq n$:

$$\delta\phi_n^P = \sum_{m=1,m\neq n}^{N} \Gamma_{mn}\phi_m, \tag{4.3.11}$$

  where the $\Gamma_{mn}$'s satisfy $\Gamma_{mn} + \overline{\Gamma_{nm}} = \Delta_{mn}$. An additional gauge choice has to be made as these constraints do not yet define $\Gamma_{nm}$ uniquely. We refer to this term as the occupied-occupied contribution.

- $\delta\phi_n^Q$ is obtained as the solution of the Sternheimer equation (4.3.8). However, this linear system is possibly very ill-conditioned if $\varepsilon_{N+1} - \varepsilon_N$ is small. We refer to this term as the unoccupied-occupied contribution.

Note that, at zero temperature, $\delta\phi_n^P$ vanishes so that $\delta\phi_n = \delta\phi_n^Q \in \mathrm{Ran}\left(Q\right)$ and only the Sternheimer equation (4.3.8) needs to be solved. In the next two sections, we detail the practical computation of these two contributions.

### 4.3.2 Occupied-occupied contributions

In this section we discuss possible gauge choices for $\Gamma_{mn}$ to obtain a unique solution to (4.3.7). Throughout this section we assume $m \neq n$ and $\Gamma_{nn} = 0$.

#### Orthogonal gauge

The orthogonal gauge choice is motivated from the zero temperature setting, where $\delta\phi_n^P = 0$ allows to trivially preserve the orthogonality amongst the computed orbitals $\phi_n$ under the perturbation. For the case involving temperature, we additionally impose

$$0 = \delta\langle\phi_m, \phi_n\rangle = \langle\phi_m, \delta\phi_n\rangle + \langle\delta\phi_m, \phi_n\rangle, \tag{4.3.12}$$

and therefore

$$\frac{1}{f_n}\Gamma_{mn} + \frac{1}{f_m}\overline{\Gamma_{nm}} = 0, \tag{4.3.13}$$

yielding

$$\Gamma_{mn}^{\text{orth}} = \frac{f_n}{\varepsilon_n - \varepsilon_m}\delta V_{mn} \text{ for } m \neq n. \tag{4.3.14}$$

As a result $f_n\delta\phi_n$ features a possibly large contribution $\Gamma_{mn}^{\text{orth}}$, which is going to be almost compensated in (4.3.9) by the large contribution $\overline{\Gamma_{nm}^{\text{orth}}}$ to $f_n\delta\phi_n^*$ due the requirement to sum to the moderate-size contribution $\Gamma_{mn}^{\text{orth}} + \overline{\Gamma_{nm}^{\text{orth}}} = \Delta_{mn}$. This can lead to numerical instabilities because small errors, *e.g.* due to the fact that the $\phi_n$'s in (4.3.9) are eigenvectors only up to the solver tolerance, will get amplified by the $\Gamma_{mn}$. The next gauge choices provide solutions to this issue.

#### Simple gauge choice

Possibly the simplest gauge choice is $\Gamma_{mn}^{\text{simple}} = \frac{1}{2}\Delta_{mn}$. Since $(\Delta_{mn})_{1\leqslant n,m\leqslant N}$ is Hermitian, (4.3.7) is immediately satisfied.

#### Quantum Espresso gauge

The DFPT framework presented in [14] and implemented in Quantum Espresso [78] suggests choosing

$$\Gamma_{mn}^{\text{QE}} = f_{\text{FD}}\left(\frac{\varepsilon_n - \varepsilon_m}{T}\right)\Delta_{mn}, \tag{4.3.15}$$

where $f_{\text{FD}} = \frac{1}{2}f$ is the Fermi–Dirac functional. Since $f_{\text{FD}}(x) + f_{\text{FD}}(-x) = 1$, we have $\Gamma_{mn}^{\text{QE}} + \overline{\Gamma_{mn}^{\text{QE}}} = \Delta_{mn}$.

#### Abinit gauge

In the Abinit software [84, 175], the choice is

$$\Gamma_{mn}^{\text{Ab}} = \mathbf{1}_{\{f_n > f_m\}}\Delta_{mn}. \tag{4.3.16}$$

#### Minimal gauge

Motivated by our analysis of the instabilities we suggest minimizing $\delta\phi_n$, that is to ensure $\Gamma_{mn}/f_n$ to stay as small as possible. This leads to the minimization problem

$$\begin{aligned} \min \quad & \sum_{n,m=1,m\neq n}^{N} \frac{1}{f_n^2}|\Gamma_{mn}|^2, \\ \text{s.t.} \quad & \Gamma_{mn} + \overline{\Gamma_{nm}} = \Delta_{mn}, \quad \forall\, 1 \leqslant n, m \leqslant N,\ m \neq n. \end{aligned} \tag{4.3.17}$$

As the constraint (4.3.7) only couples $(n, m)$ and $(m, n)$, this translates into an uncoupled system of constrained minimization problems: for $1 \leqslant n, m \leqslant N$, $m \neq n$, solve

$$\begin{aligned} \min \quad & \frac{1}{f_n^2}|\Gamma_{mn}|^2 + \frac{1}{f_m^2}|\overline{\Gamma_{nm}}|^2, \\ \text{s.t.} \quad & \Gamma_{mn} + \overline{\Gamma_{nm}} = \Delta_{mn}, \end{aligned} \tag{4.3.18}$$

whose solution is

$$\Gamma_{mn}^{\min} = \frac{f_n^2}{f_n^2 + f_m^2}\Delta_{mn}. \tag{4.3.19}$$

This gauge choice is implemented by default in the DFTK software [101]. Another gauge choice inspired from this one would be to directly minimize $|\Gamma_{mn}|^2 + |\overline{\Gamma_{nm}}|^2$ but it can be shown that this leads to the simple gauge choice $\Gamma_{mn}^{\text{simple}} = \frac{1}{2}\Delta_{mn}$.

### Comparison of gauge choices

From (4.2.15) we can see that the growth of $\delta\rho$ with respect to $\delta V$ can not be higher than the growth of $\Delta_{mn}$ with respect to $\delta V$. The latter is of the order of $\max_{x\in\mathbb{R}} \frac{1}{T}|f'(x)| = \frac{1}{2T}$, which thus provides an intrinsic limit to the conditioning of the problem. For all gauge choices but the orthogonal one easily verifies

$$|\Gamma_{mn}| \leqslant |\Delta_{mn}| \leqslant \max_{x\in\mathbb{R}} \frac{1}{T}|f'(x)||\delta V_{mn}| = \frac{1}{2T}|\delta V_{mn}|. \tag{4.3.20}$$

If we make an error on $\delta V$ it is thus at most amplified by a factor of $\frac{1}{2T}$. All choices but the orthogonal one thus manage to stay within the intrinsic conditioning limit, see Figure 4.2.

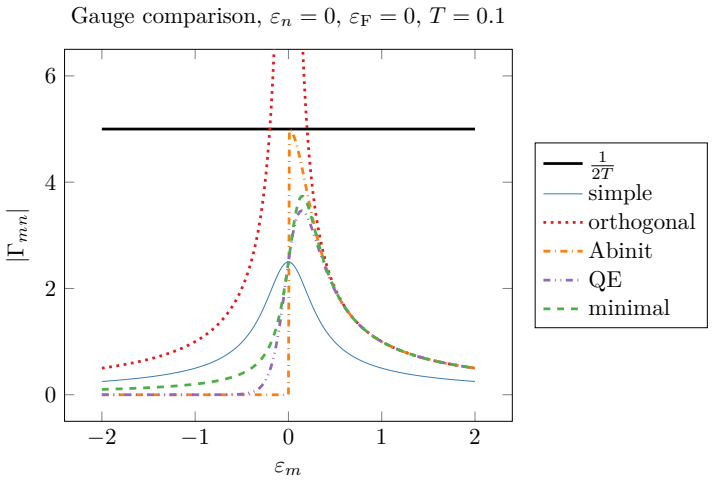



FIGURE 4.2 – Comparison of gauge choices for $\delta V_{mn} = 1$. Except the orthogonal gauges, all contributions $\Gamma_{mn}$ are bounded by $\frac{1}{2T}$.

## 4.3.3 Computation of unoccupied-occupied contributions employing a Schur complement

Since the $\phi_m$ for $m > N$ are not exactly known, a different approach is needed for obtaining the contribution $\delta\phi_n^Q$. Usually one resorts to solving the Sternheimer equation

$$\forall\ 1 \leqslant n \leqslant N, \quad Q(H_\rho - \varepsilon_n)Q\delta\phi_n^Q = -Q\delta V\phi_n =: b_n \tag{4.3.21}$$

using iterative schemes restricted to $\text{Ran}\,(Q)$. However, for $n = N$ the difference $\varepsilon_{N+1} - \varepsilon_N$ can become small, which deteriorates conditioning and increases the number of iterations required for convergence.

We overcome this issue by making use of the $N_{\text{ex}}$ extra bands, which are anyway available after the SCF algorithm has completed. Following Section 4.2.3 the $N_{\text{ex}}$ extra bands can be divided into two categories:

1. Some (usually the lower-energy ones) have been discarded during the response calculation because they have a too small occupation. Up to the eigensolver tolerance these are exact eigenvectors.

2. The remaining ones have served to accelerate the successive diagonalization steps during the SCF. These have not yet been fully converged.

In any case these extra bands thus offer (at least) approximate information about some $\phi_m$ for $m > N$, which is the underlying reason why the following approach accelerates the computation of $\delta\phi_n^Q$.

For the sake of clarity, we place ourselves here in the discrete setting: $H_\rho \in \mathbb{C}^{N_{\text{b}} \times N_{\text{b}}}$, $\Phi \in \mathbb{C}^{N_{\text{b}} \times N}$ and $\widetilde{\Phi} \in \mathbb{C}^{N_{\text{b}} \times N_{\text{ex}}}$. We assume that the number of computed bands $N + N_{\text{ex}}$ is larger than the number of occupied states $N$ and that we trust $\Phi = (\phi_1, \dots, \phi_N)$ but not $\widetilde{\Phi} = (\widetilde{\phi}_{N+1}, \dots, \widetilde{\phi}_{N+N_{\text{ex}}})$ to be eigenvectors. These $N_{\text{ex}}$ extra bands consist of both contributions (1) and (2) described at the beginning of this section. We assume in addition that $(\Phi, \widetilde{\Phi})$ forms an orthonormal family and that $\widetilde{\Phi}^* H_\rho \widetilde{\Phi}$ is a diagonal matrix whose elements, denoted by $(\widetilde{\varepsilon}_n)_{n=N+1,\dots,N+N_{\text{ex}}}$, are not necessarily all exact eigenvalues. Note that Rayleigh–Ritz based iterative methods such as the LOBPCG algorithm fit exactly in this framework. We

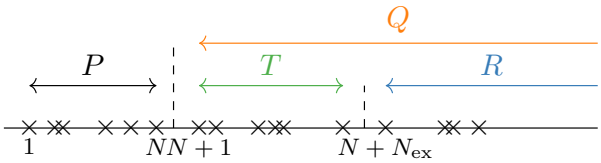



FIGURE 4.3 – Graphical representation of the Schur decomposition to solve the Sternheimer equation. $P$ is the orthogonal projector onto the occupied states. $Q$ is the orthogonal projector onto the unoccupied states, and we decompose it as the sum of $T$ (extra states which we can use) and $R$ (remaining states).

decompose

$$\text{Ran}\,(Q) = \text{Ran}\,(T) \oplus \text{Ran}\,(R), \tag{4.3.22}$$

where $T$ is the orthogonal projector onto $\text{Span}\,(\widetilde{\phi}_m)_{N<m\leqslant N+N_{\text{ex}}}$ and $R = Q - T$ is the projector onto the remaining (uncomputed) states, see Figure 4.3. Then, as $\delta\phi_n^Q \in \text{Ran}\,(Q)$, we can decompose

$$\delta\phi_n^Q = \widetilde{\Phi}\alpha_n + \delta\phi_n^R, \tag{4.3.23}$$

where $\alpha_n \in \mathbb{C}^{N_{\text{ex}}}$ and $\delta\phi_n^R \in \text{Ran}\,(R)$. Plugging this into (4.3.21) we get

$$Q(H_\rho - \varepsilon_n)\widetilde{\Phi}\alpha_n + Q(H_\rho - \varepsilon_n)\delta\phi_n^R = b_n. \tag{4.3.24}$$

Using a Schur complement we deduce

$$\alpha_n = \left(\widetilde{\Phi}^*(H_\rho - \varepsilon_n)\widetilde{\Phi}\right)^{-1}\left(\widetilde{\Phi}^* b_n - \widetilde{\Phi}^*(H_\rho - \varepsilon_n)\delta\phi_n^R\right). \tag{4.3.25}$$

Inserting (4.3.25) into (4.3.24) and projecting on $\text{Ran}\,(R)$ yields an equation in $\delta\phi_n^R$:

$$\begin{aligned} R(H_\rho - \varepsilon_n)\left[1 - \widetilde{\Phi}\left(\widetilde{\Phi}^*(H_\rho - \varepsilon_n)\widetilde{\Phi}\right)^{-1}\widetilde{\Phi}^*(H_\rho - \varepsilon_n)\right]R\delta\phi_n^R \\ = Rb_n - R(H_\rho - \varepsilon_n)\widetilde{\Phi}\left(\widetilde{\Phi}^*(H_\rho - \varepsilon_n)\widetilde{\Phi}\right)^{-1}\widetilde{\Phi}^* b_n. \end{aligned} \tag{4.3.26}$$

This equation can then be solved for $\delta\phi_n^R$ with a Conjugate Gradient (CG) method which is enforced to stay in $\text{Ran}\,(R)$ at each iteration. Afterwards we compute $\alpha_n$ from (4.3.25), which yields $\delta\phi_n^Q$ from (4.3.23). This scheme has been implemented as the default solver for the Sternheimer equation in DFTK.

## 4.4 Numerical tests

For all the numerical tests, we use the DFTK software [101], a recent plane-wave DFT package in `Julia`. All the codes to run the simulation of this chapter are available online[1]. The Brillouin zone is discretized using a uniform Monkhorst–Pack grid [150]. We use the PBE exchange-correlation functional [160] and GTH pseudopotentials [79, 91]. The other parameters of the calculation will be specified for each example. In all the tests, we generate a perturbation $\delta V$ from atomic displacements, with local and nonlocal contributions. Then, we perform two response calculations: one with the standard approach to solve directly the Sternheimer equation (4.3.21) to compute $\delta\phi_n^Q$, the other with the (new) Schur complement approach (4.3.26). Both linear systems are solved using the conjugate gradient (CG) algorithm, with kinetic energy preconditioning (the linear solver is preconditioned with the inverse Laplacian, which is diagonal in Fourier representation), and we compare the number of iterations required to converge the norm of the residual below $10^{-9}$. Note that the Sternheimer equation is solved for all $N$ occupied orbitals and for each $k$-point.

If $T > 0$ the contribution $\delta\phi_n^P$ is nonzero and has been computed using the sum-over-states formula with the minimal gauge choice (4.3.19). In terms of runtime we expect only negligible differences between the gauge choices. Moreover, since the time for this contribution is much smaller compared to the time required to solve the Sternheimer equation, we do not report a detailed performance comparison on this step in the following.

### 4.4.1 Insulators and semiconductors

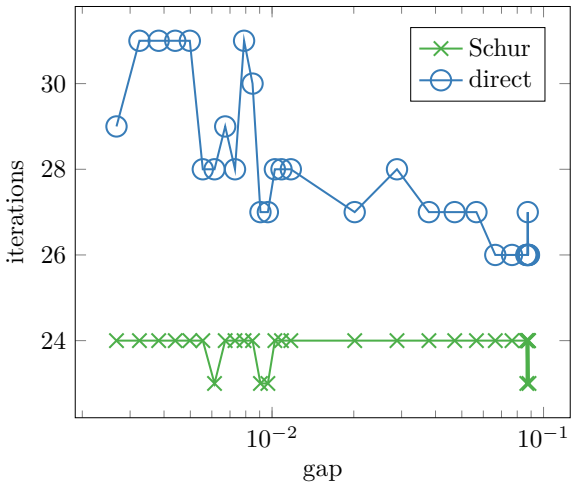



FIGURE 4.4 – Number of iterations of the linear solver for the Sternheimer equation for $n = N = 4$ *vs* the gap, with and without the Schur complement (4.3.26).

For insulators and semiconductors the gap between occupied and virtual states is usually large. One would therefore not expect a large gain from using the Schur complement (4.3.26) when computing $\delta\phi_n = \delta\phi_n^Q$. However, for distorted semiconductor structures or semiconductors with defects the gap can be made arbitrarily small, such that one would expect to see the Schur complement approach to be in the advantage. We test this using an FCC Silicon crystal for which we increase the lattice constant from 10 bohrs to 11.4 bohrs to artificially decrease and eventually close the gap. All calculations have been performed using a cut-off energy of $E_{\text{cut}} = 50$ Ha and a single $k$-point (the $\Gamma$-point). In Figure 4.4 we plot the number of iterations required for the linear solver of the Sternheimer equation to converge, for $n = N = 4$. Using the Schur complement the number of iterations stays almost constant even when the gap decreases. In contrast, with a direct approach, the linear solver requires about 30% more iterations near the closing gap.

[1] https://github.com/gkemlin/response-calculations-metals

### 4.4.2 Metals

The real advantage of using the Schur complement (4.3.26) instead of directly solving the Sternheimer equation (4.3.21) becomes apparent when computing response properties for metals at finite temperature. We use a standard heuristic which suggests to fully converge 20% more bands than the number of electrons pairs of the system, with 3 additional extra bands that are not converged by the successive eigensolvers of the SCF. We then select the "occupied" orbitals with an occupation threshold of $10^{-8}$. In addition to the number of iterations, we also compare the cost of the response calculations with and without the Schur complement (4.3.26). For this we consider the total number of Hamiltonian applications which was required to compute the response $\delta\rho$. For the small to medium-sized systems we consider here, the Hamiltonian-vector-product is the most expensive step in an DFT calculation and thus provides a representative cost indicator. Notice that both the implementation of the Schur complement and the direct method require exactly one Hamiltonian application per iteration of the CG. Additionally the Schur approach requires the computation of $H_\rho\widetilde{\phi}$, which is only a negligible additional cost as this is only needed once per $k$-point.

#### Aluminium

We start by considering an elongated aluminium supercell with 40 atoms. We use a cut-off energy $E_{\text{cut}} = 40$ Ha, a temperature $T = 10^{-3}$ Ha with Fermi–Dirac smearing and a $3 \times 3 \times 1$ discretization of the Brillouin zone. To ensure convergence of the SCF iterations we employ the Kerker preconditioner [112]. Since the system has 120 electrons per unit cell our usual heuristic converges 72 bands up to the tolerance of the eigensolver accompanied by 3 bands, which are not fully converged.

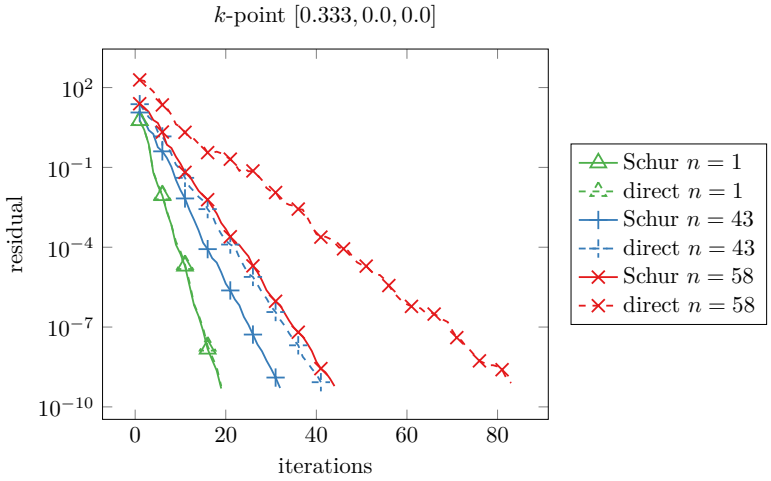



FIGURE 4.5 – Convergence of the Sternheimer solver for three different orbitals for $Al_{40}$. Each curve represents the convergence of the CG which solves the Sternheimer equation for one orbital: those with the slowest convergence are associated to the occupied orbitals with the highest energy.

| $k$-point – coordinate | $1 - [0, 0, 0]$ | $2 - [1/3, 0, 0]$ | $5 - [1/3, 1/3, 0]$ |
|---|---|---|---|
| $N$ | 69 | 58 | 67 |
| $\varepsilon_{N+1} - \varepsilon_N$ | 0.0320 | 0.0134 | 0.0217 |
| #iterations $n = N$ Schur | 48 | **44** | 41 |
| #iterations $n = N$ direct | 56 | **83** | 58 |

TABLE 4.1 – Convergence data for $k$-points 1, 2 and 5 for $Al_{40}$. Other $k$-points are not displayed but they all behave as one of these by symmetry. $N$ is the number of occupied bands, for an occupation threshold of $10^{-8}$.

The convergence behaviour when solving the Sternheimer equation for $k$-points of particular interest is shown in Table 4.1 and Figure 4.5. As expected, for $k$-points with a small difference $\varepsilon_{N+1} - \varepsilon_N$, the Schur complement (4.3.26) brings a noteworthy improvement with roughly 50% fewer iterations required to achieve convergence. Since the system we consider here has numerous occupied bands – between 60

and 70 depending on the $k$-point – most bands already feature a well-conditioned Sternheimer equation. Considering the cost for computing the total response, the Schur approach therefore overall only achieves a reduction by 17% in the number of Hamiltonian applications, from about $17,800$ (direct) to $14,800$ (with Schur). However, it should be noted that this improvement essentially comes for free as the extra bands are anyway provided by the SCF computation as a byproduct.

### Heusler system

Next we study the response calculation of Heusler-type transition-metal alloys. We focus mainly on the $Fe_2MnAl$ system but other compounds, such as the $Fe_2CrGa$ and CoFeMnGa alloy systems, have been tested and similar results were obtained. Heusler alloys are of considerable practical interest due to their rich and unusual magnetic and electronic properties. For instance, $Fe_2MnAl$ shows halfmetallic behaviour: the majority spin channel (denoted by $\uparrow$) behaves like a metal whereas the minority spin channel (denoted by $\downarrow$) behaves like an insulator as it has a vanishing density of states at the Fermi level. See [99], and reference therein, for more details as well as an analysis of the SCF convergence on such systems. For these systems we use a cut-off $E_{\text{cut}} = 45$ Ha, a temperature $T = 10^{-2}$ Ha with Gaussian smearing and a $13 \times 13 \times 13$ discretization of the Brillouin zone. The SCF was converged using a Kerker preconditioner [112]. Moreover, as we deal with a spin-polarized system, the numerical simulation slightly differs. The orbitals $\phi^{\sigma}_{(n,\boldsymbol{k})}$ and the occupation numbers $f^{\sigma}_{(n,\boldsymbol{k})}$ depend on the spin orientation $\sigma \in \{\uparrow, \downarrow\}$ and the $f^{\sigma}_{(n,\boldsymbol{k})}$'s belong to $[0, 1)$ instead of $[0, 2)$. Furthermore we modify the heuristic to determine the number of bands to be computed: $Fe_2MnAl$ has $N_{\text{el}} = 50$ electrons per unit cell and we use $25 + 0.2 \times 50 = 35$ fully converged bands per $k$-point, complemented by 3 additional bands, which are not checked for convergence.

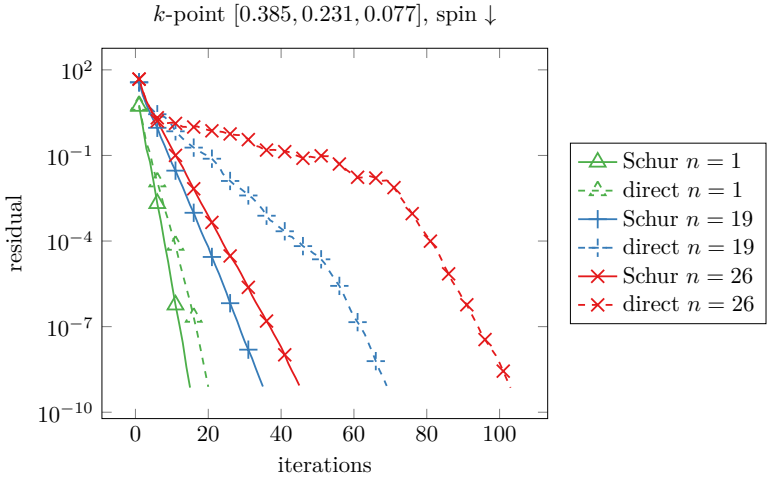



FIGURE 4.6 – Convergence of the Sternheimer solver for three different orbitals for $Fe_2MnAl$. Each curve represents the convergence of the CG which solves the Sternheimer equation for one orbital: those with the slowest convergence are associated to the occupied orbitals with the highest energy.

| spin channel | $\uparrow$ | $\downarrow$ |
|---|---|---|
| $N$ | 28 | 26 |
| $\varepsilon_{N+1} - \varepsilon_N$ | 0.0423 | 0.0154 |
| #iterations $n = N$ Schur | **45** | **45** |
| #iterations $n = N$ direct | **86** | **103** |

TABLE 4.2 – Convergence data for the two spin channels of the $k$-point with reduced coordinates $[0.385, 0.231, 0.077]$ for $Fe_2MnAl$. $N$ is the number of occupied bands, for an occupation threshold of $10^{-8}$.

We show in Table 4.2 and Figure 4.6 the results for the two spin channels of the $k$-point with reduced coordinates $[0.385, 0.231, 0.077]$. The other $k$-points behave similarly. Since both channels feature a small difference $\varepsilon_{N+1} - \varepsilon_N$ using the Schur complement (4.3.26) to solve the Sternheimer equation has a significant impact: for the orbitals with highest energy it reduces the number of iterations by half. For the

direct approach we notice a plateau where the solver encounters difficulties to converge the Sternheimer equation for the $N$-th orbital due to the small gap.

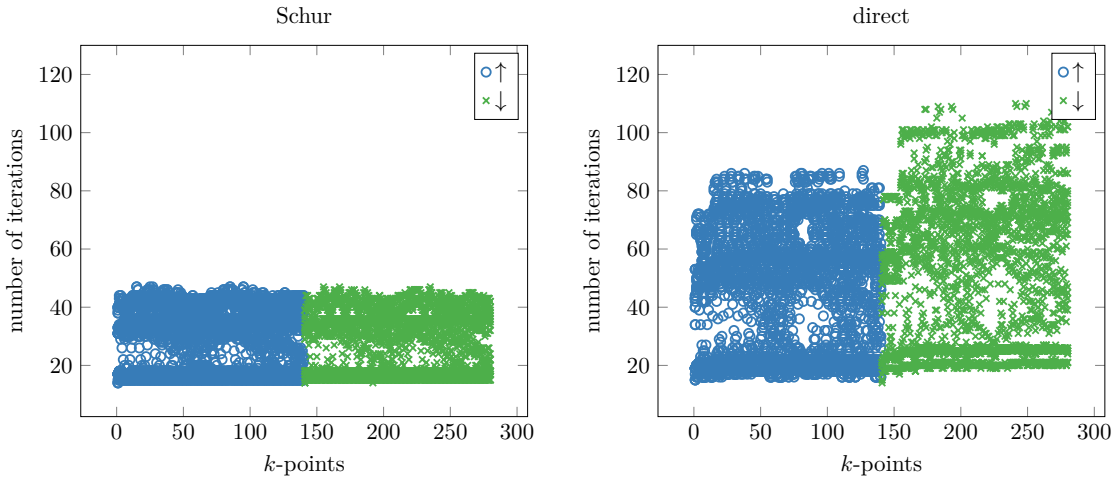



FIGURE 4.7 – Histogram of the number of iterations of the CG to solve the Sternheimer equation, with and without the Schur complement (4.3.26). On the $x$-axis, the $k$-point index number: the first 140 (blue $\circ$) have spins up, and the last 140 (green $\times$) have the same coordinates but with spins down. For each of these $k$-points, we plot the number of iterations for every occupied band of the $k$-point.

Unlike the aluminium case the improvements observed for the Heusler alloys are not restricted to a small number of bands. In Figure 4.7 we contrast the number of iterations required to solve the Sternheimer equation for every band at every $k$-point with and without using the Schur complement. Notice that lattice symmetries allow to reduce the number of explicitly treated $k$-points to 140 albeit we are using a $13 \times 13 \times 13$ $k$-point grid. In terms of the total number of Hamiltonian applications required for the response calculation, the Schur complement achieves a reduction by roughly 40%, from around $344,000$ (without Schur) to $208,000$ (with Schur). It should be noted that in this system the standard heuristic caused a large portion of the available extra bands to be fully converged, thus providing an ideal setting for the Schur complement approach to be effective. For example for the $k$-point discussed in Table 4.2 seven extra bands have been fully converged and an additional three partially. Given the enormous importance of Heusler systems and the known numerical difficulties for computing response properties in these systems, our result is encouraging and motivates the development of a more economical heuristic for choosing the number of converged bands in future work.

### 4.4.3 Comparison to shifted Sternheimer approaches

In the literature other strategies for computing $\delta\rho$ have been reported. We briefly consider the approach proposed in [14], where the response is computed as

$$\delta\rho(\boldsymbol{r}) = \sum_{n=1}^{N} 2\phi_n^*(\boldsymbol{r})\delta\phi_n(\boldsymbol{r}) - f_n'\delta\varepsilon_{\mathrm{F}}|\phi_n(\boldsymbol{r})|^2. \tag{4.4.1}$$

Instead of splitting $\delta\phi_n$ into two contributions, the full $\delta\phi_n$ is computed for all $n \leqslant N$ by solving the *shifted* Sternheimer equation

$$(H_\rho + S - \varepsilon_n)\delta\phi_n = -(f_n - S_n)\delta V\phi_n. \tag{4.4.2}$$

Here $S : \mathrm{Ran}\,(P) \to \mathrm{Ran}\,(P)$ is a shift operator acting on the space of occupied orbitals, chosen so that the linear system is nonsingular (for any $n \leqslant N$, $H_\rho - \varepsilon_n$ is not invertible). Then, $S_n$ is chosen for every $n \leqslant N$ such that $\delta\rho$ from (4.4.1) satisfies (4.2.15). However, as $S$ only acts on $\mathrm{Ran}\,(P)$, equation still becomes badly conditioned if $\varepsilon_{N+1} - \varepsilon_N$ is too small. This becomes apparent when solving the shifted Sternheimer equation (4.4.2) for the $Fe_2MnAl$ system, see Figure 4.8. For the orbital responses of the highest-energy occupied bands the CG iterations on the shifted Sternheimer equation converge very slowly – in contrast to the Schur complement approach (4.3.26) we proposed in this work. In terms of the number of Hamiltonian applications, the shifted Sternheimer strategy required around $492,000$ applications versus $208,000$ for the Schur complement approach.

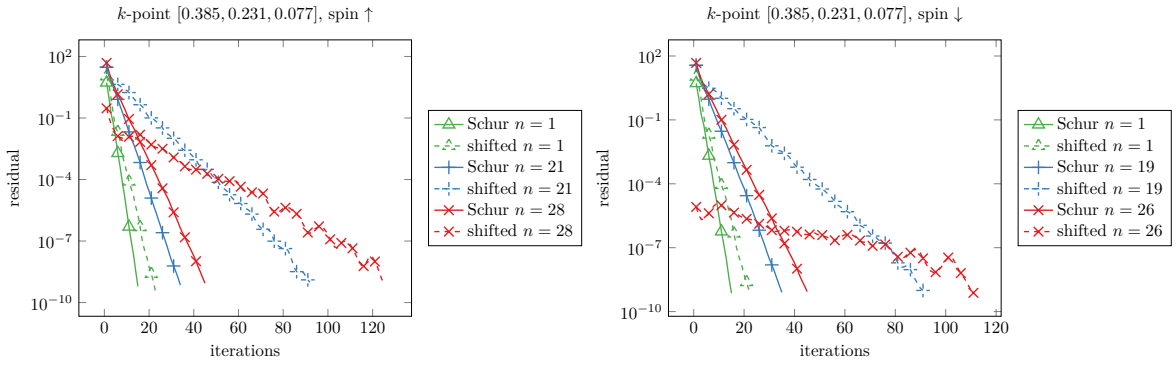



FIGURE 4.8 – Resolution of the Sternheimer equation for both spin channels of one specific $k$-point for the $Fe_2MnAl$ system, with the Schur approach (4.3.26) and the shifted approach (4.4.2). Note that for this particular $k$-point, the spin ↓ channel has a starting point with already small residual for the highest occupied energy level. This is due to the term $f_n$ appearing in (4.4.2), and the convergence is still slow.

## 4.5 Conclusion

In density functional theory, the simulation of many physical properties requires the computation of the response of the ground-state density to an external perturbation. In this work we have reviewed the standard formalism of such response calculations from the point of view of numerical analysis. We provided an overview of the possible gauge choices for representing the density response, summarizing and contrasting the approaches employed by state-of-the-art codes such as Quantum Espresso [78] or Abinit [175] in a common framework.

Based on our analysis we furthermore suggested two novel approaches for DFT response calculations. For the occupied-occupied part of the response we developed a gauge choice based on the idea to maximize numerical stability in the involved sums by minimizing the numerical range of the individual orbital contributions. For the occupied-unoccupied part of the response we suggested a novel approach to solving the Sternheimer equation based on a Schur complement. Key idea of this approach is to make use of the additional (partially) converged bands, which are available as a byproduct from the preceding self-consistent field (SCF) procedure (which yields the ground-state density). Without additional computational effort this allows to improve the conditioning of the Sternheimer equation and thus accelerate its convergence. We demonstrated this numerically on a number of practically relevant problems, including response calculations on small-gapped semiconductors, elongated metallic slabs or numerically challenging Heusler alloy systems. Overall the Schur complement approach allowed to obtain a converged response saving up to 40% in the required Hamiltonian applications – the cost-dominating step in small to medium-sized DFT problems. For larger systems we similarly expect savings from introducing a Schur complement technique, even though algorithms commonly employ different trade-offs.

In this work we followed standard heuristics for selecting the number of extra bands to employ in the SCF calculations and thus the number of additional bands available when solving the response problem. However, our results emphasize the need for a more robust understanding between the computed number of bands and the observed rate of convergence. We have provided some initial ideas for such an analysis in the appendix, but leave a more exhaustive discussion for future work.

## Appendix: Choosing the number of extra bands

In this chapter, we saw through various numerical examples that using a Schur complement to compute the unoccupied-occupied contributions to the orbitals' response improves the convergence of the Sternheimer equation. In this appendix, we quantify this acceleration and discuss how this idea can be used to select the number of bands to be computed. Considering the straight convergence curves from Figures 4.5–4.6 suggest that the convergence of the CG is indeed led by the square root of the condition number of the system matrix (see [184, Section 9]) when using the Schur complement. Key idea will thus be to estimate the condition number of the linear system (4.3.26).

## Numerical analysis

To analyse the condition number of the Schur complement, we consider the following specific setting

$$\begin{cases} H_\rho \phi_n = \varepsilon_n \phi_n, \quad \varepsilon_1 \leqslant \varepsilon_2 \leqslant \cdots \\ \langle \phi_n, \phi_m \rangle = \delta_{nm}, \end{cases} \tag{4.5.1}$$

where $H_\rho \in \mathbb{C}^{N_\mathrm{b} \times N_\mathrm{b}}$ is typically the discretized self-consistent Hamiltonian of the system, at some $k$-point. We assume that we have $N$ occupied orbitals that have an occupation number higher than the threshold we fixed and that we have $N_\mathrm{ex}$ extra bands, as explained in Section 4.3.3. In summary, we have at our disposal $N + N_\mathrm{ex}$ bands in total: $\Phi = (\phi_1, \ldots, \phi_N)$ are occupied, fully converged bands and the extra bands $\Phi_\mathrm{ex}^\ell = (\phi_{N+1}^\ell, \ldots, \phi_{N+N_\mathrm{ex}}^\ell)$ are not necessarily all converged. We added here the exponent $\ell$ as we make the following assumptions:

- for any $\ell \in \mathbb{N}$, $(\Phi, \Phi_\mathrm{ex}^\ell)$ is an orthonormal family;
- for any $\ell \in \mathbb{N}$, $(\Phi_\mathrm{ex}^\ell)^* H_\rho \Phi_\mathrm{ex}^\ell \in \mathbb{C}^{N_\mathrm{ex} \times N_\mathrm{ex}}$ is a diagonal matrix whose elements are labelled $\varepsilon_m^\ell := \langle \phi_m^\ell, H_\rho \phi_m^\ell \rangle$ for $N + 1 \leqslant m \leqslant N + N_\mathrm{ex}$;
- as $\ell \to +\infty$, $(\phi_m^\ell, \varepsilon_m^\ell) \to (\phi_m, \varepsilon_m)$.

All these assumptions are satisfied for instance if the sequence $(\Phi, \Phi_\mathrm{ex}^\ell)_{l \in \mathbb{N}}$ is generated by any Rayleigh–Ritz based eigensolver (for instance the LOBPCG eigensolver [115]), which is the case by default in DFTK. For every $\ell$, we can thus decompose the plane-wave approximation space $\mathcal{H} = X_{N_\mathrm{b}}$ (with $N_\mathrm{b} \gg N + N_\mathrm{ex}$) in two different ways:

$$\mathcal{H} = \mathrm{Ran}\,(P) \oplus \mathrm{Ran}\,(T) \oplus \mathrm{Ran}\,(R) \quad \text{and} \quad \mathcal{H} = \mathrm{Ran}\,(P) \oplus \mathrm{Ran}\,(T^\ell) \oplus \mathrm{Ran}\,(R^\ell), \tag{4.5.2}$$

where

$$P := \sum_{n=1}^{N} \phi_n \phi_n^* \quad \text{and} \quad \begin{cases} T := \sum_{n=N+1}^{N_\mathrm{ex}} \phi_n \phi_n^*, \quad R := 1 - P - T \\ T^\ell := \sum_{n=N+1}^{N_\mathrm{ex}} \phi_n^\ell (\phi_n^\ell)^*, \quad R^\ell := 1 - P - T^\ell, \end{cases} \tag{4.5.3}$$

are all orthogonal projectors. In these two decompositions, $H_\rho$ has the associated block representations:

$$H_\rho = \begin{pmatrix} E & 0 & 0 \\ 0 & E_\mathrm{ex} & 0 \\ 0 & 0 & \ddots \end{pmatrix} \quad \text{and} \quad H_\rho = \begin{pmatrix} E & 0 & 0 \\ 0 & E_\mathrm{ex}^\ell & R^\ell H_\rho T^\ell \\ 0 & T^\ell H_\rho R^\ell & R^\ell H_\rho R^\ell \end{pmatrix} \tag{4.5.4}$$

where $E := \mathrm{Diag}(\varepsilon_1, \ldots, \varepsilon_n)$, $E_\mathrm{ex} := \mathrm{Diag}(\varepsilon_{N+1}, \ldots, \varepsilon_{N+N_\mathrm{ex}})$ and $E_\mathrm{ex}^\ell := \mathrm{Diag}(\varepsilon_{N+1}^\ell, \ldots, \varepsilon_{N+N_\mathrm{ex}}^\ell)$ are diagonal matrices. Moreover, note that as $\Phi_\mathrm{ex}^\ell \to \Phi_\mathrm{ex}$, the residuals $R^\ell H_\rho T^\ell$ converge to 0.

Now, we fix $n \leqslant N$ and we compute the condition number of the linear system (4.3.26). Enforcing the CG to stay at each iteration in $\mathrm{Ran}\,(R^\ell)$, this condition number is given by the ratio of the largest and smallest nonzero eigenvalues of

$$H_n^\ell + X_n^\ell, \tag{4.5.5}$$

where

$$\begin{aligned} H_n^\ell := R^\ell (H_\rho - \varepsilon_n) R^\ell \quad \text{and} \quad X_n^\ell &= -R^\ell (H_\rho - \varepsilon_n) \Phi_\mathrm{ex}^\ell (E_\mathrm{ex}^\ell - \varepsilon_n)^{-1} (\Phi_\mathrm{ex}^\ell)^* (H_\rho - \varepsilon_n) R^\ell \\ &= -R^\ell H_\rho \Phi_\mathrm{ex}^\ell (E_\mathrm{ex}^\ell - \varepsilon_n)^{-1} (\Phi_\mathrm{ex}^\ell)^* H_\rho R^\ell. \end{aligned} \tag{4.5.6}$$

Here $E_\mathrm{ex}^\ell - \varepsilon_n$ is diagonal and thus explicitly invertible if $\ell$ is large enough as $\varepsilon_{N+1}^\ell \to \varepsilon_{N+1} > \varepsilon_N \geqslant \varepsilon_n$. We focus for the moment on the smallest nonzero eigenvalue, that is $\varepsilon_{N+N_\mathrm{ex}+1}^\ell - \varepsilon_n$. The condition number being proportional to the inverse of the smallest eigenvalue, we now derive a lower bound of $\varepsilon_{N+N_\mathrm{ex}+1}^\ell - \varepsilon_n$ in order to get an upper bound on the condition number of (4.5.5). When $\ell \to +\infty$, we have $X_n^\ell \to 0$ (as $R H_\rho \Phi_\mathrm{ex} = 0$) and $H_n^\ell \to H_n := R(H_\rho - \varepsilon_n)R$ whose smallest nonzero eigenvalue is $\varepsilon_{N+N_\mathrm{ex}+1} - \varepsilon_n$. We use next a perturbative approach to effectively approximate the condition number of (4.5.5).

We use a standard eigenvalue perturbation result, whose proof is recalled for the sake of completeness. It is directly adapted from the general case of self-adjoint bounded below operators with symmetric perturbations studied for instance in [71].

**Proposition 4.1.** *Let $N \in \mathbb{N}$, $H_0, W \in \mathbb{C}^{N\times N}$ be Hermitian matrices and $\alpha \geqslant 0$ such that $H_0 + \alpha > 0$. Then, the eigenvalues of $H \coloneqq H_0 + W$ and $H_0$ satisfy*

$$|\nu_i(H) - \nu_i(H_0)| \leqslant (\nu_i(H_0) + \alpha)\|W\|_{H_0,\alpha}, \tag{4.5.7}$$

*where $\|W\|_{H_0,\alpha}$ is the operator norm of $(H_0+\alpha)^{-1/2}W(H_0+\alpha)^{-1/2}$ and $\nu_i(A)$ is the $i$-th lowest eigenvalue of the matrix $A$.*

*Proof.* Let $u \in \mathbb{C}^N$ and define $v \coloneqq (H_0 + \alpha)^{1/2}u$. Then,

$$\begin{aligned} |\langle u, Hu\rangle - \langle u, H_0u\rangle| = |\langle u, Wu\rangle| &= \left|\left\langle v, (H_0 + \alpha)^{-1/2}W(H_0 + \alpha)^{-1/2}v\right\rangle\right| \\ &\leqslant \|W\|_{H_0,\alpha}\langle v, v\rangle = \|W\|_{H_0,\alpha}\langle u, (H_0 + \alpha)u\rangle. \end{aligned} \tag{4.5.8}$$

Therefore,

$$(1-\|W\|_{H_0,\alpha})\langle u, H_0u\rangle - \alpha\|W\|_{H_0,\alpha}\langle u, u\rangle \leqslant \langle u, Hu\rangle \leqslant (1+\|W\|_{H_0,\alpha})\langle u, H_0u\rangle + \alpha\|W\|_{H_0,\alpha}\langle u, u\rangle. \tag{4.5.9}$$

The min-max theorem then yields for $i = 1, \dots, N$,

$$(1 - \|W\|_{H_0,\alpha})\nu_i(H_0) - \alpha\|W\|_{H_0,\alpha} \leqslant \nu_i(H) \leqslant (1 + \|W\|_{H_0,\alpha})\nu_i(H_0) + \alpha\|W\|_{H_0,\alpha}, \tag{4.5.10}$$

which gives the desired inequality. □

In our case, we can apply this result to

$$H_n^\ell + X_n^\ell = H_n + (H_n^\ell - H_n) + X_n^\ell, \tag{4.5.11}$$

with $H_0 = H_n$, $W = W_n^\ell \coloneqq (H_n^\ell - H_n) + X_n^\ell$ and $\alpha = \varepsilon_{N+N_{\text{ex}}+1} - \varepsilon_n > 0$. Proposition 4.1 applied to the $(N + N_{\text{ex}} + 1)$-th eigenvalues then yields

$$\varepsilon^\ell_{N+N_{\text{ex}}+1} - \varepsilon_n \geqslant (\varepsilon_{N+N_{\text{ex}}+1} - \varepsilon_n)\left(1 - 2\left\|W_n^\ell\right\|_{H_n,\varepsilon_{N+N_{\text{ex}}+1}-\varepsilon_n}\right) \approx (\varepsilon_{N+N_{\text{ex}}+1} - \varepsilon_n), \tag{4.5.12}$$

where we assume that $2\left\|W_n^\ell\right\|_{H_n,\varepsilon_{N+N_{\text{ex}}+1}-\varepsilon_n}$ is small enough to be negligible with respect to 1, which is the case if the extra states are sufficiently converged. Now, if this bound is valid in theory, in practice we do not have access to $\varepsilon_{N+N_{\text{ex}}+1}$ as we work with $N + N_{\text{ex}}$ bands only. However, up to loosing sharpness, we can use that $\varepsilon_{N+N_{\text{ex}}+1} \geqslant \varepsilon_{N+N_{\text{ex}}}$ where $\varepsilon_{N+N_{\text{ex}}}$ can be estimated using the last extra band. Indeed, using for instance the Bauer–Fike bound ([100, Theorem 1] or [178]), we obtain

$$\varepsilon_{N+N_{\text{ex}}} \geqslant \varepsilon^\ell_{N+N_{\text{ex}}} - \left\|r^\ell_{N+N_{\text{ex}}}\right\|, \tag{4.5.13}$$

where $r^\ell_{N+N_{\text{ex}}}$ is the residual associated to the last extra band. Of course, this estimate is not sharp as we expect the error on the eigenvalue to behave as the square of the residual, but this requires to estimate the gap to the rest of the spectrum, see for instance the Kato–Temple bound [100, Theorem 2]. In the end, we have the following lower bound for $\varepsilon^\ell_{N+N_{\text{ex}}+1} - \varepsilon_n$:

$$\varepsilon^\ell_{N+N_{\text{ex}}+1} - \varepsilon_n \geqslant \left(\varepsilon^\ell_{N+N_{\text{ex}}} - \varepsilon_n - \left\|r^\ell_{N+N_{\text{ex}}}\right\|\right) \approx \varepsilon^\ell_{N+N_{\text{ex}}} - \varepsilon_n, \tag{4.5.14}$$

where we assume again that $\left\|r^\ell_{N+N_{\text{ex}}}\right\|$ is small enough with respect to $\varepsilon^\ell_{N+N_{\text{ex}}} - \varepsilon_n$.

We can now derive an upper bound on $\kappa_n^\ell$, the condition number of (4.5.5). It is given by the ratio of its highest eigenvalue and $\varepsilon^\ell_{N+N_{\text{ex}}+1} - \varepsilon_n$. Since the Laplace operator is the higher-order term in the Kohn–Sham Hamiltonian, the highest eigenvalue is, as usually in plane-wave simulations, of order $E_{\text{cut}}$. With proper kinetic preconditioning, we can assume that its contribution to the condition number of the linear system is constant with respect to $E_{\text{cut}}$ and $n$ so that, finally,

$$\kappa_n^\ell \lesssim \frac{C}{\varepsilon^\ell_{N+N_{\text{ex}}+1} - \varepsilon_n} \lesssim \frac{C}{\varepsilon^\ell_{N+N_{\text{ex}}} - \varepsilon_n}. \tag{4.5.15}$$

Therefore the condition number is bounded from above by $C/(\varepsilon^{\ell}_{N+N_{\text{ex}}} - \varepsilon_n)$ to first-order. The number of CG iterations to solve the linear system (4.3.26) with a given accuracy is then proportional to the square root of the condition number of the matrix (4.5.5) (see [184]):

$$\sqrt{\kappa^{\ell}_n} \lesssim \sqrt{\frac{C}{\varepsilon^{\ell}_{N+N_{\text{ex}}} - \varepsilon_n}}. \tag{4.5.16}$$

Note that this upper bound is valid provided that the extra bands are converged enough, not necessarily fully, and proper kinetic preconditioning is employed.

Estimate (4.5.16) leads, as expected, to the qualitative conclusion that the more extra bands we use, the higher the difference $\varepsilon^{\ell}_{N+N_{\text{ex}}} - \varepsilon_n$ and the faster the convergence. However, note that it is not possible to evaluate directly the convergence speed as the constant $C$ is *a priori* unknown, in particular if we use preconditioners.

## An adaptive strategy to choose the number of extra bands

The main bottleneck of (4.5.16) is the estimation of the constant $C$. However, one can reasonably assume that this constant does not depend too much on $n$, so that the ratio between the number of iterations to reach convergence between the last occupied band ($n = N$) and the first band ($n = 1$) can be estimated by

$$\xi^{\ell}_{N_{\text{ex}}} := \sqrt{\frac{\varepsilon^{\ell}_{N+N_{\text{ex}}} - \varepsilon_1}{\varepsilon^{\ell}_{N+N_{\text{ex}}} - \varepsilon_N}}, \tag{4.5.17}$$

This ratio can be of interest as (4.5.16) suggests that the Sternheimer solver converges the fastest for $n = 1$ and the slowest for $n = N$.

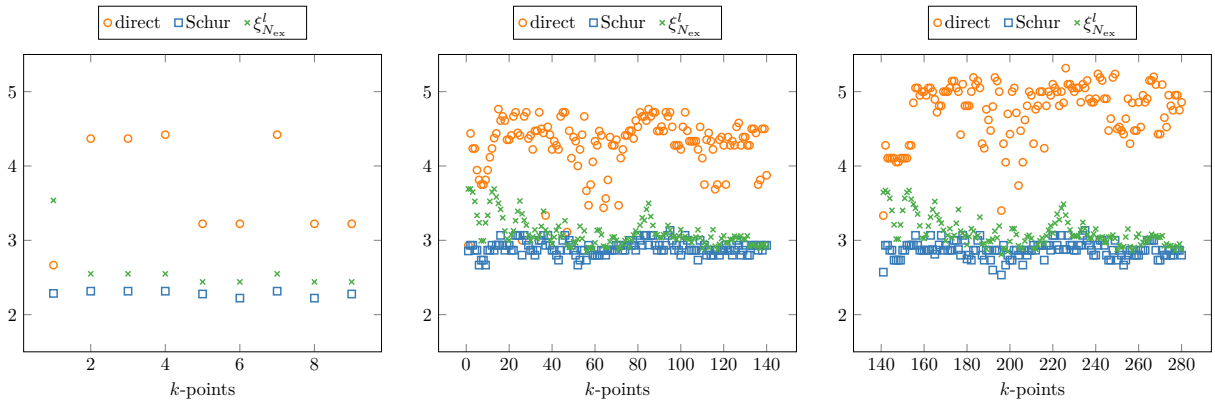



FIGURE 4.9 – Comparison between the ratio $\xi^{\ell}_{N_{\text{ex}}}$ (×) and the ratios of the experimental number of iterations between the first and last occupied bands, with (□) and without (○) the Schur complement (4.3.26). On the $x$-axis is the index of the $k$-point. [Left] $Al_{40}$ [Middle] $Fe_2MnAl$ spin up channels [Right] $Fe_2MnAl$ spin down channels.

We plot in Figure 4.9 the upper bound $\xi^{\ell}_{N_{\text{ex}}}$ as well as the computed ratios between the number of iterations of the first and last bands for the systems we considered in Section 4.4. These plots show that $\xi^{\ell}_{N_{\text{ex}}}$ is indeed an upper bound of the actual ratio. This bound does not seem to be sharp however. This is due to the successive approximations we made to obtain this estimate. Plots in Figure 4.9 also confirm that if, for every $k$-point, the ratio of the number of iterations between the first and last occupied bands is assumed to be an accurate indicator of the efficiency of the Sternheimer solver, then using the Schur complement (4.3.26) always make this ratio smaller.

If we want the ratio of the number of iterations between the first and the last occupied bands to be lower than some target ratio $\xi_T$ (for instance 3), Figure 4.9 suggests that the computable ratio $\xi^{\ell}_{N_{\text{ex}}}$ can help in choosing the number of extra bands to reach this target ratio. We propose in Algorithm 4.1 an adaptive algorithm to select the number of extra bands as a post-processing step after termination of the SCF. The basic idea is that, given the initial output $(\Phi, \Phi^{\ell}_{\text{ex}})$ with $\ell = 0$ of an SCF calculation, one iterates $\Phi^{\ell}_{\text{ex}} \to \Phi^{\ell+1}_{\text{ex}}$ where $\Phi^{\ell}_{\text{ex}}$ gathers the extra bands. At each iteration $\ell$, we compute $\xi^{\ell}_{N_{\text{ex}}}$ and check

if it is below the target ratio. If not, we compute more approximated eigenvectors, that we converge until the residual $\|r^\ell_{N+N_{\text{ex}}}\|$ is negligible with respect to $\varepsilon^\ell_{N+N_{\text{ex}}} - \varepsilon_N$, and so on. To generate such a residual, after adding a random extra band properly orthonormalized, we update the extra bands using a LOBPCG with tolerance

$$\text{tol} = (\varepsilon^\ell_{N+N_{\text{ex}}-1} - \varepsilon_N)/50. \tag{4.5.18}$$

Note that we use $\varepsilon^\ell_{N+N_{\text{ex}}-1}$ instead of $\varepsilon^\ell_{N+N_{\text{ex}}}$: this is done for the sake of simplicity, instead of updating the tolerance on the fly with $\varepsilon^\ell_{N+N_{\text{ex}}}$ changing at each iteration of the LOBPCG.

ALGORITHM 4.1 – Adaptive choice of the number of extra bands

**Data:** target ratio $\xi_T$, $N_{\text{ex}}$, $\ell$, $\xi^\ell_{N_{\text{ex}}}$
**while** $\xi^\ell_{N_{\text{ex}}} > \xi_T$ **do**
  add random extra band $\phi_{\text{new}}$ in the orthogonal of Span $(\Phi, \Phi^\ell_{\text{ex}})$;
  $N_{\text{ex}} \leftarrow N_{\text{ex}} + 1$;
  update on the fly the extra bands with tolerance from (4.5.18) using the LOBPCG method;
  $\Phi^{\ell+1}_{\text{ex}} \leftarrow (\Phi^\ell_{\text{ex}}, \phi_{\text{new}})$ and $E^{\ell+1}_{\text{ex}} \leftarrow (\Phi^{\ell+1}_{\text{ex}})^* H_\rho \Phi^{\ell+1}_{\text{ex}}$;
  $\ell \leftarrow \ell + 1$;
  compute $\xi^\ell_{N_{\text{ex}}}$ with (4.5.17);
**end**

## Numerical tests

We test this strategy on the systems investigated in Section 4.4, with different values for the target ratio $\xi_T$ in order to see a noticeable improvement for each system. In practice, we suggest this ratio to be between 2 and 3.

We first start with the $Al_{40}$ system. Figure 4.9 [Left] suggests that the default choice of extra bands already gives satisfying results by reaching a ratio of approximately 2.5 for all $k$-points but the $\Gamma$-point (for which there is no real issue with the Sternheimer equation, according to Table 4.1). We thus run Algorithm 4.1 with a smaller target ratio $\xi_T = 2.2$. We use as initial value for $N_{\text{ex}}$ the default value for each $k$-point. Results are plotted in Table 4.3 [Left] and suggests adding 15 extra bands. Running again the simulations from Section 4.4 with 72 fully converged bands and 18 additional, not fully converged, bands yields indeed an improvement in the convergence of the CG when solving the Sternheimer equation with the Schur complement method. Moreover, in Figure 4.10, the ratio $\xi^\ell_{N_{\text{ex}}}$ indeed lies below the target ratio $\xi_T = 2.2$, and matches this ratio for the $k$-points that caused difficulties for the Sternheimer equation solver to converge. In terms of computational time, the number of Hamiltonian applications to compute the response has been reduced from $\sim 14,800$ with the default number of extra bands to $\sim 12,800$. However, running the algorithm required $\sim 3,400$ additional Hamiltonian applications, making the total amount of Hamiltonian applications higher than that of the Schur approach with the standard heuristic.

$Al_{40}$, $\xi_T = 2.2$

| $k$-point | 1 | 2 | 5 |
|---|---|---|---|
| $N$ | 69 | 58 | 67 |
| default $N_{\text{ex}}$ | 6 | 17 | 8 |
| suggested $N_{\text{ex}}$ | 21 | 29 | 12 |
| #iterations $n = 1$ Schur | 21 | 19 | 18 |
| #iterations $n = N$ Schur | 32 | 36 | 28 |

$Fe_2MnAl$, $\xi_T = 2.5$

| $k$-point / spin | 96 ↑ | 96 ↓ | 72 ↑ | 72 ↓ |
|---|---|---|---|---|
| $N$ | 28 | 26 | 29 | 26 |
| default $N_{\text{ex}}$ | 10 | 12 | 9 | 12 |
| suggested $N_{\text{ex}}$ | 16 | 18 | 17 | 20 |
| #iterations $n = 1$ Schur | 15 | 15 | 15 | 15 |
| #iterations $n = N$ Schur | 36 | 35 | 35 | 35 |

TABLE 4.3 – Suggested number of extra bands for $Al_{40}$ and $Fe_2MnAl$ to reach the target ratio $\xi_T$, obtained with Algorithm 4.1 with default $N_{\text{ex}}$ as starting point, as well as the number of iterations to reach convergence with the newly suggested $N_{\text{ex}}$. Note that the ratio between iterations indeed lies below the target ratio $\xi_T$.

Similarly, for $Fe_2MnAl$, we run Algorithm 4.1 with target ratio $\xi_T = 2.5$ as well as initial value the default $N_{\text{ex}}$ for all the 140 $k$-points and spin polarizations. We present in Table 4.3 [Right] the output for both spin polarizations of two particular $k$-points. The results are similar for the rest of the $k$-points and the maximum additional extra bands suggested by the algorithm is 8. We thus run the same simulations

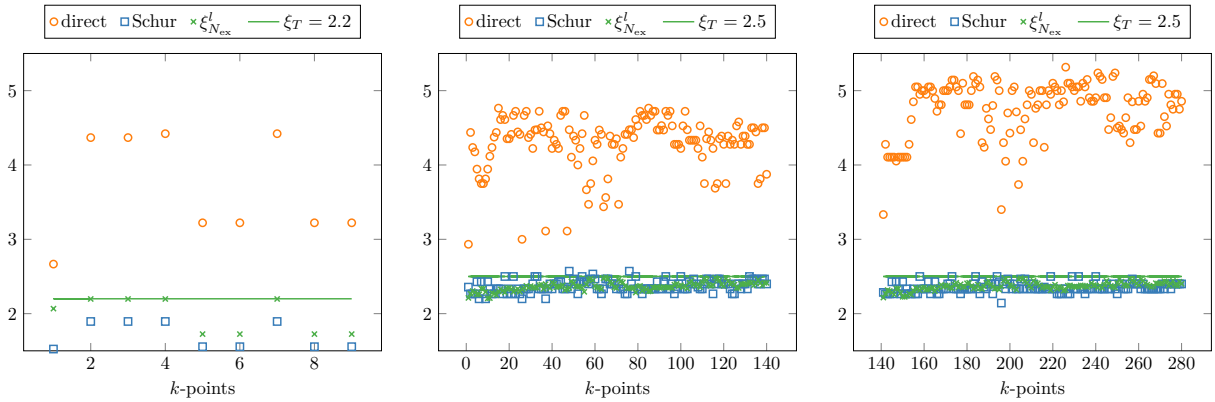



FIGURE 4.10 – Comparison between the ratio $\xi^{\ell}_{N_{\text{ex}}}$ ($\times$) and the ratios of the measured number of iterations between the first and last occupied bands, with ($\square$) and without ($\circ$) the Schur complement (4.3.26). On the $x$-axis is the index of the $k$-point. [Left] $Al_{40}$ with 15 additional extra bands. [Left] $Fe_2MnAl$ spin $\uparrow$ with 8 additional bands. [Right] $Fe_2MnAl$ spin $\downarrow$ with 8 additional bands.

as in Section 4.4 but this time with 35 fully converged bands and 11 extra, nonnecessarily converged, bands. We indeed see for these two $k$-points that the target ratio has been reached, and that the number of iterations to converge is smaller than for the default choice we made in Table 4.2. In Figure 4.10, we plot the ratios $\xi^{\ell}_{N_{\text{ex}}}$ as well as the actual ratios and they almost all lie below the target ratio. Contrarily to $Al_{40}$, we note however that the actual measured ratios are not always below the indicator $\xi^{\ell}_{N_{\text{ex}}}$. In terms of computational time, the number of Hamiltonian applications has been reduced from $\sim 208,000$ with the default choice of $N_{\text{ex}}$ to $\sim 179,000$. Again, running the algorithm required $\sim 49,000$ additional Hamiltonian applications, making it more expensive than using the default number of extra bands.

It appears that Algorithm 4.1 can be used to choose the number of extra bands in order to reach a given ratio $\xi_T$. However, using the algorithm as such is not useful in practice as it requires a too high number of Hamiltonian applications, making this strategy less interesting than the Schur approach we proposed with the default choice of extra bands. Strategies to reduce the number of Hamiltonian applications in order to choose an appropriate number of extra bands will be subject of future work.

Chapter 5

# A priori error analysis of linear and nonlinear periodic Schrödinger equations with analytic potentials

This chapter is a preliminary version of [GKip1], and is currently in preparation (to be submitted by the end of 2022):

*Eric Cancès, Gaspard Kemlin and Antoine Levitt. A priori analysis of linear and nonlinear periodic Schrödinger equations with analytic potentials.* https://arxiv.org/abs/2206.04954.

**Abstract** This chapter is concerned with the numerical analysis of linear and nonlinear Schrödinger equations with analytic potentials. While the regularity of the potential (and the source term when there is one) automatically conveys to the solution in the linear cases, this is no longer true in general in the nonlinear case. We also study the rate of convergence of the plane-wave (Fourier) discretization method for computing numerical approximations of the solution.

## Contents

**5.1 Introduction** . . . . . **110**
**5.2 Spaces of analytic functions** . . . . . **111**
**5.3 The linear case** . . . . . **112**
5.3.1 The linear elliptic problem . . . . . 112
5.3.2 The linear eigenvalue problem . . . . . 114
5.3.3 Plane-wave approximation of the linear Schrödinger equation . . . . . 114
**5.4 The nonlinear case: a counter-example** . . . . . **115**
**5.5 Extension to the multidimensional case with application to Kohn–Sham models.** . . . . . **121**

## 5.1 Introduction

Kohn–Sham density functional theory (KS-DFT) is currently the most popular model in quantum chemistry and materials science as it offers a good compromise between accuracy and computational efficiency. KS-DFT aims at computing, for a given configuration of the nuclei of the molecular system or material of interest, the electronic ground-state energy and density. From the latter, it is possible to compute the effective forces acting on the nuclei in this configuration, and thus to identify the (meta)stable equilibrium configurations of the system, or to simulate the dynamics of the molecular system in various thermodynamic conditions. In materials science applications, computations are commonly done in a periodic simulation cell, which can be either the unit cell of a crystal (for the special case of perfect crystals), or a supercell (for all the other cases: crystals with defects, disordered alloys, glassy materials, liquids...).

We denote by $\mathbb{L} = \mathbb{Z}\boldsymbol{a}_1 + \mathbb{Z}\boldsymbol{a}_2 + \mathbb{Z}\boldsymbol{a}_3$ the periodic lattice, where $(\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3)$ is a nonnecessarily orthonormal basis of $\mathbb{R}^3$, and by $\Omega = [0,1)\boldsymbol{a}_1 + [0,1)\boldsymbol{a}_2 + [0,1)\boldsymbol{a}_3$ the simulation cell. Let us denote by

$$\mathrm{L}^2_{\#,\mathbb{L}} \coloneqq \{u \in \mathrm{L}^2_{\mathrm{loc}}(\mathbb{R}^3, \mathbb{C}) \mid u \ \text{ is } \mathbb{L}\text{-periodic}\}$$

the Hilbert space of complex-valued $\mathbb{L}$-periodic locally square integrable functions on $\mathbb{R}^3$, endowed with its usual inner product. The KS-DFT equations read

$$H_\rho \varphi_i = \lambda_i \varphi_i, \quad (\varphi_i, \varphi_j)_{\mathrm{L}^2_{\#,\mathbb{L}}} = \delta_{ij}, \quad \rho(\boldsymbol{x}) = \sum_{i=1}^{N_\mathrm{p}} |\varphi_i(\boldsymbol{x})|^2, \tag{5.1.1}$$

where $H_\rho$ is the Kohn–Sham Hamiltonian, a self-adjoint operator on $\mathrm{L}^2_{\#,\mathbb{L}}$ bounded below and with compact resolvent. The $\varphi_i$'s are the Kohn–Sham orbitals, and the $\lambda_i$'s their energies. Since $H_\rho$ depends on $\rho$, which in turn depends on the eigenfunctions $\varphi_i$, $1 \leqslant i \leqslant N_\mathrm{p}$, (5.1.1) is a nonlinear eigenproblem. The parameter $N_\mathrm{p}$ represents physically the number of valence electron pairs per simulation cell and $\rho$ the ground-state electronic density. We assume here, and this is the case for most physical systems, that $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_{N_\mathrm{p}}$ are the lowest $N_\mathrm{p}$ eigenvalues of $H_\rho$ (*Aufbau* principle). The Kohn–Sham Hamiltonian with pseudopotentials reads

$$H_\rho = -\frac{1}{2}\Delta + V_\mathrm{nl} + V^\mathrm{Hxc}_{\mathrm{loc},\rho}$$

where $V_\mathrm{nl}$ is a finite-rank self-adjoint operator (the nonlocal part of the pseudopotential), and

$$V^\mathrm{Hxc}_{\mathrm{loc},\rho}(\boldsymbol{x}) = V_\mathrm{loc}(\boldsymbol{x}) + V_{\mathrm{H},\rho}(\boldsymbol{x}) + V_{\mathrm{xc},\rho}(\boldsymbol{x})$$

is a periodic real-valued function depending (nonlocally) on $\rho$. The function $V_\mathrm{loc}$ is the local component of the pseudopotential, the Hartree potential $V_{\mathrm{H},\rho}$ is the unique solution with zero mean to the periodic Poisson equation

$$-\Delta V_{\mathrm{H},\rho}(\boldsymbol{x}) = 4\pi\left(\rho(\boldsymbol{x}) - \frac{1}{|\Omega|}\int_\Omega \rho\right), \qquad \int_\Omega V_{\mathrm{H},\rho} = 0,$$

and the function $V_{\mathrm{xc},\rho}$, called the exchange-correlation potential, depends on the chosen approximation of the exchange-correlation energy functional. In the simple X$\alpha$ model [186], $V_{\mathrm{xc},\rho}(\boldsymbol{x}) = -C_\mathrm{D}\rho(\boldsymbol{x})^{1/3}$, where $C_\mathrm{D} > 0$ is the Dirac constant.

It is not mandatory to use pseudopotentials in KS-DFT calculations. Some software allow for all-electrons calculations in which the total pseudopotential operator $V_\mathrm{nl} + V_\mathrm{loc}$ is replaced with a local potential with Coulomb singularities at the positions of the nuclei. However, most calculations are done with pseudopotentials, or use the formally similar Projector Augmented Wave (PAW) method [23], for three reasons: (i) core electrons are barely affected by the chemical environment and can usually be considered to occupy "frozen states", (ii) in heavy atoms, core electrons must be dealt with relativistic quantum models which makes the simulation more expensive from a computational viewpoint, (iii) due to the Coulomb singularities, all-electron Kohn–Sham orbitals have cusps at the positions of the nuclei and are therefore only Lipschitz continuous, while the Kohn–Sham orbitals computed with pseudopotentials are much more regular and can be well approximated with Fourier spectral methods (usually called plane-wave discretization methods in the field).

Several methods for constructing pseudopotentials have been proposed in the literature, leading to local and nonlocal functions of different regularities. As expected, the rate of convergence of the plane-wave discretization method is directly linked to the regularities of these functions. The *a priori* error analysis of this problem was performed in [31] for pseudopotentials with Sobolev regularity. It was proved in particular, for the simple X$\alpha$ exchange-correlation functional, but also for the much more popular local density approximation (LDA) exchange-correlation functional, that if the local and nonlocal part of the pseudopotential are in the periodic Sobolev space of order $s > 3/2$, then the Kohn–Sham orbitals $\varphi_i$ and the density $\rho$ are in the periodic Sobolev space of order $s+2$, and (optimal) polynomial convergence rates were obtained in any Sobolev spaces of order $r$ with $-s < r < s+2$. In addition, as for linear second-order elliptic eigenproblems, the error on the eigenvalues converges to zero as the square of the error on the eigenfunctions evaluated in $\mathrm{H}^1$-norm. The analysis in [31] covers for example the case of Troullier–Martins pseudopotentials [193], for which $s = \frac{7}{2} - \varepsilon$. On the other hand, these estimates are not sharp in the case of Goedecker–Teter–Hutter (GTH) pseudopotentials [79, 91], for which the local and nonlocal contributions are periodic sums of Gaussian-polynomial functions, and therefore have entire continuations to the whole complex plane. Such pseudopotentials are implemented in different DFT software, such as BigDFT [170], Quantum Espresso [78] or Abinit [84, 175], as well as DFTK, a recent electronic structure package in the `Julia` language [101].

The purpose of this chapter is to investigate this case. While it has been known for a long time (see *e.g.* [18, 76, 163] and references therein for historical insight or [21, 92] for more recent developments) that the solutions to elliptic equations on $\mathbb{R}^d$ with real-analytic data have an analytic continuation in a complex neighbourhood of $\mathbb{R}^d$, the size of this neighbourhood is *a priori* unknown. In the periodic case we are considering, the latter directly impacts the decay rate of the Fourier coefficients of the solution, hence the convergence rate of the plane-wave discretization method. For pedagogical reasons, we will work most of the time with one dimensional linear or nonlinear Schrödinger equations, because (i) it is easier to visualize analytic or entire continuations of functions originally defined on the real space $\mathbb{R}^d$ when $d = 1$, and (ii) exponential convergence rates of plane-wave discretization methods are easier to spot in 1D. However, most of our arguments extend to the multidimensional case. In Section 5.2, we introduce a hierarchy of spaces $(\mathcal{H}_A)_{A>0}$ of complex-valued $2\pi$-periodic functions on the real line having analytic continuations to the strip $\mathbb{R} + \mathrm{i}(-A, A)$. We then pick a real-valued function $V \in \mathcal{H}_B$ for some $B > 0$ and consider the one-dimensional Schrödinger operator $H = -\Delta + V$. A low *vs* high-frequency decomposition of the periodic $\mathrm{L}^2$ space allows to prove that for all $0 < A < B$, the solution $u$ to the linear equation $Hu = f$ lays in $\mathcal{H}_A$ whenever $f \in \mathcal{H}_A$ (see Section 5.3.1), and that the eigenfunctions of $H$ are in $\mathcal{H}_A$ (see Section 5.3.2). We rely on this result to prove in Section 5.3.3 that the plane-wave discretization method converges exponentially in this case. We turn in Section 5.4 to the nonlinear setting, where we expose a counter-example for which we show that such results are not true any more. Finally, we consider in Section 5.5 the multidimensional case, which is an immediate extension, and its application to Kohn–Sham models.

## 5.2 Spaces of analytic functions

Let us first introduce some notation. We denote by $\mathrm{L}^2_\#(\mathbb{R},\mathbb{C})$ the space of square integrable complex-valued $2\pi$-periodic functions on $\mathbb{R}$, endowed with its natural inner product

$$(u,v)_{\mathrm{L}^2_\#} := \int_0^{2\pi} \overline{u(x)}\, v(x)\, dx,$$

and by $\mathcal{S}'_\#(\mathbb{R},\mathbb{C})$ the space of tempered complex-valued $2\pi$-periodic distributions on $\mathbb{R}$. For each $u \in \mathcal{S}'_\#(\mathbb{R},\mathbb{C})$, we denote by $(\widehat{u}_k)_{k\in\mathbb{Z}}$ the Fourier coefficients of $u$ with the following normalization convention:

$$\forall\, u \in \mathrm{L}^2_\#(\mathbb{R},\mathbb{C}), \quad \forall\, k \in \mathbb{Z}, \quad \widehat{u}_k := (e_k, u)_{\mathrm{L}^2_\#} = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} u(x) \mathrm{e}^{-\mathrm{i}kx} \mathrm{d}x,$$

where $e_k(x) := \frac{1}{\sqrt{2\pi}} \mathrm{e}^{\mathrm{i}kx}$ is the $\mathrm{L}^2_\#$-normalized Fourier mode with wave-vector $k \in \mathbb{Z}$. Recall that the $2\pi$-periodic Sobolev spaces are the Hilbert spaces $\mathrm{H}^s_\#(\mathbb{R},\mathbb{C})$, $s \in \mathbb{R}$, defined by

$$\mathrm{H}^s_\#(\mathbb{R},\mathbb{C}) := \left\{ u \in \mathrm{L}^2_\#(\mathbb{R},\mathbb{C}) \;\middle|\; \sum_{k\in\mathbb{Z}} (1+|k|^2)^s |\widehat{u}_k|^2 < \infty \right\}, \qquad (u,v)_{\mathrm{H}^s_\#} := \sum_{k\in\mathbb{Z}} (1+|k|^2)^s\, \overline{\widehat{u}_k}\, \widehat{v}_k.$$

We will also use the self-explanatory notation $C^k_\#(\mathbb{R},\mathbb{R})$, $C^k_\#(\mathbb{R},\mathbb{C})$, $\mathrm{L}^p_\#(\mathbb{R},\mathbb{R})$, $\mathrm{L}^p_\#(\mathbb{R},\mathbb{C})$ for $k \in \mathbb{N}\cup\{\infty\}$ and $1 \leqslant p \leqslant \infty$, all these spaces being endowed with their natural norms or topologies. We now introduce, for any $A > 0$, the space

$$\mathcal{H}_A \coloneqq \left\{ u \in \mathrm{L}^2_\#(\mathbb{R},\mathbb{C}) \;\middle|\; \sum_{k\in\mathbb{Z}} w_A(k)|\widehat{u}_k|^2 < \infty \right\} \quad \text{where} \quad w_A(k) \coloneqq \cosh(2Ak),$$

endowed with the inner product

$$(u,v)_A \coloneqq \sum_{k\in\mathbb{Z}} w_A(k)\,\overline{\widehat{u}_k}\,\widehat{v}_k.$$

Note that $\mathcal{H}_A$ can be canonically identified with the space of analytic functions

$$\widetilde{\mathcal{H}}_A \coloneqq \left\{ u : \Omega_A \to \mathbb{C} \text{ analytic} \;\middle|\; \begin{array}{c} [-A,A] \ni y \mapsto u(\cdot + \mathrm{i}y) \in \mathrm{L}^2_\#(\mathbb{R},\mathbb{C}) \text{ continuous,} \\ \displaystyle\int_0^{2\pi} \Big( |u(x+\mathrm{i}A)|^2 + |u(x-\mathrm{i}A)|^2 \Big)\mathrm{d}x < \infty \end{array} \right\},$$

where $\Omega_A \coloneqq \mathbb{R} + \mathrm{i}(-A,A) \subset \mathbb{C}$ is the horizontal strip of width $2A$ of the complex plane centered on the real axis, endowed with the inner product

$$(u,v)_{\widetilde{\mathcal{H}}_A} = \frac{1}{2}\Big( (u(\cdot+\mathrm{i}A), v(\cdot+\mathrm{i}A))_{\mathrm{L}^2_\#} + (u(\cdot-\mathrm{i}A), v(\cdot-\mathrm{i}A))_{\mathrm{L}^2_\#} \Big).$$

The canonical unitary mapping $\mathcal{H}_A$ onto $\widetilde{\mathcal{H}}_A$ is the analytic continuation: any function $u \in \mathcal{H}_A$ has a unique analytic continuation $u : \Omega_A \to \mathbb{C}$ given by

$$\forall\, z = x + \mathrm{i}y \in \Omega_A, \quad u(z) = \sum_{k\in\mathbb{Z}} \widehat{u}_k \,\frac{\mathrm{e}^{\mathrm{i}kz}}{\sqrt{2\pi}} = \sum_{k\in\mathbb{Z}} \widehat{u}_k\,\mathrm{e}^{-ky}\, e_k(x).$$

It can be easily seen that the Fourier coefficients of $u(\cdot \pm \mathrm{i}A)$ are the Fourier coefficients of $u$ rescaled by a factor $\mathrm{e}^{\mp kA}$ and that the function $(-A,A) \ni y \mapsto u(\cdot+\mathrm{i}y) = \sum_{k\in\mathbb{Z}} \widehat{u}_k\,\mathrm{e}^{-ky}\,e_k(\cdot) \in \mathrm{L}^2_\#(\mathbb{R},\mathbb{C})$ has a unique continuation to $[-A,A]$ . Therefore,

$$\|u\|^2_{\widetilde{\mathcal{H}}_A} = \frac{1}{2}\Big( \|u(\cdot+\mathrm{i}A)\|^2_{\mathrm{L}^2_\#} + \|u(\cdot-\mathrm{i}A)\|^2_{\mathrm{L}^2_\#} \Big) = \frac{1}{2}\left( \sum_{k\in\mathbb{Z}} \big|\widehat{u}_k\,\mathrm{e}^{-kA}\big|^2 + \sum_{k\in\mathbb{Z}} \big|\widehat{u}_k\,\mathrm{e}^{+kA}\big|^2 \right) = \sum_{k\in\mathbb{Z}} w_A(k)|\widehat{u}_k|^2 = \|u\|^2_A.$$

**Proposition 5.1.** *Let $B > 0$. Then, for all $0 < A < B$, the multiplication by a function $V \in \mathcal{H}_B$ defines a bounded operator on $\mathcal{H}_A$.*

*Proof.* Let $V \in \mathcal{H}_B$. It holds, for all $0 < A < B$,

$$\begin{aligned} \|V\|^2_{\mathcal{L}(\mathcal{H}_A)} &= \sup_{u\in\mathcal{H}_A\setminus\{0\}} \frac{\|Vu\|^2_A}{\|u\|^2_A} = \sup_{u\in\mathcal{H}_A\setminus\{0\}} \frac{\|V(\cdot+\mathrm{i}A)u(\cdot+\mathrm{i}A)\|^2_{\mathrm{L}^2_\#} + \|V(\cdot-\mathrm{i}A)u(\cdot-\mathrm{i}A)\|^2_{\mathrm{L}^2_\#}}{\|u(\cdot+\mathrm{i}A)\|^2_{\mathrm{L}^2_\#} + \|u(\cdot-\mathrm{i}A)\|^2_{\mathrm{L}^2_\#}} \\ &\leqslant 2\max\Big\{ \|V(\cdot+\mathrm{i}A)\|^2_{\mathrm{L}^\infty_\#}, \|V(\cdot-\mathrm{i}A)\|^2_{\mathrm{L}^\infty_\#} \Big\}. \end{aligned}$$

As the right hand-side is finite for all $0 < A < B$, the proposition follows. □

## 5.3 The linear case

### 5.3.1 The linear elliptic problem

We consider in a first stage the one-dimensional linear elliptic problem

$$\text{seek } u \in \mathrm{H}^2_\#(\mathbb{R},\mathbb{C}) \quad \text{such that} \quad -\Delta u + Vu = f \text{ on } \mathbb{R}, \tag{5.3.1}$$

where $V \in \mathrm{L}^2_\#(\mathbb{R},\mathbb{R})$ and $f \in \mathrm{L}^2_\#(\mathbb{R},\mathbb{C})$ are given $2\pi$-periodic functions. For simplicity, we assume in this section that $V \geqslant 1$, a sufficient condition for the operator $-\Delta + V$ to be invertible. It is well-known that (5.3.1) has a unique solution $u$ satisfying the *a priori* bounds

$$\|u\|_{\mathrm{L}^2_\#} \leqslant \frac{\|f\|_{\mathrm{L}^2_\#}}{\alpha} \quad \text{and} \quad \|u\|_{\mathrm{H}^1_\#} \leqslant \|f\|_{\mathrm{H}^{-1}_\#}, \tag{5.3.2}$$

where $\alpha := \lambda_1(-\Delta + V) \geqslant 1$ is the smallest eigenvalue of the self-adjoint operator $H = -\Delta + V$ on $\mathrm{L}^2_\#(\mathbb{R},\mathbb{C})$. By elementary bootstrap arguments, $u \in \mathrm{H}^{s+2}_\#(\mathbb{R},\mathbb{C})$ whenever $V$ and $f$ are in $\mathrm{H}^s_\#$, for any $s \geqslant 0$. The following result deals with the case of real-analytic potentials $V$ and right-hand sides $f$.

**Theorem 5.1.** *Let $B > 0$ and $V \in \mathcal{H}_B$ be real-valued and such that $V \geqslant 1$ on $\mathbb{R}$. Then, for all $0 < A < B$ and $f \in \mathcal{H}_A$, the unique solution $u$ of (5.3.1) is in $\mathcal{H}_A$. Moreover, we have the following estimate*

$$\exists\, C > 0 \text{ independent of } f \text{ such that } \|u\|_A \leqslant C\|f\|_A. \tag{5.3.3}$$

*As a consequence, if $V$ and $f$ are entire, then so is $u$.*

*Proof.* For $N > 0$, we consider the decomposition $\mathrm{L}^2_\#(\mathbb{R},\mathbb{C}) = X_N \oplus X_N^\perp$ where

$$X_N := \mathrm{Span}(e_k,\ |k| \leqslant N) = \big\{u \in \mathrm{L}^2_\#(\mathbb{R},\mathbb{C}) \mid \widehat{u}_k = 0,\ \forall\ |k| > N\big\}. \tag{5.3.4}$$

Let $\Pi_N$ be the orthogonal projector on $X_N$ and $\Pi_N^\perp := 1 - \Pi_N$ the orthogonal projector on $X_N^\perp$. Note that the restriction of $\Pi_N$ to the Sobolev space $\mathrm{H}^s_\#(\mathbb{R},\mathbb{C})$, $s > 0$, is also the orthogonal projector on $X_N$ for the $\mathrm{H}^s_\#$ inner product, and that the same property holds for the Hilbert spaces $\mathcal{H}_A$.

For a fixed $N$, we decompose $u$ as $u = u_1 + u_2$ with $u_1 \in X_N$ and $u_2 \in X_N^\perp$. As $u_1$ has compact Fourier support, it obviously belongs to $\mathcal{H}_A$ and we have the estimate

$$\|u_1\|_A \leqslant \|u\|_{\mathrm{L}^2_\#}\sqrt{w_A(N)} \leqslant \frac{\|f\|_{\mathrm{L}^2_\#}}{\alpha}\sqrt{w_A(N)}. \tag{5.3.5}$$

Let us show that, for $N$ large enough, $u_2$ also belongs to $\mathcal{H}_A$. Projecting $-\Delta u + Vu = f$ onto $X_N^\perp$, we get

$$T_{22}u_2 + V_{22}u_2 = f_2 - V_{21}u_1, \tag{5.3.6}$$

where $T_{22}$ is the restriction to the invariant subspace $\mathcal{H}_A \cap X_N^\perp$ of the self-adjoint operator $-\Delta$ on $\mathrm{L}^2_\#(\mathbb{R},\mathbb{C})$, $V_{22} := \Pi_N^\perp V \Pi_N^\perp \in \mathcal{L}(\mathcal{H}_A \cap X_N^\perp)$, $V_{21} := \Pi_N^\perp V \Pi_N \in \mathcal{L}(\mathcal{H}_A \cap X_N, \mathcal{H}_A \cap X_N^\perp)$ and $f_2 := \Pi_N^\perp f \in \mathcal{H}_A \cap X_N^\perp$. The operator $T_{22}$ is bounded from below by $N^2$ and is therefore invertible with inverse $T_{22}^{-1}$ bounded by $N^{-2}$ in $\mathcal{L}(\mathcal{H}_A \cap X_N^\perp)$. As $f_2, V_{21}u_1 \in \mathcal{H}_A \cap X_N^\perp$, we can therefore rewrite (5.3.6) as

$$(1 + T_{22}^{-1}V_{22})u_2 = T_{22}^{-1}(f_2 - V_{21}u_1). \tag{5.3.7}$$

Since $\|T_{22}^{-1}\|_{\mathcal{L}(\mathcal{H}_A \cap X_N^\perp)} \leqslant N^{-2}$ and $\|V_{22}\|_{\mathcal{L}(\mathcal{H}_A \cap X_N^\perp)} \leqslant \|V\|_{\mathcal{L}(\mathcal{H}_A)}$, the operator $(1 + T_{22}^{-1}V_{22}) \in \mathcal{L}(\mathcal{H}_A \cap X_N^\perp)$ is invertible for $N$ large enough and it holds

$$u_2 = (1 + T_{22}^{-1}V_{22})^{-1}T_{22}^{-1}(f_2 - V_{21}u_1) = \left(\sum_{n=0}^{+\infty}(-1)^n(T_{22}^{-1}V_{22})^n\right)T_{22}^{-1}(f_2 - V_{21}u_1). \tag{5.3.8}$$

Putting things together with Neumann series, we get

$$\begin{aligned}\|u_2\|_A &\leqslant \frac{\|f_2\|_A + \|V_{21}u_1\|_A}{N^2 - \|V\|_{\mathcal{L}(\mathcal{H}_A)}} \leqslant \frac{\|f\|_A + \|V\|_{\mathcal{L}(\mathcal{H}_A)}\|u_1\|_A}{N^2 - \|V\|_{\mathcal{L}(\mathcal{H}_A)}} \\ &\leqslant \frac{\|f\|_A}{N^2 - \|V\|_{\mathcal{L}(\mathcal{H}_A)}} + \|V\|_{\mathcal{L}(\mathcal{H}_A)}\frac{\|f\|_{\mathrm{L}^2_\#}}{\alpha}\frac{\sqrt{w_A(N)}}{N^2 - \|V\|_{\mathcal{L}(\mathcal{H}_A)}},\end{aligned} \tag{5.3.9}$$

Finally, combining (5.3.5) and (5.3.9) with $\|u\|_A \leqslant \|u_1\|_A + \|u_2\|_A$ yields the bound on $\|u\|_A$. □

**Remark 5.1.** If we require in addition that $V(\cdot \pm \mathrm{i}B)$ is not only in $\mathrm{L}^2_\#(\mathbb{R},\mathbb{C})$ but also in $\mathrm{L}^\infty_\#(\mathbb{R},\mathbb{C})$, then the exact same argument as in the proof of Proposition 5.1 yields

$$\|V\|_{\mathcal{L}(\mathcal{H}_B)} \leqslant 2\max\Big\{\|V(\cdot + \mathrm{i}B)\|_{\mathrm{L}^\infty_\#}, \|V(\cdot - \mathrm{i}B)\|_{\mathrm{L}^\infty_\#}\Big\},$$

so that $f \in \mathcal{H}_B$ then implies $u \in \mathcal{H}_B$.

### 5.3.2 The linear eigenvalue problem

We now focus on the linear eigenvalue problem,

$$\begin{cases} -\Delta u + Vu = \lambda u, \\ \|u\|_{\mathrm{L}^2_\#(\mathbb{R},\mathbb{C})} = 1, \end{cases} \tag{5.3.10}$$

where $V \in \mathcal{H}_B$ for some $B > 0$. Using the same technique as for the proof of Theorem 5.1, we get the following result.

**Theorem 5.2.** *Let $B > 0$, $V \in \mathcal{H}_B$ be real-valued, and $(u, \lambda) \in \mathrm{H}^2_\#(\mathbb{R},\mathbb{C}) \times \mathbb{R}$ a normalized eigenmode of $H = -\Delta + V$, with isolated eigenvalue (*i.e. *a solution to (5.3.10)). Then, $u$ is in $\mathcal{H}_A$ for all $0 < A < B$. Moreover, we have the following estimate*

$$\|u\|_A \leqslant \Big(1 + \|V\|_{\mathcal{L}(\mathcal{H}_A)}\Big) \sqrt{w_A\Big(\sqrt{\|V\|_{\mathcal{L}(\mathcal{H}_A)} + \lambda + 1}\Big)}.$$

*As a consequence, if $V$ is entire, then so is $u$.*

*Proof.* Although the proof of Theorem 5.2 follows basically the same lines as the one of Theorem 5.1, we provide it for the sake of completeness.

Let $(u, \lambda) \in \mathrm{H}^2_\#(\mathbb{R},\mathbb{C}) \times \mathbb{R}$ be a solution to (5.3.10). Using the same notation as in the proof of Theorem 5.1, we decompose $u$ as $u = u_1 + u_2$ with $u_1 \in X_N$ and $u_2 \in X_N^\perp$, and observe that for $N$ large enough,

$$u_2 = -(1 + T_{22}^{-1}(V_{22} - \lambda))^{-1} T_{22}^{-1} V_{21} u_1,$$

with

$$\big\|T_{22}^{-1}(V_{22} - \lambda)\big\|_{\mathcal{L}(\mathcal{H}_A \cap X_N^\perp)} \leqslant \frac{\|V\|_{\mathcal{L}(\mathcal{H}_A)} + |\lambda|}{N^2}.$$

Therefore, choosing $N_\lambda = \sqrt{\|V\|_{\mathcal{L}(\mathcal{H}_A)} + |\lambda| + 1}$, we have

$$\big\|(1 + T_{22}^{-1}(V_{22} - \lambda))^{-1} T_{22}^{-1}\big\|_{\mathcal{L}(\mathcal{H}_A \cap X_{N_\lambda}^\perp)} < 1,$$

which yields

$$\|u_2\|_A \leqslant \|V\|_{\mathcal{L}(\mathcal{H}_A)} \sqrt{w_A(N_\lambda)}.$$

Combining this with $\|u_1\|_A \leqslant \sqrt{w_A(N_\lambda)}$ yields the desired estimate. □

### 5.3.3 Plane-wave approximation of the linear Schrödinger equation

Using $X_N = \mathrm{Span}(e_k,\ |k| \leqslant N) \subset \mathrm{H}^1_\#(\mathbb{R},\mathbb{C})$ as a variational approximation space for (5.3.10), we obtain the finite-dimensional problem

$$\begin{cases} \text{seek } (u_N, \lambda_N) \in X_N \times \mathbb{R} \text{ such that } \|u_N\|_{\mathrm{L}^2_\#(\mathbb{R},\mathbb{C})} = 1 \text{ and} \\ \forall\ v_N \in X_N, \quad \displaystyle\int_0^{2\pi} \overline{\nabla u_N} \cdot \nabla v_N + \int_0^{2\pi} V \overline{u_N} v_N = \lambda_N \int_0^{2\pi} \overline{u_N} v_N, \end{cases} \tag{5.3.11}$$

which is equivalent to seeking the eigenpairs of the Hermitian matrix $H_N \in \mathbb{C}^{\mathcal{N}\times\mathcal{N}}_{\mathrm{herm}}$, where $\mathcal{N} = 2\lfloor N \rfloor + 1$, with entries

$$[H_N]_{kk'} \coloneqq |k|^2 \delta_{kk'} + \widehat{V}_{k-k'}, \quad k, k' \in \mathbb{Z},\ |k| \leqslant N,\ |k'| \leqslant N.$$

The following theorem states that if $V \in \mathcal{H}_B$ for some $B > 0$, the plane-wave discretization method has an exponential convergence rate. Note that a similar result holds for the plane-wave approximation of the linear problem $-\Delta u + Vu = f$, whenever $f \in \mathcal{H}_A$.

**Theorem 5.3.** *Let $B > 0$, $V \in \mathcal{H}_B$ be real-valued, $j \in \mathbb{N}^*$ and $0 < A < B$. Let $\lambda_j$ the lowest $j^{\text{th}}$ eigenvalue of the self-adjoint operator $H = -\Delta + V$ on $\mathrm{L}^2_\#(\mathbb{R},\mathbb{C})$ counting multiplicities, and $\mathcal{E}_j = \mathrm{Ker}(H - \lambda_j)$ the corresponding eigenspace. For $N$ large enough, we denote by $\lambda_{j,N}$ the lowest $j^{\text{th}}$ eigenvalue of (5.3.11), and by $u_{j,N}$ an associated normalized eigenvector. Then, there exists a constant $c_{j,A} \in \mathbb{R}_+$ such that*

$$\forall\ N > 0 \text{ s.t. } 2\lfloor N \rfloor + 1 \geqslant j, \quad d_{\mathrm{H}^1_\#}(u_{j,N}, \mathcal{E}_j) \leqslant c_{j,A} \exp(-AN) \quad \text{and} \quad 0 \leqslant \lambda_{j,N} - \lambda_j \leqslant c_{j,A} \exp(-2AN).$$

*Proof.* First, note that $-\Delta + V$ has compact resolvent so that its eigenvalues $\lambda_j$ are isolated. Let $0 < A < B$ and $A' = \frac{A+B}{2}$. We have

$$\forall\, v \in \mathcal{H}_{A'}, \quad \|v - \Pi_N v\|_{\mathrm{H}^1_\#} \leqslant c_{A,B}\|v\|_{A'}\mathrm{e}^{-AN} \quad \text{with} \quad c_{A,B} \coloneqq \left(\max_{k\in\mathbb{Z}} \frac{(1+|k|^2)\mathrm{e}^{2A|k|}}{w_{A'}(k)}\right)^{1/2} < \infty.$$

The operator $H$ is self-adjoint on $\mathrm{L}^2_\#(\mathbb{R},\mathbb{C})$ with form domain $\mathrm{H}^1_\#(\mathbb{R},\mathbb{C})$ and it follows from Theorem 5.2 that all the eigenfunctions of the operator $H$ are in $\mathcal{H}_{A'}$. The result follows from classical arguments on the variational approximations of the eigenmodes of bounded below self-adjoint operators with compact resolvent (see *e.g.* [8, Theorems 8.1 and 8.2]). □

## 5.4 The nonlinear case: a counter-example

In the perspective of studying nonlinear elliptic problems with analytic data, we now consider the nonlinear periodic elliptic equation with cubic nonlinearity

$$-\varepsilon\Delta u_\varepsilon + u_\varepsilon + u_\varepsilon^3 = f \quad \text{in } \mathrm{H}^1_\#(\mathbb{R},\mathbb{R}), \tag{5.4.1}$$

where $\varepsilon > 0$, and $f : \mathbb{R} \to \mathbb{R}$ is a real-analytic $2\pi$-periodic function admitting an entire continuation, still denoted by $f$, to the complex plane. We will show that, in this particular case, the same kind of results are not true any more and we provide an estimation of the width of the horizontal analyticity strip of the solution, which is finite even though the source term $f$ is entire.

The singular limit $\varepsilon = 0$ gives rise to the algebraic equation $u_0(x) + u_0(x)^3 = f(x)$, which has a unique real solution for each $x \in \mathbb{R}$. The latter can be computed by Cardano's formula: the discriminant of the cubic equation is

$$R(x) = -\big(4 + 27f^2(x)\big) < 0,$$

so that

$$u_0(x) = \sqrt[3]{\frac{1}{2}\left(f(x) + \sqrt{\frac{-R(x)}{27}}\right)} + \sqrt[3]{\frac{1}{2}\left(f(x) - \sqrt{\frac{-R(x)}{27}}\right)}, \tag{5.4.2}$$

the other two roots being complex conjugates with nonzero imaginary parts.

In the rest of this section, we will use the function $f : x \mapsto \mu\sin(x)$, for a fixed $\mu > 0$. The analytic continuation of the function $u_0$ originally defined on $\mathbb{R}$ satisfies

$$u_0(z) + u_0^3(z) = f(z), \tag{5.4.3}$$

with $\sqrt{\cdot}$ and $\sqrt[3]{\cdot}$ (used in (5.4.2)) the continuations of the square root and cubic root functions with branch cut respectively $\mathbb{R}_-$ and $\mathrm{i}\mathbb{R}$. The maximal horizontal strip of the complex plane on which the function $u_0(z)$ is analytic is $\mathbb{R} + \mathrm{i}(-B_0, B_0)$ where $B_0 = \operatorname{arcsinh}\left(\mu^{-1}\sqrt{4/27}\right) > 0$ is such that the discriminant cancels for $z = \pm\mathrm{i}B_0$. More precisely, the function $u_0(z)$ has branching points on the imaginary axis at $z_\pm = \pm\mathrm{i}B_0$. It holds $u_0(z_\pm) = \pm\mathrm{i}/\sqrt{3}$, $f(z_\pm) = \pm\mathrm{i}\sqrt{4/27}$, and $R(z_\pm) = 0$. The complex number $u_0(z_\pm)$ is the threefold degenerate root of the cubic equation

$$Z^3 + Z - f(z_\pm) = (Z - u_0(z_\pm))^3 = 0,$$

and we have

$$\left|\frac{\mathrm{d}u_0}{\mathrm{d}t}(tz_\pm)\right| = \left|\frac{\frac{\mathrm{d}f}{\mathrm{d}t}(tz_\pm)}{1 + 3(\frac{\mathrm{d}u_0}{\mathrm{d}t}(tz_\pm))^2}\right| \to \infty \quad \text{when } \mathbb{R} \ni t \to 1_-.$$

In particular, although the source term $f$ has an entire continuation, the solution $u_0$ of (5.4.3) does not.

When $\varepsilon > 0$, we can approximate numerically the solution to (5.4.1) with the plane-wave approximation introduced before. The plots in Figure 5.2 suggest that increasing $\varepsilon$ increases the width $B_\varepsilon$ of the horizontal analyticity strip of $z \mapsto u_\varepsilon(z)$, but does not make it entire. Moreover, we can quantify the convergence of $u_\varepsilon$ towards $u_0$ on the real axis with the following result.

**Theorem 5.4.** *We have the following convergence estimates:* $\exists\ C_1, C_2, C_3 > 0$ *such that, for* $\varepsilon > 0$,

$$\|u_\varepsilon - u_0\|_{\mathrm{L}^2_\#} \leqslant C_1\varepsilon, \quad \|u_\varepsilon - u_0\|_{\mathrm{H}^1_\#} \leqslant C_2\varepsilon, \quad \|u_\varepsilon - u_0\|_{\mathrm{H}^2_\#} \leqslant C_3\varepsilon.$$

*Proof.* cf. appendix. □

In this particular case, we can obtain an upper bound of the value of $B_\varepsilon$. Let $\varphi_\varepsilon(y) \coloneqq u_\varepsilon(\mathrm{i}y)$. We also have $f(\mathrm{i}y) = \mu\sin(\mathrm{i}y) = \mathrm{i}\mu\sinh(y)$. Since $u_\varepsilon$ is analytic at $z = 0$, $\varphi_\varepsilon$ satisfies the second-order ODE:

$$\begin{cases} \varepsilon\varphi_\varepsilon''(y) + \varphi_\varepsilon(y) + \varphi_\varepsilon^3(y) = \mathrm{i}\mu\sinh, \\ \varphi_\varepsilon(0) = u_\varepsilon(0) = 0, \quad \varphi_\varepsilon'(0) = \mathrm{i}u_\varepsilon'(0), \end{cases}$$

where $u_\varepsilon'(0) \in \mathbb{R}$ since $u_\varepsilon$ is real-valued on $\mathbb{R}$. Decomposing $\varphi_\varepsilon$ in its real part $\theta_\varepsilon$ and imaginary part $\psi_\varepsilon$ (*i.e.* $\varphi_\varepsilon = \theta_\varepsilon + \mathrm{i}\psi_\varepsilon$), we see that $\theta_\varepsilon$ and $\psi_\varepsilon$ satisfy the coupled system of ODEs

$$\begin{cases} \varepsilon\theta_\varepsilon'' + \theta_\varepsilon + \theta_\varepsilon^3 - 3\theta_\varepsilon\psi_\varepsilon^2 = 0, \\ \theta_\varepsilon(0) = 0, \quad \theta_\varepsilon'(0) = 0, \end{cases} \qquad \begin{cases} \varepsilon\psi_\varepsilon'' + \psi_\varepsilon - \psi_\varepsilon^3 + 3\theta_\varepsilon^2\psi_\varepsilon = \mu\sinh, \\ \psi_\varepsilon(0) = 0, \quad \psi_\varepsilon'(0) = u_\varepsilon'(0). \end{cases}$$

This implies that $\theta_\varepsilon = 0$, $\psi_\varepsilon$ is odd, and $u_\varepsilon$ remains purely imaginary along the imaginary axis. We now focus on the ODE satisfied by $\psi_\varepsilon$ on $\mathbb{R}_+$, which can be rewritten as a first-order ODE on $\Psi_\varepsilon(y) \coloneqq \begin{bmatrix} \psi_\varepsilon(y) \\ \psi_\varepsilon'(y) \end{bmatrix} \in \mathbb{R}^2$

$$\Psi_\varepsilon'(y) = \begin{bmatrix} \Psi_{\varepsilon,2}(y) \\ \varepsilon^{-1}\big(\mu\sinh(y) - \Psi_{\varepsilon,1}(y) + \Psi_{\varepsilon,1}^3(y)\big) \end{bmatrix}, \qquad \Psi_\varepsilon(0) \coloneqq \begin{bmatrix} 0 \\ u_\varepsilon'(0) \end{bmatrix}. \tag{5.4.4}$$

If we can prove that $\Psi_\varepsilon$ blows up at a finite $0 < Y_\varepsilon < \infty$, then this will imply that the width $B_\varepsilon$ of the horizontal analyticity strip of $u_\varepsilon$ satisfies $B_\varepsilon \leqslant Y_\varepsilon < \infty$.

In order to estimate $Y_\varepsilon$, we need comparison theorems for systems of ODEs. We use the following simplified version of more general results on systems of differential inequalities [199, 200]. In the sequel, the inequality $a \geqslant b$ for two vectors $a, b \in \mathbb{R}^d$ means that $a_i \geqslant b_i$ for all $1 \leqslant i \leqslant d$.

**Theorem 5.5.** *Let* $d \geqslant 1$ *and* $G : \mathbb{R}^d \to \mathbb{R}^d$ *be locally Lipschitz and quasimonotone in the sense that for all* $X, Z \in \mathbb{R}^d$,

$$(Z_i = X_i \text{ and } Z_j \geqslant X_j \text{ for } j \neq i) \quad \Rightarrow \quad (G(X) \leqslant G(Z)).$$

*Let* $0 \leqslant y_0 < y_{\mathrm{M}} \leqslant +\infty$ *and* $\Phi \in C^1([y_0, y_{\mathrm{M}}), \mathbb{R}^d)$ *and* $\Psi \in C^1([y_0, y_{\mathrm{M}}), \mathbb{R}^d)$ *satisfying respectively the ODE*

$$\Phi'(y) = G(\Phi(y)), \qquad \Phi(y_0) \in \mathbb{R}^d,$$

*and the differential inequality*

$$\Psi'(y) \geqslant G(\Psi(y)), \qquad \Psi(y_0) = \Phi(y_0).$$

*Then we have*

$$\forall\ y \in [y_0, y_{\mathrm{M}}), \quad \Psi(y) \geqslant \Phi(y).$$

*Proof.* See *e.g.* [199, p. 112] for a more general result or cf. appendix for the proof of this particular case. □

To apply this result to (5.4.4), we introduce the function $G_\varepsilon : \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$\forall\ X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \in \mathbb{R}^2, \quad G_\varepsilon(X) = \begin{bmatrix} X_2 \\ \varepsilon^{-1}(-X_1 + X_1^3) \end{bmatrix},$$

and the maximal solution $\Phi_\varepsilon$ to

$$\Phi_\varepsilon'(y) = G_\varepsilon(\Phi_\varepsilon(y)), \qquad \Phi_\varepsilon(0) = \begin{bmatrix} 0 \\ u_\varepsilon'(0) \end{bmatrix}. \tag{5.4.5}$$

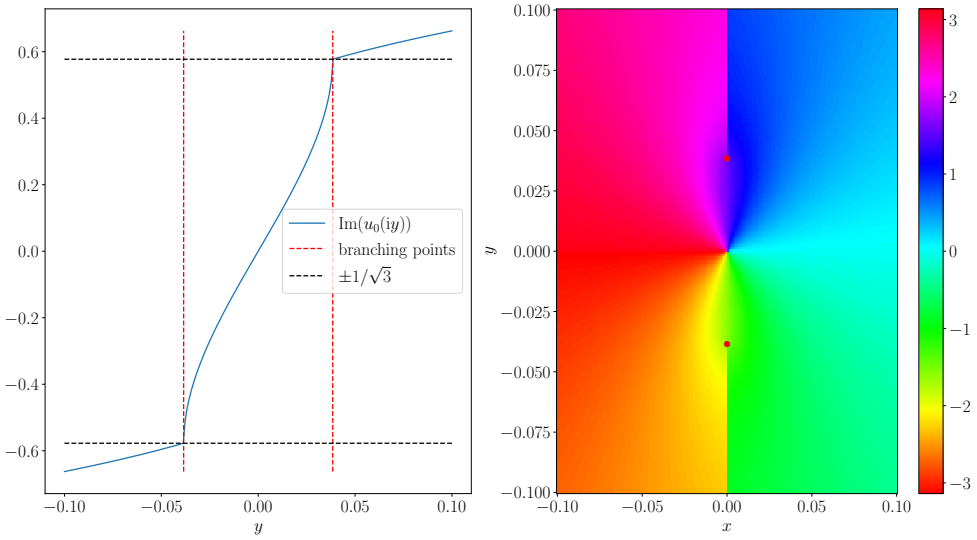



FIGURE 5.1 – Analytic continuation of $u_0$ for $\mu = 10$, for which $B_0 \approx 0.0385$. (Right) Phase of $z \mapsto u_0(z)$. A branching point, in red, appears at the expected position. (Left) Imaginary part of $y \mapsto u_0(\mathrm{i}y)$. A discontinuity also appears at the expected position.

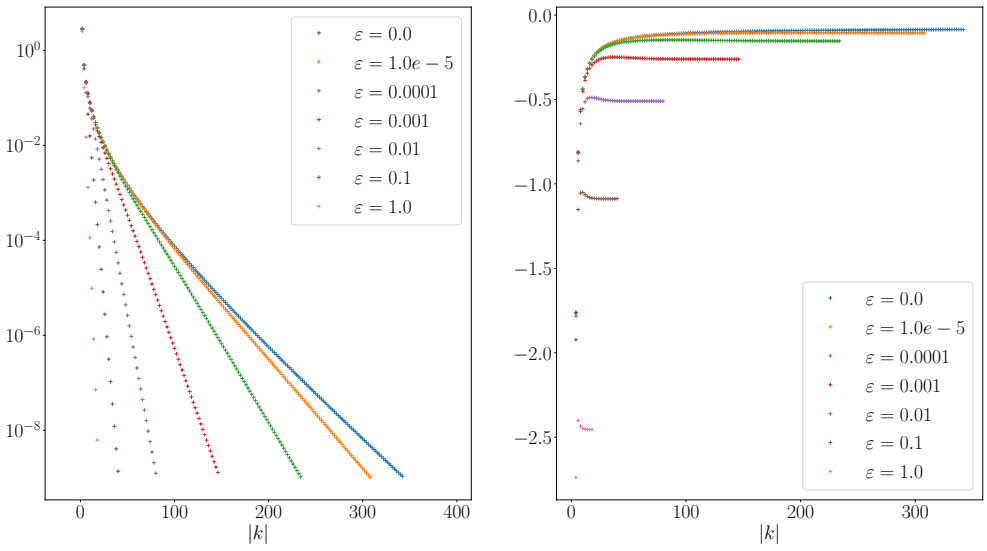



FIGURE 5.2 – (Left) Fourier coefficients of $u_\varepsilon$. (Right) Logarithm of the ratio of two successive nonzero Fourier coefficients.

As $\sinh(y) \geqslant 0$ for all $y \geqslant 0$, we have

$$\Psi_\varepsilon'(y) \geqslant G_\varepsilon(\Psi_\varepsilon(y)), \qquad \Psi_\varepsilon(0) = \Phi_\varepsilon(0).$$

However, Theorem 5.5 cannot be directly applied as the quasimonotonicity assumption is not satisfied everywhere in $\mathbb{R}^2$: $G_{\varepsilon,i}(X_1, X_2)$ is nondecreasing in $X_2$ for $i = 1$ but it is not in $X_1$ for $i = 2$. It is only the case if $X_1$ is in the domain where $x \mapsto x^3 - x$ is nondecreasing. Thus, in order to be able to apply the comparison theorem, we have to show that $\psi_\varepsilon = (\Psi_\varepsilon)_1$ is in this domain at some $y_0$ and stays in it for all $y \geqslant y_0$.

To this end, we introduce the set

$$X_\mu = \big\{(y, v) \in \mathbb{R}^2,\ \mu \sinh(y) - v + v^3 \geqslant 0\big\}.$$

This set is such that, if $(y, \psi_\varepsilon(y))$ lies strictly in $X_\mu$, then $\psi_\varepsilon$ is locally strictly convex. For $y < B_0$, $\psi_\varepsilon$ might oscillate on both sides of the boundary of $X_\mu$ [198]. To make this result more precise, we start by quoting a general lemma on second-order ODEs, whose proof is given in the appendix.

**Lemma 5.1.** *Let $T > 0$ and $\omega_\varepsilon$ be a solution to the second-order ODE on $[0, T]$*

$$\begin{cases} \varepsilon\omega_\varepsilon''(t) + \alpha(t)\omega_\varepsilon(t) = \varepsilon\beta(t, \omega_\varepsilon, \varepsilon), \\ \omega_\varepsilon(0) = 0,\ \omega_\varepsilon'(0) = O(\varepsilon), \end{cases}$$

*where $\alpha(t) \geqslant \alpha_T > 0$ on $[0, T]$ and $\exists\ c_T > 0$ such that*

$$\forall\ t \in [0, T],\ z \in \mathbb{R},\ \varepsilon > 0, \quad |\beta(t, z, \varepsilon)| \leqslant c_T\left(1 + \frac{|z|^2}{\varepsilon}\right)$$

*and*

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\beta(t, \omega_\varepsilon(t), \varepsilon)\right| \leqslant c_T\left(1 + \frac{|\omega_\varepsilon(t)\omega_\varepsilon'(t)|}{\varepsilon}\right).$$

*Then, $\exists\ C_T$ and $\varepsilon_T$ such that*

$$\forall\ \varepsilon \leqslant \varepsilon_T, \quad \begin{cases} \|\omega_\varepsilon\|_{\mathrm{L}^\infty([0,T])} \leqslant C_T\varepsilon, \\ \|\omega_\varepsilon'\|_{\mathrm{L}^\infty([0,T])} \leqslant C_T\sqrt{\varepsilon}. \end{cases}$$

*Proof.* cf. appendix. □

Recall that, by Theorem 5.4, $u_\varepsilon \to u_0$ in $\mathrm{H}^2_\#(\mathbb{R})$. As a consequence, $\psi_\varepsilon(0) = 0$ and $\psi_\varepsilon'(0) = \psi_0'(0) + O(\varepsilon)$. Now, we introduce $h(z) = z - z^3$ such that

$$\begin{cases} \varepsilon\psi_\varepsilon''(y) + h(\psi_\varepsilon(y)) = \mu\sinh(y), \\ h(\psi_0(y)) = \mu\sinh(y). \end{cases}$$

Thus, the error $\omega_\varepsilon(y) \coloneqq \psi_\varepsilon(y) - \psi_0(y)$ satisfies

$$\begin{aligned} \varepsilon\omega_\varepsilon''(y) &= \varepsilon\psi_\varepsilon''(y) - \varepsilon\psi_0''(y) = -(h(\psi_\varepsilon(y)) - h(\psi_0(y))) - \varepsilon\psi_0''(y) \\ &= -h'(\psi_0(y))\omega_\varepsilon(y) + \int_{\psi_0(y)}^{\psi_\varepsilon(y)} \frac{h''(s)}{2}(\psi_\varepsilon(y) - s)\mathrm{d}s - \varepsilon\psi_0''(y). \end{aligned}$$

By taking $T < B_0$, $h'(\psi_0(y)) = 1 - 3|\psi_0(y)|^2 \geqslant 1 - 3|\psi_0(T)|^2 > 0$ as $B_0$ is defined such that $1 = 3|\psi_0(B)|^2$. Thus, Lemma 5.1 can be applied to $\omega_\varepsilon$ and we have that

$$\forall\ T < B_0,\ \exists\ C_T > 0,\ \varepsilon_T > 0 \text{ such that } \forall\ \varepsilon \leqslant \varepsilon_T,\ \begin{cases} \|\psi_\varepsilon - \psi_0\|_{\mathrm{L}^\infty([0,T])} \leqslant C_T\varepsilon, \\ \|\psi_\varepsilon' - \psi_0\|_{\mathrm{L}^\infty([0,T])} \leqslant C_T\sqrt{\varepsilon}. \end{cases}$$

This yields the uniform convergence of $\psi_\varepsilon$ and $\psi_\varepsilon'$ towards $\psi_0$ and $\psi_0'$, unfortunately only on compact subsets $[0, T]$ of $[0, B_0)$, with a constant $C_T \to \infty$ as $T \to B_0$.

The convergence being valid only on compact subsets $[0,T] \subset [0,B_0)$, we cannot deduce properties of $\psi_\varepsilon(B_0)$ from those of $\psi_0(B_0)$. We thus investigated numerically the behaviour of this function for the set of parameters used in Figure 5.3 ($\varepsilon = 0.1$, $\mu = 0.5$), and observed that $0 < \psi_\varepsilon(B_0) < \frac{1}{\sqrt{3}}$ and $\psi'_\varepsilon(B_0) > 0$. This numerical observation can be trusted as the ODE satisfied by $\psi'_\varepsilon$ on the interval $[0,B_0]$ for $\varepsilon = 0.1$ and $\mu = 0.5$ is not stiff. It is therefore easy to solve it numerically with high accuracy with *a posteriori* error estimates guaranteeing that $\psi_\varepsilon(B_0)$ is indeed strictly between 0 and $\frac{1}{\sqrt{3}}$, and $\psi'_\varepsilon(B_0)$ is positive. Therefore, given the shape of the set $X_\mu$ (see Figure 5.3), $\psi_\varepsilon$ is strictly convex on $[B_0, Y_\varepsilon)$. It is thus always above its tangent at $y = B_0$, whose slope is $\psi'_\varepsilon(B_0) > 0$. Therefore, for any $\eta > 0$, there is $y_\eta \geqslant B_0$ such that $\psi_\varepsilon(y_\eta) = 1 + \eta$ and $\psi_\varepsilon(y) \geqslant 1 + \eta > 1$ for any $y \geqslant y_\eta$. We are now ready to compute an upper bound of $Y_\varepsilon$ with the use of Theorem 5.5, that can be applied as the quasimonotonicity assumption is now satisfied in the domain of interest. To this end, we rewrite (5.4.5) as

$$\Phi'_\varepsilon(y) = G_\varepsilon(\Phi_\varepsilon(y)), \qquad \Phi_\varepsilon(y_\eta) = \begin{bmatrix} 1+\eta \\ \psi'_\varepsilon(y_\eta) \end{bmatrix}. \tag{5.4.6}$$

Theorem 5.5 then yields:

$$\forall\, y \geqslant y_\eta,\ \Psi_\varepsilon(y) \geqslant \Phi_\varepsilon(y).$$

This leads us to study of the ODE

$$\begin{cases} \varepsilon \phi''_\varepsilon = -\phi_\varepsilon + \phi_\varepsilon^3 \\ \phi_\varepsilon(y_\eta) = 1+\eta,\ \phi'_\varepsilon(y_\eta) = \psi'_\varepsilon(y_\eta) > 0. \end{cases}$$

We have

$$\frac{\varepsilon}{2}\frac{\mathrm{d}}{\mathrm{d}y}(\phi'_\varepsilon)^2 = \frac{\mathrm{d}}{\mathrm{d}y}\left(-\frac{1}{2}\phi_\varepsilon^2 + \frac{1}{4}\phi_\varepsilon^4\right),$$

from which we deduce

$$\frac{\varepsilon}{2}(\phi'_\varepsilon)^2 = \frac{1}{4}\phi_\varepsilon^4 - \frac{1}{2}\phi_\varepsilon^2 + C(\eta),$$

with

$$\begin{aligned} C(\eta) &= -\frac{1}{4}(1+\eta)^4 + \frac{1}{2}(1+\eta)^2 + \frac{\varepsilon}{2}(\psi'_\varepsilon(y_\eta))^2 \\ &\geqslant -\frac{1}{4}(1+\eta)^4 + \frac{1}{2}(1+\eta)^2 + \frac{\varepsilon}{2}(\psi'_\varepsilon(y_0))^2, \end{aligned}$$

where $C(\eta)$ is computed from the initial conditions at $y = y_\eta$ and $B_0 < y_0 < y_\eta$ is such that $\psi_\varepsilon(y_0) = 1$. Note that $\eta \mapsto -\frac{1}{4}(1+\eta)^4 + \frac{1}{2}(1+\eta)^2 + \frac{\varepsilon}{2}(\psi'_\varepsilon(y_0))^2$ is decreasing on $\mathbb{R}_*^+$ and takes the value $\frac{1}{4} + \frac{\varepsilon}{2}(\psi'_\varepsilon(y_0))^2 > \frac{1}{4}$ at $\eta = 0$. Thus, for $\eta > 0$ small enough, $C(\eta) \geqslant \frac{1}{4}$ and we have

$$\frac{\varepsilon}{2}(\phi'_\varepsilon)^2 \geqslant \frac{1}{4}\phi_\varepsilon^4 - \frac{1}{2}\phi_\varepsilon^2 + \frac{1}{4} = \frac{1}{4}\left(\phi_\varepsilon^2 - 1\right)^2,$$

hence

$$\phi'_\varepsilon \geqslant \frac{1}{\sqrt{2\varepsilon}}\left(\phi_\varepsilon^2 - 1\right) \quad \text{on } [y_\eta, Y_\varepsilon).$$

Finally, we consider the ODE

$$\xi'_{\varepsilon,\eta} = \frac{1}{\sqrt{2\varepsilon}}(\xi_{\varepsilon,\eta}^2 - 1), \qquad \xi_{\varepsilon,\eta}(y_\eta) = 1+\eta,$$

whose solution is

$$\xi_{\varepsilon,\eta}(y) = \frac{1 + \frac{2}{\eta} + \exp\left(\frac{y-y_\eta}{\sqrt{\varepsilon/2}}\right)}{1 + \frac{2}{\eta} - \exp\left(\frac{y-y_\eta}{\sqrt{\varepsilon/2}}\right)},$$

which is defined only up to $Y_{\varepsilon,\eta} := \sqrt{\frac{\varepsilon}{2}}\log\left(1 + \frac{2}{\eta}\right) + y_\eta$ . Applying again Theorem 5.5, we have that $\phi_\varepsilon(y) \geqslant \xi_{\varepsilon,\eta}(y)$ for any $y \geqslant y_\eta$ such that both functions are still finite. Putting everything together, we obtain that $\psi_\varepsilon$ is only defined up to some $Y_\varepsilon$ with $B_0 < Y_\varepsilon \leqslant Y_{\varepsilon,\eta}$ and that

$$\forall\, y \in [y_\eta, Y_\varepsilon), \quad \psi_\varepsilon(y) \geqslant \xi_{\varepsilon,\eta}(y). \tag{5.4.7}$$

These results are illustrated on Figure 5.3, where we plotted the lower bound $\xi_{\varepsilon,\eta}$ for $\varepsilon = 0.1$ and $\eta = 0.5$.

We can deduce from these investigations that $u_\varepsilon$ is only analytic on a horizontal strip of finite width of the complex plane although the source term $f$ is an entire function: our results in the linear case are therefore no longer valid in general in the nonlinear case.

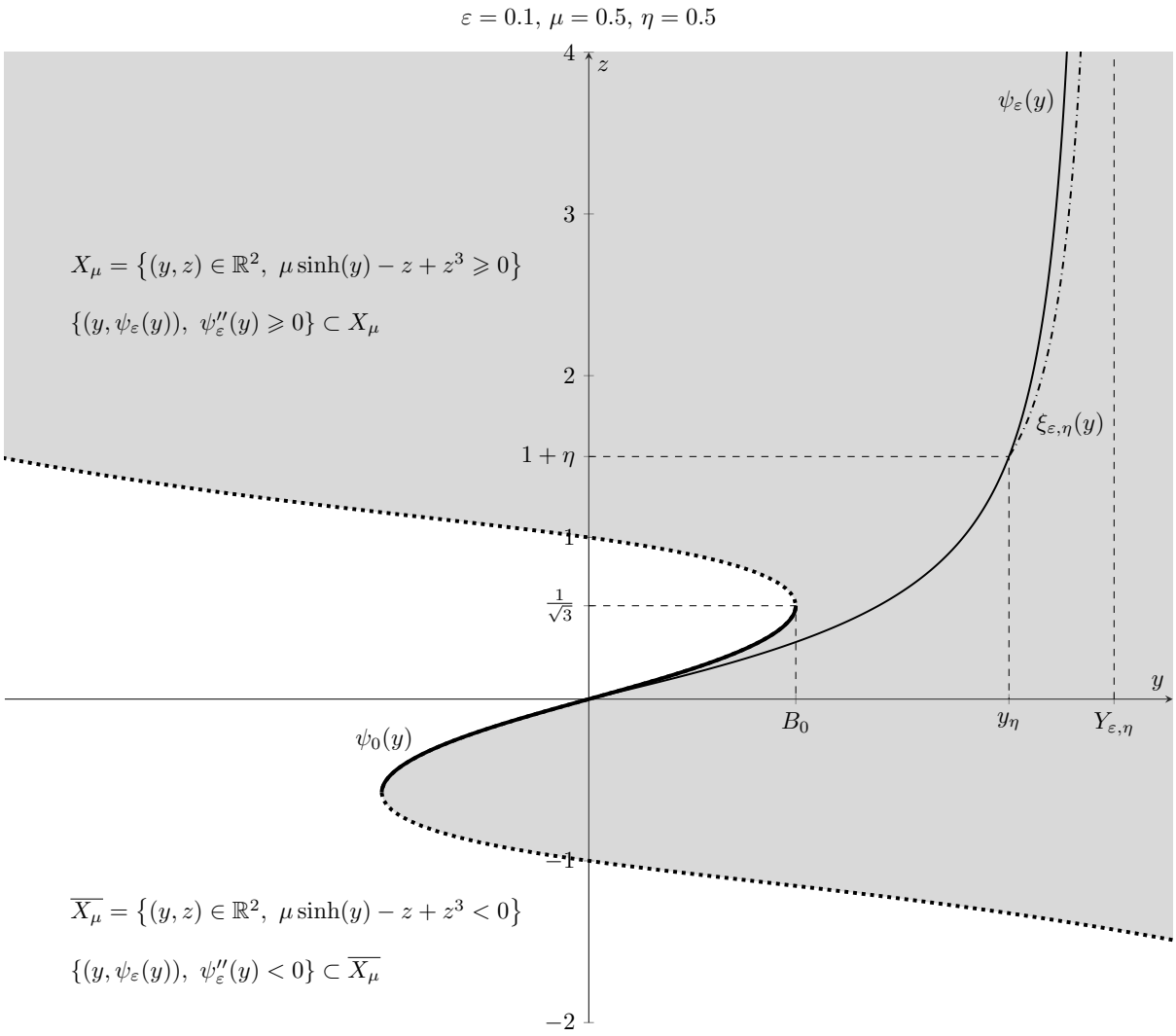



FIGURE 5.3 – Description of $X_\mu$ for $\mu = 0.5$, along with the plot of $\psi_\varepsilon$ and the lower bound $\xi_{\varepsilon,\eta}$ for $\varepsilon = 0.1$, $\eta = 0.5$. While $y < B_0$, $\psi_\varepsilon$ can possibly oscillate around $\psi_0$, but as soon as $y \geqslant B_0$, $\psi_\varepsilon$ is strictly convex and has no other choice than to explode in finite time $Y_\varepsilon \leqslant Y_{\varepsilon,\eta}$, where $Y_{\varepsilon,\eta}$ is the explosion time of the lower bound $\xi_{\varepsilon,\eta}$.

## 5.5 Extension to the multidimensional case with application to Kohn–Sham models.

The goal of this section is to extend the previous results to the multidimensional case and apply them to the linear version of the Kohn–Sham equations (5.1.1). To this end, consider a Bravais lattice $\mathbb{L} = \mathbb{Z}\boldsymbol{a}_1 + \cdots + \mathbb{Z}\boldsymbol{a}_d$ where $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d$ are linearly independent vectors of $\mathbb{R}^d$ ($d = 3$ for KS-DFT). We denote by $\Omega = [0,1)\boldsymbol{a}_1 + \cdots + [0,1)\boldsymbol{a}_d$ a unit cell, by $\mathbb{L}^*$ the reciprocal lattice, by $e_{\boldsymbol{G}}(\boldsymbol{x}) = |\Omega|^{-1/2}\mathrm{e}^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{x}}$ the Fourier mode with wave-vector $\boldsymbol{G} \in \mathbb{L}^*$, and by

$$\mathrm{H}^s_{\#,\mathbb{L}} \coloneqq \left\{ u = \sum_{\boldsymbol{G}\in\mathbb{L}^*} \widehat{u}_{\boldsymbol{G}} e_{\boldsymbol{G}} \in \mathcal{S}(\mathbb{R}^d, \mathbb{C}) \;\middle|\; \sum_{\boldsymbol{G}\in\mathbb{L}^*} (1+|\boldsymbol{G}|^2)^s |\widehat{u}_{\boldsymbol{G}}|^2 < \infty \right\}$$

the $\mathbb{L}$-periodic Sobolev spaces endowed with their usual inner products. All the arguments in Sections 5.3.1-5.3.3 can be extended to the multidimensional case by introducing the Hilbert spaces

$$\mathcal{H}_{A,\mathbb{L}} \coloneqq \left\{ u \in \mathrm{L}^2_{\#,\mathbb{L}} \;\middle|\; \sum_{\boldsymbol{G}\in\mathbb{L}^*} w_{A,\mathbb{L}}(\boldsymbol{G}) |\widehat{u}_{\boldsymbol{G}}|^2 < \infty \right\}, \quad (u,v)_{A,\mathbb{L}} \coloneqq \sum_{\boldsymbol{G}\in\mathbb{L}^*} w_{A,\mathbb{L}}(\boldsymbol{G}) \overline{\widehat{u}_{\boldsymbol{G}}} \widehat{v}_{\boldsymbol{G}},$$

where $w_{A,\mathbb{L}}(\boldsymbol{G}) = \displaystyle\sum_{n=1}^{d} w_A((2\pi)^{-1}\boldsymbol{G}\cdot\boldsymbol{a}_n)$ . Note that the notation $w_{A,\mathbb{L}}$ is slightly misleading as the $w_{A,\mathbb{L}}$'s actually depend on the chosen basis $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d$ of the lattice $\mathbb{L}$. Each $u \in \mathcal{H}_{A,\mathbb{L}}$ can be extended to an analytic function $u(z_1, \ldots, z_d)$ of $d$ complex variables defined on a neighbourhood on $\mathbb{R}^d$, and it holds

$$\sum_{\boldsymbol{G}\in\mathbb{L}^*} w_{A,\mathbb{L}}(\boldsymbol{G}) |\widehat{u}_{\boldsymbol{G}}|^2 = \frac{1}{2} \sum_{n=1}^{d} \int_\Omega \left| u(\boldsymbol{x} + \mathrm{i}(2\pi)^{-1} A\boldsymbol{a}_n) \right|^2 + \left| u(\boldsymbol{x} - \mathrm{i}(2\pi)^{-1} A\boldsymbol{a}_n) \right|^2 \mathrm{d}\boldsymbol{x}.$$

The extension of Proposition 5.1 follows with the operator norm, for any $0 < A < B$,

$$\forall\, V \in \mathcal{H}_{B,\mathbb{L}}, \quad \|V\|_{\mathcal{L}(\mathcal{H}_{A,\mathbb{L}})} = \max_{1\leqslant n\leqslant d} \left\| V(\cdot + \mathrm{i}(2\pi)^{-1} A\boldsymbol{a}_n) \right\|_{\mathrm{L}^\infty_\#}.$$

The approximation space $X_{N,\mathbb{L}}$ is then defined as

$$X_{N,\mathbb{L}} \coloneqq \mathrm{Span}(e_{\boldsymbol{G}},\ \boldsymbol{G} \in \mathbb{L}^*,\ |\boldsymbol{G}| \leqslant N),$$

and the inverse $T^{-1}_{22,\mathbb{L}}$ of the restriction $T_{22,\mathbb{L}}$ of the operator $-\Delta$ on $\mathrm{L}^2_{\#,\mathbb{L}}$ to the invariant subspace $X^\perp_{N,\mathbb{L}} = \mathrm{Span}(e_{\boldsymbol{G}},\ \boldsymbol{G} \in \mathbb{L}^*,\ |\boldsymbol{G}| > N)$ satisfies

$$\| T_{22,\mathbb{L}} \|_{\mathcal{L}(X^\perp_{N,\mathbb{L}})} = \| T_{22,\mathbb{L}} \|_{\mathcal{L}(\mathcal{H}_{A,\mathbb{L}} \cap X^\perp_{N,\mathbb{L}})} \leqslant N^{-2}.$$

The proofs of Theorem 5.1, Theorem 5.2 and Theorem 5.3 can thus be straightforwardly adapted to the multidimensional case.

Lastly, if $V \in \mathcal{H}_{B,\mathbb{L}}$ for some $B > 0$, the Schrödinger operator $H = -\Delta + V$ *considered this time as a Schrödinger operator on* $\mathrm{L}^2(\mathbb{R}^d, \mathbb{C})$ *with an* $\mathbb{L}$*-periodic potential*, can be decomposed by the Bloch transform and its Bloch fibers are the self-adjoint operators on $\mathrm{L}^2_{\#,\mathbb{L}}$ with domain $\mathrm{H}^2_{\#,\mathbb{L}}$ and form domain $\mathrm{H}^1_{\#,\mathbb{L}}$ defined as $H_{\boldsymbol{k}} = (-\mathrm{i}\nabla + \boldsymbol{k})^2 + V$. The following result is concerned with the Bloch eigenmodes of $H$.

**Theorem 5.6.** *Let $B > 0$ and $V \in \mathcal{H}_{B,\mathbb{L}}$. For each $\boldsymbol{k} \in \mathbb{R}^d$, the eigenfunctions of the Bloch fibers $H_{\boldsymbol{k}} = (-\mathrm{i}\nabla + \boldsymbol{k})^2 + V$ of the periodic Schrödinger operator $H = -\Delta + V$ are in $\mathcal{H}_{A,\mathbb{L}}$ for any $0 < A < B$. Let $\lambda_{1,\boldsymbol{k}} \leqslant \lambda_{2,\boldsymbol{k}} \leqslant \cdots$ be the eigenvalues of $H_{\boldsymbol{k}}$ counted with multiplicities and ranked in nondecreasing order, and $\lambda_{1,\boldsymbol{k},N} \leqslant \lambda_{2,\boldsymbol{k},N} \leqslant \cdots \leqslant \lambda_{d_{\mathbb{L},N},\boldsymbol{k},N}$ the eigenvalues of the variational approximation of $H_{\boldsymbol{k}}$ in the $d_{\mathbb{L},N}$-dimensional space*

$$X_{\mathbb{L},\boldsymbol{k},N} \coloneqq \mathrm{Span}(e_{\boldsymbol{G}},\ \boldsymbol{G} \in \mathbb{L}^*,\ |\boldsymbol{G} + \boldsymbol{k}| \leqslant N).$$

*Then, for each $0 < A < B$ and $n \in \mathbb{N}^*$, there exists a constant $C \in \mathbb{R}_+$ such that*

$$0 \leqslant \max_{\boldsymbol{k}\in\Omega^*} (\lambda_{n,\boldsymbol{k},N} - \lambda_{n,\boldsymbol{k}}) \leqslant C\mathrm{e}^{-2AN}, \tag{5.5.1}$$

*where $\Omega^*$ is the first Brillouin zone (*i.e. *the Voronoi cell of the lattice $\mathbb{L}$ of $\mathbb{R}^d$ containing the origin).*

*Proof.* It suffices to replace in the proofs of Theorem 5.2 and Theorem 5.3 $X_N$ with $X_{\mathbb{L},\boldsymbol{k},N}$ and $T_{22}$ with the restriction $T_{22,\mathbb{L},\boldsymbol{k}}$ of the operator $(-\mathrm{i}\nabla+\boldsymbol{k})^2$ to the invariant space $X^\perp_{\mathbb{L},\boldsymbol{k},N}$. The latter is invertible and such that $\|T^{-1}_{22,\mathbb{L},\boldsymbol{k}}\|_{\mathcal{L}(X^\perp_{\mathbb{L},\boldsymbol{k},N})} \leqslant N^{-2}$ and $\|T^{-1}_{22,\mathbb{L},\boldsymbol{k}}\|_{\mathcal{L}(\mathcal{H}_{A,\mathbb{L}}\cap X^\perp_{\mathbb{L},\boldsymbol{k},N})} \leqslant N^{-2}$. □

# Appendix

## Proof of Theorem 5.4

First, recall that, for all $\varepsilon > 0$,

$$-\varepsilon u_\varepsilon'' + u_\varepsilon + u_\varepsilon^3 = f \tag{5.5.2}$$

and that

$$u_0 + u_0^3 = f. \tag{5.5.3}$$

**$\mathrm{L}^2$-norm convergence** By subtracting (5.5.3) to (5.5.2) and adding $\varepsilon u_0''$ to each side, we get

$$-\varepsilon(u_\varepsilon - u_0)'' + \big(1 + u_\varepsilon^2 + u_0u_\varepsilon + u_0^2\big)(u_\varepsilon - u_0) = \varepsilon u_0''.$$

Multiplying on both sides by $(u_\varepsilon - u_0)$ and integrating over $[0, 2\pi]$ gives

$$\varepsilon\int_0^{2\pi} |(u_\varepsilon - u_0)'|^2 + \int_0^{2\pi} \big(1 + u_\varepsilon^2 + u_0u_\varepsilon + u_0^2\big)|u_\varepsilon - u_0|^2 = \varepsilon\int_0^{2\pi} u_0''(u_\varepsilon - u_0).$$

As $1 + a^2 + ab + b^2 \geqslant 1$ for any $a, b \in \mathbb{R}$, we finally have (using Cauchy–Schwarz inequality for the right-hand side)

$$\varepsilon\|(u_\varepsilon - u_0)'\|^2_{\mathrm{L}^2_\#} + \|u_\varepsilon - u_0\|^2_{\mathrm{L}^2_\#} \leqslant \varepsilon\|u_0''\|_{\mathrm{L}^2_\#}\|u_\varepsilon - u_0\|_{\mathrm{L}^2_\#}. \tag{5.5.4}$$

Thus, we have that

$$\|u_\varepsilon - u_0\|_{\mathrm{L}^2_\#} \leqslant C_1\varepsilon, \quad \text{with} \quad C_1 = \|u_0''\|_{\mathrm{L}^2_\#}.$$

**$\mathrm{H}^1$-norm convergence** From (5.5.4), we already have that there exists $C > 0$ such that

$$\|(u_\varepsilon - u_0)'\|_{\mathrm{L}^2_\#} \leqslant C\sqrt{\varepsilon},$$

which gives the $\mathrm{H}^1$ convergence of $u_\varepsilon$ towards $u_0$, but not at the announced rate. However, as we are working in a 1D setting, this still implies the uniform convergence of $u_\varepsilon$ towards $u_0$. Hence, $\sup_\varepsilon \|u_\varepsilon\|_{\mathrm{L}^\infty_\#} < +\infty$. Starting from here, we can introduce $w_\varepsilon := u_\varepsilon'$ and differentiate (5.5.3) and (5.5.2) to get

$$\begin{cases} -\varepsilon w_\varepsilon'' + (1 + 3u_\varepsilon^2)w_\varepsilon = f'; \\ (1 + 3u_0^2)w_0 = f'. \end{cases}$$

Subtracting both equations yields

$$-\varepsilon(w_\varepsilon - w_0)''\big(1 + 3u_\varepsilon^2\big)(w_\varepsilon - w_0) = \varepsilon w_0'' + 3w_0\big(u_0^2 - u_\varepsilon^2\big).$$

Multiplying on both sides by $(w_\varepsilon - w_0)$ and integrating over $[0, 2\pi]$ gives

$$\varepsilon\int_0^{2\pi} |(w_\varepsilon - w_0)'|^2 + \int_0^{2\pi} \big(1 + 3u_\varepsilon^2\big)|w_\varepsilon - w_0|^2 = \varepsilon\int_0^{2\pi} w_0''(w_\varepsilon - w_0) + 3\int_0^{2\pi} w_0\big(u_0^2 - u_\varepsilon^2\big)(w_\varepsilon - w_0).$$

Then, using first that $1 + 3u_\varepsilon^2 \geqslant 1$ and then that $u_0^2 - u_\varepsilon^2 = (u_0 - u_\varepsilon)(u_0 + u_\varepsilon)$ along with $\mathrm{L}^\infty$ and Cauchy–Schwarz bounds, we have the following inequality:

$$\begin{aligned}\varepsilon\|(w_\varepsilon - w_0)'\|^2_{\mathrm{L}^2_\#} + \|w_\varepsilon - w_0\|^2_{\mathrm{L}^2_\#} &\leqslant \varepsilon\|w_0''\|_{\mathrm{L}^2_\#}\|w_\varepsilon - w_0\|_{\mathrm{L}^2_\#} + 3\|w_0\|_{\mathrm{L}^\infty_\#}\|u_0 + u_\varepsilon\|_{\mathrm{L}^\infty_\#}\|u_0 - u_\varepsilon\|_{\mathrm{L}^2_\#}\|w_\varepsilon - w_0\|_{\mathrm{L}^2_\#} \\ &\leqslant \varepsilon\|w_0''\|_{\mathrm{L}^2_\#}\|w_\varepsilon - w_0\|_{\mathrm{L}^2_\#} + 6\|w_0\|_{\mathrm{L}^\infty_\#}\sup_{\varepsilon'}\|u_{\varepsilon'}\|_{\mathrm{L}^\infty_\#}\|u_0 - u_\varepsilon\|_{\mathrm{L}^2_\#}\|w_\varepsilon - w_0\|_{\mathrm{L}^2_\#}.\end{aligned} \tag{5.5.5}$$

Finally, we have

$$\|u_\varepsilon - u_0\|_{\mathrm{H}^1_\#} \leqslant C_2\varepsilon \quad \text{with} \quad C_2^2 := C_1^2 + \left(\|w_0''\|^2_{\mathrm{L}^2_\#} + 6C_1\|w_0\|_{\mathrm{L}^\infty_\#}\sup_{\varepsilon'}\|u_{\varepsilon'}\|_{\mathrm{L}^\infty_\#}\right)^2.$$

**$\mathrm{H}^2$-norm convergence** Similarly, from (5.5.5), we already have that there exists $C > 0$ such that

$$\|(u_\varepsilon - u_0)''\|_{\mathrm{L}^2_\#} \leqslant C\sqrt{\varepsilon},$$

which gives the $\mathrm{H}^2$ convergence of $u_\varepsilon$ towards $u_0$, but not at the announced rate. However, as we are working in a 1D setting, this still implies the $\mathrm{H}^1$ convergence and thus the uniform convergence of $u'_\varepsilon$ towards $u'_0$. Hence, $\sup_\varepsilon \|u'_\varepsilon\|_{\mathrm{L}^\infty_\#} < \infty$. Starting from here, we can introduce $w_\varepsilon \coloneqq u''_\varepsilon$ and differentiate twice (5.5.3) and (5.5.2) to get

$$\begin{cases} -\varepsilon w''_\varepsilon + w_\varepsilon + 3(2u'^2_\varepsilon u_\varepsilon + u^2_\varepsilon u''_\varepsilon) = f''; \\ w_0 + 3(2u'^2_0 u_0 + u^2_0 w_0) = f'. \end{cases}$$

Subtracting both equations yields

$$-\varepsilon(w_\varepsilon - w_0)''\big(1 + 3u^2_\varepsilon\big)(w_\varepsilon - w_0) = \varepsilon w''_0 + 3w_0\big(u^2_0 - u^2_\varepsilon\big) + 6(u'^2_0 - u'^2_\varepsilon)u_0 + 6(u_0 - u_\varepsilon)u'^2_\varepsilon.$$

Multiplying on both sides by $(w_\varepsilon - w_0)$ and integrating over $[0, 2\pi]$ gives

$$\begin{aligned} \varepsilon \int_0^{2\pi} |(w_\varepsilon - w_0)'|^2 + \int_0^{2\pi} \big(1 + 3u^2_\varepsilon\big)|w_\varepsilon - w_0|^2 &= \varepsilon \int_0^{2\pi} w''_0(w_\varepsilon - w_0) + 3\int_0^{2\pi} w_0\big(u^2_0 - u^2_\varepsilon\big)(w_\varepsilon - w_0) \\ &\quad + 6\int_0^{2\pi} (u'^2_0 - u'^2_\varepsilon)u_0(w_\varepsilon - w_0) + 6\int_0^{2\pi} (u_0 - u_\varepsilon)u'^2_\varepsilon(w_\varepsilon - w_0) \end{aligned}$$

Then, using first that $1+3u^2_\varepsilon \geqslant 1$ and then that $u^2_0 - u^2_\varepsilon = (u_0 - u_\varepsilon)(u_0 + u_\varepsilon)$ and $u'^2_0 - u'^2_\varepsilon = (u'_0 - u'_\varepsilon)(u'_0 + u'_\varepsilon)$ along with $\mathrm{L}^\infty$ and Cauchy–Schwarz bounds, we have the following inequality:

$$\begin{aligned} \varepsilon\|(w_\varepsilon - w_0)'\|^2_{\mathrm{L}^2_\#} + \|w_\varepsilon - w_0\|^2_{\mathrm{L}^2_\#} &\leqslant \varepsilon\|w''_0\|_{\mathrm{L}^2_\#}\|w_\varepsilon - w_0\|_{\mathrm{L}^2_\#} + 3\|w_0\|_{\mathrm{L}^\infty_\#}\|u_0 + u_\varepsilon\|_{\mathrm{L}^\infty_\#}\|u_0 - u_\varepsilon\|_{\mathrm{L}^2_\#}\|w_\varepsilon - w_0\|_{\mathrm{L}^2_\#} \\ &\quad + 6\|u_0\|_{\mathrm{L}^\infty_\#}\|u'_0 + u'_\varepsilon\|_{\mathrm{L}^\infty_\#}\|u'_0 - u'_\varepsilon\|_{\mathrm{L}^2_\#}\|w_\varepsilon - w_0\|_{\mathrm{L}^2_\#} \\ &\quad + 6\big\|u'^2_\varepsilon\big\|_{\mathrm{L}^\infty_\#}\|u_0 - u_\varepsilon\|_{\mathrm{L}^2_\#}\|w_\varepsilon - w_0\|_{\mathrm{L}^2_\#} \\ &\leqslant \varepsilon\|w''_0\|_{\mathrm{L}^2_\#}\|w_\varepsilon - w_0\|_{\mathrm{L}^2_\#} + 6\|w_0\|_{\mathrm{L}^\infty_\#}\sup_{\varepsilon'}\|u_{\varepsilon'}\|_{\mathrm{L}^\infty_\#}\|u_0 - u_\varepsilon\|_{\mathrm{L}^2_\#}\|w_\varepsilon - w_0\|_{\mathrm{L}^2_\#} \\ &\quad + 12\|u_0\|_{\mathrm{L}^\infty_\#}\sup_{\varepsilon'}\|u'_{\varepsilon'}\|_{\mathrm{L}^\infty_\#}\|u'_0 - u'_\varepsilon\|_{\mathrm{L}^2_\#}\|w_\varepsilon - w_0\|_{\mathrm{L}^2_\#} \\ &\quad + 6\sup_{\varepsilon'}\big\|u'^2_{\varepsilon'}\big\|_{\mathrm{L}^\infty_\#}\|u_0 - u_\varepsilon\|_{\mathrm{L}^2_\#}\|w_\varepsilon - w_0\|_{\mathrm{L}^2_\#} \end{aligned} \tag{5.5.6}$$

Finally, we have

$$\|u_\varepsilon - u_0\|_{\mathrm{H}^2_\#} \leqslant C_3\varepsilon$$

$$\text{with} \quad C_3^2 \coloneqq C_2^2 + \left(\|w''_0\|^2_{\mathrm{L}^2_\#} + 6C_1\left(\|w_0\|_{\mathrm{L}^\infty_\#}\sup_{\varepsilon'}\|u_{\varepsilon'}\|_{\mathrm{L}^\infty_\#} + \sup_{\varepsilon'}\big\|u'^2_{\varepsilon'}\big\|_{\mathrm{L}^\infty_\#}\right) + 12C_2\|u_0\|_{\mathrm{L}^\infty_\#}\sup_{\varepsilon'}\|u'_{\varepsilon'}\|_{\mathrm{L}^\infty_\#}\right)^2.$$

## Proof of Theorem 5.5

Let first start by recalling the proof when $d = 1$. We are thus looking at the system

$$\begin{cases} w'(t) = g(t, w(t)), & w(t_0) \in \mathbb{R}, \\ v'(t) \geqslant g(t, v(t)), & v(t_0) = w(t_0), \end{cases}$$

where $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Note that in this case, we only need $x \mapsto g(t, x)$ to be locally Lipschitz and that no quasimonotonicity assumption is needed as there is only one variable. Let $T$ be such that both solutions are defined on $[t_0, T]$. We assume by contradiction that the set $\{t \in (t_0, T],\ v(t) < w(t)\}$ is not empty. We can then define its infimum

$$t_* \coloneqq \inf\{t \in (t_0, T],\ v(t) < w(t)\}.$$

By continuity of $v$ and $w$, we then have $v(t_*) = w(t_*)$ and there exists $\delta > 0$ such that $v(s) < w(s)$ on $(t_*, t_* + \delta]$. Let us now look at $h \coloneqq w - v$. It satisfies

$$h(t_*) = 0 \quad \text{and} \quad h(s) > 0 \text{ on } (t_*, t_* + \delta].$$

However, using the local Lipschitz assumption, it holds, for some $L > 0$,

$$h'(s) = w'(s) - v'(s) \leqslant g(s, w(s)) - g(s, v(s)) \leqslant L|w(s) - v(s)| = Lh(s),$$

because $h(s) \geqslant 0$ for $s \in [t_*, t_* + \delta]$. Therefore, by Grönwall's lemma,

$$\forall\ s \in [t_*, t_* + \delta],\ h(s) \leqslant h(t_*) \exp(L\delta) = 0,$$

which leads to a contradiction. The set $\{t \in (t_0, T],\ v(t) < w(t)\}$ is thus empty and $v(t) \geqslant w(t)$ on $[t_0, T]$.

We now consider the case when $d > 1$, and present a proof adapted from [199, Chapter 3]. We consider solutions on $[t_0, T]$ of the differential system

$$\begin{cases} w'(t) = g(t, w(t)), & w(t_0) \in \mathbb{R}, \\ v'(t) \geqslant g(t, v(t)), & v(t_0) = w(t_0), \end{cases}$$

where $g : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ is such that $g(t, \cdot)$ is locally Lipschitz and quasimonotone for any $t \in \mathbb{R}$. Let assume again by contradiction that the set

$$\{t \in (t_0, T],\ \exists\ i,\ v_i(t) < w_i(t)\}$$

is not empty. Therefore, it has an infimum that we denote by $t_*$. Using the definition of $t_*$ and the continuity of $v$ and $w$, it holds, for some index $i$ and $\delta > 0$,

$$v_i(t_*) = w_i(t_*), \quad v_i(s) < w_i(s),\ \forall\ s \in (t_*, t_* + \delta] \quad \text{and} \quad v_j(t_*) \geqslant w_j(t_*),\ j \neq i.$$

Hence, the quasimonotonicity of $g$ implies in particular that $g_i(t_*, v(t_*)) \geqslant g_i(t_*, w(t_*))$. Moreover, for any $s \in (0, \delta]$,

$$\frac{v_i(t_* + s) - v_i(t_*)}{s} \leqslant \frac{w_i(t_* + s) - w_i(t_*)}{s},$$

which implies, by letting $s \to 0$, that $v_i'(t_*) \leqslant w_i'(t_*)$. Compiling everything thus yields

$$g_i(t_*, v(t_*)) \leqslant v_i'(t_*) \leqslant w_i'(t_*) = g_i(t_*, w(t_*)) \leqslant g_i(t_*, v(t_*)). \tag{5.5.7}$$

Now, two different situations are possible:

- The first case is if $v_i'(t_*) > g_i(t_*, v(t_*))$. This implies that (5.5.7) is false and we reach a contradiction. Thus, $t_*$ cannot be defined and we have $v(t) \geqslant w(t)$ for any $t \in [0, T]$.
- The second case is if $v_i'(t_*) = g_i(t_*, v(t_*))$ and is more subtle. Let us introduce $L$, a Lipschitz constant for $g$ such that
$$|g(t, v + h) - g(t, v)| \leqslant L|h|, \tag{5.5.8}$$
for the maximum norm. Note that $g$ being locally Lipschitz in space, $L$ depends on $T$ (which is fixed) *via* the upper and lower bounds of the $v_i$'s on $[0, T]$. Then, we define
$$h(t) = (\mathrm{e}^{2Lt}, \dots, \mathrm{e}^{2Lt}) \in \mathbb{R}^d.$$
For any $\varepsilon > 0$, (5.5.8) thus yields
$$\varepsilon h'(t) = 2L\varepsilon h(t) > g(t, v(t) + \varepsilon h(t)) - g(t, v(t)) \geqslant g(t, v(t) + \varepsilon h(t)) - v'(t),$$
from which we have $v'(t) + \varepsilon h'(t) > g(t, v(t) + \varepsilon h(t))$. We can thus apply everything that precedes to the function $v + \varepsilon h$: it satisfies the same differential inequality on $[0, T]$, so the reasoning is valid and when we reach the case disjunction, this time we have that, for the associated $t_*$, $v'(t_*) + \varepsilon h'(t_*) > g(t_*, v(t_*) + \varepsilon h(t_*))$ which is the case that implies a contradiction. Thus, $v + \varepsilon h \geqslant w$ on $[0, T]$. As $\varepsilon$ has been chosen arbitrarily, it holds $v(t) \geqslant w(t)$ for any $t \in [0, T]$.

The theorem then follows with $g(t, X) = G(X)$, $\Psi = v$ and $\Phi = w$.

## Proof of Lemma 5.1

We introduce the path in the complex plane, for $t \in [0,T]$,

$$x_\varepsilon(t) \coloneqq \sqrt{\alpha(t)}\omega_\varepsilon(t) + \mathrm{i}\sqrt{\varepsilon}\omega_\varepsilon'(t).$$

Then, we have

$$\begin{aligned}
\mathrm{i}\sqrt{\varepsilon}x_\varepsilon'(t) &= \sqrt{\alpha(t)}\mathrm{i}\sqrt{\varepsilon}\omega_\varepsilon'(t) - \varepsilon\omega_\varepsilon''(t) + \frac{\mathrm{i}\sqrt{\varepsilon}\omega_\varepsilon(t)\alpha'(t)}{2\sqrt{\alpha(t)}} \\
&= \sqrt{\alpha(t)}\mathrm{i}\sqrt{\varepsilon}\omega_\varepsilon'(t) + \alpha(t)\omega_\varepsilon(t) - \varepsilon\beta(t,\omega_\varepsilon(t),\varepsilon) + \frac{\mathrm{i}\sqrt{\varepsilon}\omega_\varepsilon(t)\alpha'(t)}{2\sqrt{\alpha(t)}} \\
&= \sqrt{\alpha(t)}x_\varepsilon(t) + \varepsilon\gamma(t,\omega_\varepsilon(t),\varepsilon), \qquad \text{with} \qquad \gamma(t,\omega_\varepsilon(t),\varepsilon) = -\beta(t,\omega_\varepsilon(t),\varepsilon) + \frac{\mathrm{i}\omega_\varepsilon(t)\alpha'(t)}{2\sqrt{\varepsilon\alpha(t)}}
\end{aligned}$$

and where, because $\alpha(t) \geqslant \alpha_T > 0$, there exists $M_T > 0$ such that

$$|\gamma(t,\omega_\varepsilon(t),\varepsilon)| = \left|\beta(t,\omega_\varepsilon(t),\varepsilon) + \frac{\mathrm{i}\omega_\varepsilon(t)\alpha'(t)}{2\sqrt{\varepsilon\alpha(t)}}\right| \leqslant M_T\left(1 + \frac{|\omega_\varepsilon(t)|}{\sqrt{\varepsilon}} + \frac{|\omega_\varepsilon(t)|^2}{\varepsilon}\right).$$

The constant $M_T$ defined here will be used all along the proof and might change implicitly.

Now, we make the following ansatz:

$$x_\varepsilon(t) = A(t)\exp\left(-\frac{\mathrm{i}}{\sqrt{\varepsilon}}\int_0^t \sqrt{\alpha(s)}\mathrm{d}s\right),$$

where $A : [0,T] \to \mathbb{R}^+$ is the modulus of $x_\varepsilon(t)$ and $A(0) = |x_\varepsilon(0)| = O(\varepsilon^{3/2})$. For this ansatz to be valid, we need

$$\mathrm{i}\sqrt{\varepsilon}x_\varepsilon'(t) = \sqrt{\alpha(t)}x_\varepsilon(t) + \varepsilon\gamma(t,\omega_\varepsilon(t),\varepsilon)$$

to be valid. As we have

$$\mathrm{i}\sqrt{\varepsilon}x_\varepsilon'(t) = \mathrm{i}\sqrt{\varepsilon}A'(t)\exp\left(-\frac{\mathrm{i}}{\sqrt{\varepsilon}}\int_0^t \sqrt{\alpha(s)}\mathrm{d}s\right) + \sqrt{\alpha(t)}x_\varepsilon(t),$$

this implies that $A$ must satisfy

$$\mathrm{i}\sqrt{\varepsilon}A'(t)\exp\left(-\frac{\mathrm{i}}{\sqrt{\varepsilon}}\int_0^t \sqrt{\alpha(s)}\mathrm{d}s\right) = \varepsilon\gamma(t,\omega_\varepsilon(t),\varepsilon) \quad \Leftrightarrow \quad A'(t) = \gamma(t,\omega_\varepsilon(t),\varepsilon)\frac{\sqrt{\varepsilon}}{\mathrm{i}}\exp\left(\frac{\mathrm{i}}{\sqrt{\varepsilon}}\int_0^t \sqrt{\alpha(s)}\mathrm{d}s\right).$$

Direct integration yields

$$A(t) - A(0) = \frac{\sqrt{\varepsilon}}{\mathrm{i}}\int_0^t B(r)C(r)\mathrm{d}r$$

where

$$B(r) \coloneqq \gamma(r,\omega_\varepsilon(r),\varepsilon) \text{ and } C(r) \coloneqq \exp\left(\frac{\mathrm{i}}{\sqrt{\varepsilon}}\int_0^r \sqrt{\alpha(s)}\mathrm{d}s\right).$$

By integrating by parts, we get

$$A(t) - A(0) = \frac{\sqrt{\varepsilon}}{\mathrm{i}}\left(B(t)D(t) - \int_0^t B'(r)D(r)\mathrm{d}r\right) \quad \text{where} \quad D(t) = \int_0^t C(r)\mathrm{d}r. \tag{5.5.9}$$

The first term is easy to deal with. Indeed, if we introduce

$$\widetilde{D}(r) \coloneqq \frac{\sqrt{\varepsilon}}{\mathrm{i}\sqrt{\alpha(r)}}\exp\left(\frac{\mathrm{i}}{\sqrt{\varepsilon}}\int_0^r \sqrt{\alpha(s)}\mathrm{d}s\right),$$

then $\widetilde{D}(r) = O(\sqrt{\varepsilon})$ and

$$\widetilde{D}'(r) = C(r) - \frac{1}{2\alpha(r)^{3/2}}\frac{\sqrt{\varepsilon}}{\mathrm{i}}\exp\left(\frac{\mathrm{i}}{\sqrt{\varepsilon}}\int_0^r \sqrt{\alpha(s)}\mathrm{d}s\right).$$

Thus,

$$D(t) = \int_0^t C(r)\mathrm{d}r = \int_0^t \widetilde{D}'(r) + O(\sqrt{\varepsilon}) = O(\sqrt{\varepsilon}).$$

From (5.5.9) we obtain, with $A(0) = O(\varepsilon^{3/2})$ and $|\omega_\varepsilon(t)| \leqslant A(t)/\sqrt{\alpha_T}$,

$$\begin{aligned}
A(t) &\leqslant M_T\left(\varepsilon^{3/2} + \varepsilon\gamma(t,\omega_\varepsilon(t),\varepsilon) + \varepsilon\int_0^t |B'(r)|\mathrm{d}r\right) \\
&\leqslant M_T\left(\varepsilon^{3/2} + \varepsilon + \sqrt{\varepsilon}|\omega_\varepsilon(t)| + |\omega_\varepsilon(t)|^2 + \varepsilon\int_0^t |B'(r)|\mathrm{d}r\right) \\
&\leqslant M_T\left(\varepsilon^{3/2} + \varepsilon + \sqrt{\varepsilon}A(t) + A(t)^2 + \varepsilon\int_0^t |B'(r)|\mathrm{d}r\right) \\
&\leqslant M_T\left(\varepsilon + \sqrt{\varepsilon}A(t) + A(t)^2 + \varepsilon\int_0^t |B'(r)|\mathrm{d}r\right),
\end{aligned}$$

the last inequality being true for $\varepsilon$ small enough.

We now focus on $B'(r)$: the assumptions on $\beta$ and the definition of $\gamma$ yield

$$|B'(r)| = \left|\frac{\mathrm{d}}{\mathrm{d}r}\gamma(r,\omega_\varepsilon(r),\varepsilon)\right| \leqslant M_T\left(1 + \frac{|\omega_\varepsilon'(r)\omega_\varepsilon(r)|}{\varepsilon} + \frac{|\omega_\varepsilon(r)|}{\sqrt{\varepsilon}} + \frac{|\omega_\varepsilon'(r)|}{\sqrt{\varepsilon}}.\right)$$

As $|\omega_\varepsilon(r)| \leqslant A(r)/\sqrt{\alpha_T}$ and $|\omega_\varepsilon'(r)| \leqslant A(r)/\sqrt{\varepsilon}$, we have

$$|B'(r)| \leqslant M_T\left(1 + \frac{A(r)^2}{\varepsilon^{3/2}} + \frac{A(r)}{\sqrt{\varepsilon}} + \frac{A(r)}{\varepsilon},\right)$$

from which we finally get, for $\varepsilon$ small enough,

$$A(t) \leqslant M_T\left(\varepsilon + \sqrt{\varepsilon}A(t) + A(t)^2 + \int_0^t \frac{A(r)^2}{\sqrt{\varepsilon}} + \sqrt{\varepsilon}A(r) + A(r)\mathrm{d}r\right). \tag{5.5.10}$$

We conclude with a bootstrapping argument. Let assume that $A(t) \leqslant C_T\varepsilon$ on $[0,T]$ for some $C_T > 0$. We are going to show that, for a well chosen $C_T$, we can find $\varepsilon_T$ small enough such that, for all $\varepsilon \leqslant \varepsilon_T$, we have indeed $A(t) \leqslant C_T\varepsilon$ on $[0,T]$. If such a bound is true, then (5.5.10) implies that

$$\begin{aligned}
A(t) &\leqslant M_T\left(\varepsilon + C_T\varepsilon^{3/2} + C_T^2\varepsilon^2 + 2TC_T\varepsilon^{3/2} + \int_0^t A(r)\mathrm{d}r\right) \\
&\leqslant \varepsilon M_T\big(1 + C_T\sqrt{\varepsilon}(1 + C_T\sqrt{\varepsilon} + 2TC_T)\big) + M_T\int_0^t A(r)\mathrm{d}r.
\end{aligned}$$

Thus, by applying the integral form of Grönwall's lemma, we get that

$$\forall\ t \in [0,T],\ A(t) \leqslant \varepsilon M_T\big(1 + C_T\sqrt{\varepsilon}(1 + C_T\sqrt{\varepsilon} + 2TC_T)\big)\exp(TM_T).$$

Therefore, by choosing for instance $C_T = 2M_T\exp(TM_T)$, we can find $\varepsilon_T$ small enough such that for any $\varepsilon \leqslant \varepsilon_T$,

$$\forall\ t \in [0,T],\ A(t) \leqslant C_T\varepsilon,$$

which validates the initial assumption. The proof of the lemma follows by recalling that $|\omega_\varepsilon(t)| \leqslant A(t)/\sqrt{\alpha_T}$ and $|\omega_\varepsilon'(t)| \leqslant A(t)/\sqrt{\varepsilon}$.

Chapter 6

# On basis set optimization in quantum chemistry

This chapter, which is the result of the CEMRACS 2021 research school, has been published in the article [GK3]:

*Eric Cancès, Geneviève Dusson, Gaspard Kemlin and Laurent Vidal. On basis set optimization in quantum chemistry. Accepted in ESAIM Proceedings.* https://arxiv.org/abs/2207.12190.

**Abstract** In this article, 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

**6.1 Introduction** . . . . . **128**
**6.2 Optimization criteria** . . . . . **129**
**6.3 Application to 1D toy model** . . . . . **131**
6.3.1 Description of the model . . . . . 131
6.3.2 Variational approximation in AO basis sets . . . . . 132
6.3.3 Overcompleteness of Hermite Basis Sets . . . . . 134
6.3.4 Practical computation of the criterion $J_A$ and $J_E$ . . . . . 134
**6.4 Numerical results** . . . . . **136**
6.4.1 Numerical setting and first results . . . . . 136
6.4.2 Influence of numerical parameters . . . . . 142

## 6.1 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 nonvariational (*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 [95, 156] 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 $\chi_\mu^z \in \mathrm{H}^1(\mathbb{R}^3)$ a fast decaying function centered at the origin called an atomic orbital. Consider an atomic configuration $\omega$ consisting of $M$ nuclei with nuclear charges $z_1,\dots,z_M$ (in atomic units) and positions $\boldsymbol{R}_1,\dots,\boldsymbol{R}_M$ in the three dimensional physical space. If the AO basis set $\chi$ 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 $\omega$ is discretized is

$$\mathcal{X}_\omega \coloneqq \mathrm{Span}(\chi_1^{z_1}(\cdot-\boldsymbol{R}_1),\dots,\chi_{n_{z_1}}^{z_1}(\cdot-\boldsymbol{R}_1),\dots,\chi_1^{z_M}(\cdot-\boldsymbol{R}_M),\dots,\chi_{n_{z_M}}^{z_M}(\cdot-\boldsymbol{R}_M)).$$

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 [46], and share common features with the modal synthesis method [51, Chapter 7], [50]. 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 [140]. 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 [185], with exponential decay and a cusp at the origin. It was then realized by Boys [25] 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

$$\int_{\mathbb{R}^3} \chi_i^{z_a}(\boldsymbol{r}-\boldsymbol{R}_a)\,\chi_j^{z_b}(\boldsymbol{r}-\boldsymbol{R}_b)\,\mathrm{d}\boldsymbol{r}, \qquad \int_{\mathbb{R}^3} \nabla\chi_i^{z_a}(\boldsymbol{r}-\boldsymbol{R}_a)\cdot\nabla\chi_j^{z_b}(\boldsymbol{r}-\boldsymbol{R}_b)\,\mathrm{d}\boldsymbol{r},$$
$$\int_{\mathbb{R}^3} \frac{\chi_i^{z_a}(\boldsymbol{r}-\boldsymbol{R}_a)\,\chi_j^{z_b}(\boldsymbol{r}-\boldsymbol{R}_b)}{|\boldsymbol{r}-\boldsymbol{R}_k|}\,\mathrm{d}\boldsymbol{r}, \qquad \int_{\mathbb{R}^3\times\mathbb{R}^3} \frac{\chi_i^{z_a}(\boldsymbol{r}-\boldsymbol{R}_a)\,\chi_j^{z_b}(\boldsymbol{r}-\boldsymbol{R}_b)\,\chi_k^{z_c}(\boldsymbol{r}'-\boldsymbol{R}_c)\,\chi_\ell^{z_d}(\boldsymbol{r}'-\boldsymbol{R}_d)}{|\boldsymbol{r}-\boldsymbol{r}'|}\,\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{r}',$$

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 [146], 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 50s: STO-$n$g basis sets [93] were built as the contraction of $n$ Gaussians that fit Clementi STO SCF AOs in an $\mathrm{L}^2$ least-squares sense [189]. 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. [20]. Basis sets better suited for correlated electronic structure methods were then introduced, notably Atomic Natural Orbitals (ANO) [2] and Dunning basis sets [66]. 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 [120], and later on in [182, 183]. An *a priori* error estimate on the approximation of Slater-type functions by Hermite and even-tempered Gaussian functions was derived in [10]. A construction of Gaussian bases combined with wavelets was proposed on a one-dimensional toy model in [164].

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 [55, 180] 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 chapter, we propose a different approach, which is adaptable to any criterion one might be interested in, and involves molecular configurations. In Section 6.2, 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 $\Omega$;
2. a probability measure $\mathbb{P}$ on $\Omega$;
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 $\Omega$, $\mathbb{P}$, $\mathcal{B}$, and $j$. As a proof of concept (Section 6.3), we apply this strategy to a simple toy model of a 1D homonuclear diatomic "molecule" with two 1D noninteracting spinless "electrons", which allows for extremely accurate reference calculations. Finally, we present numerical results in Section 6.4, 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".

## 6.2 Optimization criteria

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 $\Omega$ 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 $\Omega$ 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 $\omega$'s in the support of $\mathbb{P}$. For practical reasons we therefore need to choose $\mathbb{P}$ of the form

$$\mathbb{P} = \sum_{n=1}^{N_c} \beta_n \delta_{\omega_n}, \tag{6.2.1}$$

where $\{\omega_1, \ldots, \omega_{N_c}\}$ is a finite (not too large) subset of $\Omega$, $\delta_{\omega_n}$ the Dirac mass at $\omega_n$, 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 $\Omega$ 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 $\omega_n$'s and $\beta_n$'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 $N_{\mathrm{c}}$ according to the available computational means;

iii) compute the $\omega_n$'s and $\beta_n$'s using *e.g.* quantization algorithms [147] possibly based on optimal transport or clustering algorithms [158].

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 $\Omega$, 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 polyhedron of $\mathbb{R}^d$.

Given an atomic configuration $\omega \in \Omega$ and an AO basis set $\chi \in \mathcal{B}$, we denote by $\chi_\omega$ the one-electron finite-dimensional space obtained by using the AO basis set $\chi$ to describe the electronic structure of a molecular system with atomic configuration $\omega$ 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 $\omega$. 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

$$j_E(\chi,\omega) \coloneqq |E_\omega - E_\omega^\chi|^2, \tag{6.2.2}$$

where $E_\omega$ is the exact ground-state energy of the neutral system with atomic configuration $\omega$ for the chosen continuous model (*e.g.* Hartree–Fock, MCSCF, Kohn–Sham B3LYP…) and $E_\omega^\chi$ the ground-state energy obtained with the model discretized in the AO basis set $\chi$. 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

$$j_A(\chi,\omega) \coloneqq -\mathrm{Tr}\big(\Pi_{\chi_\omega}^A \gamma_\omega \Pi_{\chi_\omega}^A A\big), \tag{6.2.3}$$

where $A$ is a given self-adjoint, positive, definite operator on the one-particle state space $\mathcal{H}$ with form domain $Q(A)$, $\gamma_\omega$ the exact ground-state 1-RDM of the neutral system with atomic configuration $\omega$ 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 $Q(A) = \mathcal{H}$ 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 $\mathrm{H}^1(\mathbb{R}^3)$, and $\Pi_{\chi_\omega}^A$ is the orthogonal projector on $\mathcal{X}_\omega$ for the $\mathrm{H}^1$-inner product. Diagonalizing $\gamma_\omega$ 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 $\omega$, 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 $\Omega$ with respect to the probability measure $\mathbb{P}$

$$J(\chi) \coloneqq \int_\Omega j(\chi,\omega)\mathrm{d}\mathbb{P}(\omega), \tag{6.2.4}$$

and the problem of basis set optimization can be formulated as

$$\boxed{\text{find } \chi_0 \in \operatorname*{argmin}_{\chi\in\mathcal{B}} J(\chi)}$$

In the following, $J_E$ and $J_A$ denote the evaluation of the criterion (6.2.4) with $j = j_E$ and $j = j_A$ respectively.

**Remark 6.1** (Reference solutions)**.** The evaluation of criteria $J_E$ and $J_A$ hinges on the knowledge of exact GS energy $E_\omega$ or 1-RDM $\gamma_\omega$ for $\omega$ 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 $\omega$. This is the reason why the probability measure $\mathbb{P}$ can only be a finite weighted sum of Dirac distributions, as defined in (6.2.1).

## 6.3 Application to 1D toy model

In this section, we focus on a linear one-dimensional toy model, mimicking a homonuclear diatomic molecule.

### 6.3.1 Description of the model

Let us consider a system of two 1D point-like "nuclei" and two 1D spinless noninteracting quantum "electrons". The one-particle state space is then $\mathcal{H} = \mathrm{L}^2(\mathbb{R})$ and the configuration space $\Omega = (0, +\infty)$. In this section, a configuration of $\Omega$ 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

$$H_a = -\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + V_a, \tag{6.3.1}$$

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

$$\forall\, x \in \mathbb{R}, \quad V_a(x) = \frac{1}{8a^2+4}(x-a)^2(x+a)^2. \tag{6.3.2}$$

Several considerations led us to define the potential as such. First, $V_a$ is designed so that i) each $H_a$ admits a nondegenerate 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

$$E_a = \mathrm{Tr}(H_a\gamma_a) = \min_{\gamma\in\mathcal{G}_2} \mathrm{Tr}(H_a\gamma), \tag{6.3.3}$$

where

$$\mathcal{G}_2 \coloneqq \big\{\gamma \in \mathcal{L}(\mathrm{L}^2(\mathbb{R})),\ \gamma^2 = \gamma = \gamma^*,\ \mathrm{Tr}(\gamma) = 2\big\},$$

$\mathcal{L}(\mathrm{L}^2(\mathbb{R}))$ denoting the space of bounded linear operators on $\mathrm{L}^2(\mathbb{R})$. The existence and uniqueness of the solution to problem (6.3.3) 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 (6.3.1) corresponds to the quartic oscillator, for which we have reference numerical solutions (*e.g.* [22]). 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 (6.3.3) 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 6.1 the potential $V_a$ for two different values of $a$.

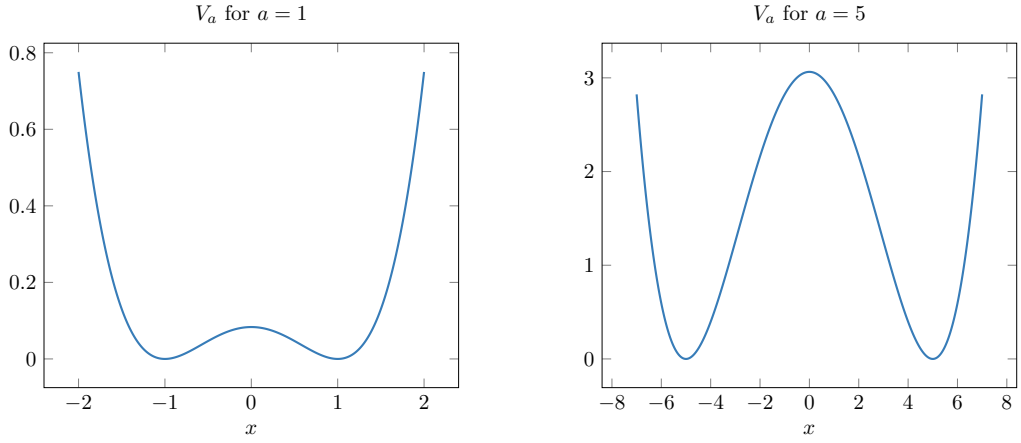



FIGURE 6.1 – $x \mapsto V_a(x)$ for $a = 1$ and $a = 5$.

In practice, it is convenient to compute $\gamma_a$ 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

$$\begin{cases} H_a \varphi_{a,i} = \lambda_{a,i} \varphi_{a,i}, & i = 1, 2 \\ \langle \varphi_{a,i} | \varphi_{a,j} \rangle = \delta_{ij}, & i, j = 1, 2, \end{cases} \tag{6.3.4}$$

$\langle \cdot | \cdot \rangle$ denoting the $\mathrm{L}^2$ inner product. We indeed have

$$E_a = \lambda_{a,1} + \lambda_{a,2} \quad \text{and} \quad \gamma_a = |\varphi_{a,1}\rangle\langle\varphi_{a,1}| + |\varphi_{a,2}\rangle\langle\varphi_{a,2}|. \tag{6.3.5}$$

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 (6.3.4) for the configurations $a_n$ in the support of the chosen atomic probability measure

$$\mathbb{P} = \sum_{n=1}^{N_c} \beta_n \delta_{a_n}, \quad 0 < a_1 < a_2 < \cdots < a_{N_c}, \quad \beta_n > 0, \quad \sum_{n=1}^{N_c} \beta_n = 1. \tag{6.3.6}$$

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 (6.3.9) gives rise to the eigenvalue problem

$$\begin{cases} H_a^{\mathrm{FD}} \varphi_{a,i}^{FD} = \lambda_{a,i}^{\mathrm{FD}} \varphi_{a,i}^{\mathrm{FD}} & i = 1, 2 \\ \delta x \big(\varphi_{a,i}^{\mathrm{FD}}\big)^T \varphi_{a,j}^{\mathrm{FD}} = \delta_{ij}, \end{cases} \tag{6.3.7}$$

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

$$E_a^{\mathrm{FD}} = \lambda_{a,1}^{\mathrm{FD}} + \lambda_{a,2}^{\mathrm{FD}} \quad \text{and} \quad P_a^{\mathrm{FD}} = \varphi_{a,1}^{\mathrm{FD}} \big(\varphi_{a,1}^{\mathrm{FD}}\big)^T + \varphi_{a,2}^{\mathrm{FD}} \big(\varphi_{a,2}^{\mathrm{FD}}\big)^T \in \mathbb{R}_{\mathrm{sym}}^{N_g \times N_g}, \tag{6.3.8}$$

where $P_a^{\mathrm{FD}}$ 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 $\gamma_a$ at the grid points.

### 6.3.2 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 (6.3.4) 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 $\chi$, the first

one translated to $a$, and the second one to $-a$:

$$\chi_{a,1} = \chi_1(\cdot - a), \dots, \chi_{a,N_\mathrm{b}} = \chi_{N_\mathrm{b}}(\cdot - a),\ \chi_{a,N_\mathrm{b}+1} = \chi_1(\cdot + a), \dots, \chi_{a,2N_\mathrm{b}} = \chi_{N_\mathrm{b}}(\cdot + a).$$

Defining the Hamiltonian matrix

$$H_a^\chi = \left( \left\langle \chi_{a,\mu} \middle| \left( -\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + V_a \right) \middle| \chi_{a,\nu} \right\rangle \right)_{1 \leqslant \mu,\nu \leqslant 2N_\mathrm{b}}$$

and the overlap matrix

$$S_a^\chi = \left( \langle \chi_{a,\mu} | \chi_{a,\nu} \rangle \right)_{1 \leqslant \mu,\nu \leqslant 2N_\mathrm{b}},$$

the discretization of problem (6.3.4) in the AO basis set $\chi$ then reads as the generalized eigenvalue problem: find $\left(C_{a,i}^\chi, \lambda_{a,i}^\chi\right) \in \mathbb{R}^{2N_\mathrm{b}} \times \mathbb{R}$, $i = 1, 2$ such that

$$\begin{cases} H_a^\chi C_{a,i}^\chi = \lambda_{a,i}^\chi S_a^\chi C_{a,i}^\chi & i = 1, 2 \\ \left(C_{a,i}^\chi\right)^T S_a^\chi C_{a,j}^\chi = \delta_{ij}. \end{cases} \tag{6.3.9}$$

The approximation $\varphi_{a,i}^\chi$ of $\varphi_{a,i}$ in the AO basis set $\chi$ can then be recovered as the linear combination of atomic orbitals (LCAO)

$$\forall x \in \mathbb{R}, \quad \varphi_{a,i}^\chi(x) = \sum_{\mu=1}^{2N_\mathrm{b}} [C_{a,i}^\chi]_\mu \chi_{a,\mu}(x). \tag{6.3.10}$$

One way to compare the LCAO ground-state 1-RDM to the reference FD solution $P_a^\mathrm{FD}$ is to simply evaluate the former at the grid points $x_j$, which gives rise to the matrix $P_a^\chi \in \mathbb{R}_\mathrm{sym}^{N_\mathrm{g} \times N_\mathrm{g}}$ 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 $P_a^\chi$ 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 $\chi_a$ is represented by a matrix $X_a \in \mathbb{R}^{N_\mathrm{g} \times 2N_\mathrm{b}}$. For any vectors $Y_1, Y_2 \in \mathbb{R}^{N_\mathrm{g}}$, 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 $\mathbb{R}^{N_\mathrm{g}}$. Solutions to (6.3.9) are then obtained by approximating respectively the Hamiltonian and overlap matrix by

$$H_a^\chi \simeq H_a^X \coloneqq \left( \delta x X_{a,\mu}^T H_a^\mathrm{FD} X_{a,\nu} \right)_{1 \leqslant \mu,\nu \leqslant 2N_\mathrm{b}}, \quad S_a^\chi \simeq S_a^X \coloneqq \left( \delta x X_{a,\mu}^T X_{a,\nu} \right)_{1 \leqslant \mu,\nu \leqslant 2N_\mathrm{b}},$$

and finding $(C_{a,i}^X, \lambda_{a,i}^X) \in \mathbb{R}^{2N_\mathrm{b}} \times \mathbb{R}$, $i = 1, 2$, such that

$$\begin{cases} H_a^X C_{a,i}^X = \lambda_{a,i}^X S_a^X C_{a,i}^X, & i = 1, 2 \\ (C_{a,i}^X)^T S_a^X C_{a,j}^X = \delta_{ij}, & i, j = 1, 2, \end{cases} \tag{6.3.11}$$

from which we get the discrete approximations

$$\varphi_{a,i}^X = X_a C_{a,i}^X, \quad i = 1, 2. \tag{6.3.12}$$

Let us gather the coefficients $C_{a,i}^X$ into the $2N_\mathrm{b} \times 2$ matrix $C_a^X = \left( C_{a,1}^X \,|\, C_{a,2}^X \right)$. The ground-state density matrix in the basis $\chi_a$ is approximated by

$$P_a^X = \varphi_{a,1}^X (\varphi_{a,1}^X)^T + \varphi_{a,2}^X (\varphi_{a,2}^X)^T = \left( X_a C_a^X \right) \left( X_a C_a^X \right)^T \in \mathbb{R}_\mathrm{sym}^{N_\mathrm{g} \times N_\mathrm{g}}. \tag{6.3.13}$$

### 6.3.3 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

$$h_n(x) = c_n p_n(x)\exp\left(-\frac{x^2}{2}\right), \quad \varepsilon_n = n+\frac{1}{2}, \quad n\in\mathbb{N}, \tag{6.3.14}$$

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 $\mathrm{L}^2(\mathbb{R})$. 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 $N_\mathrm{b}$ Hermite functions

$$\chi^{\mathrm{HBS}} = \{\chi_\mu^{\mathrm{HBS}}\}_{1\leqslant\mu\leqslant N_\mathrm{b}} = \{h_n\}_{0\leqslant n\leqslant N_\mathrm{b}-1}.$$

The overlap matrix for the configuration $a$ then is of the form

$$S_a^{\chi^{\mathrm{HBS}}} = \begin{pmatrix} I_{N_\mathrm{b}} & \Sigma_a \\ \Sigma_a^T & I_{N_\mathrm{b}} \end{pmatrix} \quad \text{where} \quad \Sigma_a \coloneqq (\langle h_n(\cdot-a)|h_m(\cdot+a)\rangle)_{0\leqslant n,m\leqslant N_\mathrm{b}-1}.$$

The matrix $\Sigma_a$ corresponds to the overlap of functions that are localized at different atomic positions. It satisfies $\Sigma_a \simeq 0$ when $a$ is large and $\Sigma_a \simeq I_{N_\mathrm{b}}$ 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 6.2, 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 [140] or more recent works based on a Cholesky decomposition of the overlap matrix [123]. Such methods are however not directly related to the optimization procedure presented in this paper.

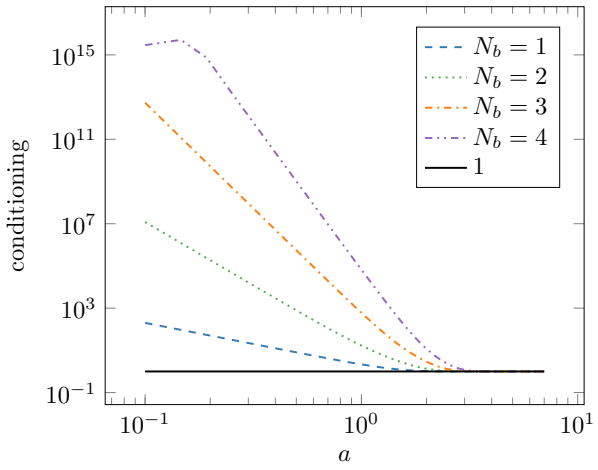



FIGURE 6.2 – Condition number of the HBS overlap matrix $S_a^{\chi^{\mathrm{HBS}}}$ for different values of $a$ in log-log scale. The larger the basis set, the faster the condition number blows up for small values of $a$.

### 6.3.4 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.

### Reference orthonormal basis

In order to avoid potential numerical stability issues, each of the $N_\mathrm{b}$ atomic orbital $\chi_\mu$ is decomposed on a given truncated orthonormal basis of $\mathrm{L}^2(\mathbb{R})$ of size $\mathcal{N}$ such that $N_\mathrm{b} \ll \mathcal{N} \ll N_\mathrm{g}$. We choose here the orthonormal basis introduced in (6.3.14). Hence, the matrix $X_a \in \mathbb{R}^{N_\mathrm{g} \times 2N_\mathrm{b}}$ is written as

$$X_a = B_a I_R, \tag{6.3.15}$$

with

$$B_a = \big(h_0(x_\cdot - a)| \cdots |h_{\mathcal{N}-1}(x_\cdot - a)|h_0(x_\cdot + a)| \cdots |h_{\mathcal{N}-1}(x_\cdot + a)\big) \in \mathbb{R}^{N_\mathrm{g} \times 2\mathcal{N}},$$

and

$$I_R = \begin{pmatrix} R & 0 \\ 0 & R \end{pmatrix} \in \mathbb{R}^{(2\mathcal{N}) \times (2N_\mathrm{b})}, \tag{6.3.16}$$

where $R \in \mathbb{R}^{\mathcal{N} \times N_\mathrm{b}}$ gathers the coefficients of the atomic orbitals $\chi_\mu$ 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 $R^T R = I_{N_\mathrm{b}}$, so that the overlap matrix of $X_a$, denoted by $S(X_a)$, has the same form as in Section 6.3.3, that is

$$S(X_a) \coloneqq \delta x X_a^T X_a = \begin{pmatrix} I_{N_\mathrm{b}} & \Sigma_a \\ \Sigma_a^T & I_{N_\mathrm{b}} \end{pmatrix}, \tag{6.3.17}$$

where $\Sigma_a$ 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 $a_\mathrm{min}$ 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 $\mathbb{R}^{N_\mathrm{g}}$ is rescaled by a factor $\sqrt{\delta x}$ so that for any given $Y_1, Y_2 \in \mathbb{R}^{N_\mathrm{g}}$ 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.

### 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 $Y \in \mathbb{R}^{N_\mathrm{g} \times d}$. Since the columns of $X_a[S^A(X_a)]^{-\frac{1}{2}}$ are orthonormal for the $A$ inner product, that is

$$\Big(X_a[S^A(X_a)]^{-\frac{1}{2}}\Big)^T A \Big(X_a[S^A(X_a)]^{-\frac{1}{2}}\Big) = I,$$

the projection $\Pi_{X_a}^A$ takes the simple form

$$\Pi_{X_a}^A = \Big(X_a[S^A(X_a)]^{-\frac{1}{2}}\Big)\Big(X_a[S^A(X_a)]^{-\frac{1}{2}}\Big)^T A = X_a[S^A(X_a)]^{-1} X_a^T A. \tag{6.3.18}$$

Hence, using the cyclicity of the trace and definitions (6.2.3), (6.3.8) and (6.3.18), one has

$$\begin{aligned} j_A(\chi, a) &\simeq -\mathrm{Tr}\big(P_a^\mathrm{FD}\, \Pi_{X_a}^A A \Pi_{X_a}^A\big) \\ &= -\mathrm{Tr}\big(P_a^\mathrm{FD} \times (AB_aI_R)[S^A(B_aI_R)]^{-1}(AB_aI_R)^T\big) \\ &= -\mathrm{Tr}\big(M_A^\mathrm{offline}(a) I_R [S^A(B_aI_R)]^{-1} I_R^T\big), \end{aligned}$$

where we have collected in the last expression all matrices independent of $R$ into the matrix

$$M_A^\mathrm{offline}(a) = (AB_a)^T P_a^\mathrm{FD} AB_a \in \mathbb{R}^{2\mathcal{N} \times 2\mathcal{N}}. \tag{6.3.19}$$

Then, using the probability measure $\mathbb{P}$ in (6.3.6), we get

$$J_A(R) = -\int_\Omega \mathrm{Tr}\big(M_A^\mathrm{offline}(a) I_R [S^A(B_aI_R)]^{-1} I_R^T\big)\, \mathrm{d}\mathbb{P}(a) = -\sum_{n=1}^{N_\mathrm{c}} \beta_n \mathrm{Tr}\big(M_A^\mathrm{offline}(a_n) I_R [S^A(B_{a_n}I_R)]^{-1} I_R^T\big)$$

and the optimization problem finally writes, with unknown $R \in \mathbb{R}^{\mathcal{N}\times N_\mathrm{b}}$ and for a given inner product $A$

$$\boxed{\text{Find } R_\mathrm{opt} \in \underset{R\in\mathbb{R}^{\mathcal{N}\times N_\mathrm{b}},\ R^TR=I_{N_\mathrm{b}}}{\operatorname{argmin}} J_A(R)} \tag{6.3.20}$$

**Criterion $J_E$**

Let again $a \in \mathbb{R}_+$ be fixed. We denote by

$$G(N_\mathrm{g},2) \coloneqq \{P \in \mathbb{R}^{N_\mathrm{g}\times N_\mathrm{g}} \mid P^2 = P = P^T, \mathrm{Tr}(P) = 2\}$$

the discrete counterpart of the Grassmann manifold $\mathcal{G}_2$, and write $E_a^R$ (resp. $H_a^R$) instead of $E_a^\chi$ (resp. $H_a^\chi$), so that the dependence in the matrix $R$ appears explicitly. Equation (6.3.3) reads in the discrete setting

$$\begin{aligned} E_a^R = \min_{P\in G(N_\mathrm{g},2)} \mathrm{Tr}\big(PH_a^R\big) &= \min_{\substack{C\in\mathbb{R}^{2N_\mathrm{b}\times 2}\\ (C)^TS(B_aI_R)C=I_2}} \mathrm{Tr}\big(CC^T \times (B_aI_R)^T H_a^\mathrm{FD}(B_aI_R)\big) \\ &= \mathrm{Tr}\big(C_a^R(C_a^R)^T \times I_R^T M_E^\mathrm{offline}(a) I_R\big) \end{aligned} \tag{6.3.21}$$

where, as for the previous case, all matrices independent of $R$ have been gathered in the matrix

$$M_E^\mathrm{offline}(a) = B_a^T H_a^\mathrm{FD} B_a, \tag{6.3.22}$$

and the matrix $C_a^R$ is solution to the minimization problem

$$\min_{\substack{C^R\in\mathbb{R}^{2N_\mathrm{b}\times 2}\\ (C^R)^TS(B_aI_R)C^R=I_2}} \mathrm{Tr}\big(C^R(C^R)^T \times I_R^T M_E^\mathrm{offline}(a) I_R\big) \tag{6.3.23}$$

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

$$[S(B_aI_R)]^{-\frac{1}{2}} I_R M_E^\mathrm{offline}(a) I_R^T [S(B_aI_R)]^{-\frac{1}{2}}.$$

From (6.3.6) and (6.3.21), one can compute

$$J_E(R) = \int_\Omega \left|E_a^\mathrm{FD} - E_a^R\right|^2 \mathrm{d}\mathbb{P}(a) = \sum_{n=1}^{N_\mathrm{c}} \beta_n \left|E_{a_n}^\mathrm{FD} - E_{a_n}^R\right|^2$$

and the optimization problem reads

$$\boxed{\text{Find } R_\mathrm{opt} \in \underset{R\in\mathbb{R}^{\mathcal{N}\times N_\mathrm{b}},\ R^TR=I_{N_\mathrm{b}}}{\operatorname{argmin}} J_E(R)} \tag{6.3.24}$$

## 6.4 Numerical results

### 6.4.1 Numerical setting and first results

Problems (6.3.20) and (6.3.24) are solved by direct minimization algorithms over the Stiefel manifold [1]

$$\mathrm{St}(\mathcal{N}, N_\mathrm{b}) = \{R \in \mathbb{R}^{\mathcal{N}\times N_\mathrm{b}} \mid R^TR = I_{N_\mathrm{b}}\}.$$

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 [149] in the `Julia` language [19]. As initial guess, we picked the first $N_\mathrm{b}$ Hermite functions introduced in Section 6.3.3.

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 6.3.3. More precisely, all the results in this subsection are obtained with the probability

$$\mathbb{P} = \frac{1}{10}\sum_{n=1}^{10} \delta_{a_n} \quad \text{with } a_n = 1.5 + (n-1)\frac{3.5}{9}. \tag{6.4.1}$$

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 6.4.2 and 6.4.2.

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 $\mathrm{L}^2(\mathbb{R})$ of size $\mathcal{N} = 10$. Regarding the choice of the inner product for the first criterion $J_A$, we used the standard $\mathrm{L}^2(\mathbb{R})$ and the $\mathrm{H}^1(\mathbb{R})$ inner products, and denoted by $J_{\mathrm{L}^2}$ and $J_{\mathrm{H}^1}$ the corresponding. This translates at the discrete level by choosing $A = I_{N_g}$ for $J_{\mathrm{L}^2}$ and $A = I_{N_g} - \Delta$ for $J_{\mathrm{H}^1}$ where $\Delta$ 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 (6.3.11) 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. $\rho_a$) stands for the reference energy (resp. density) on the finite difference grid. Note that we write HBS for the (nonoptimized) Hermite basis set, and $\mathrm{L}^2$-OBS, $\mathrm{H}^1$-OBS or $E$-OBS for optimized basis sets with respect to the criterion $J_{\mathrm{L}^2}$, $J_{\mathrm{H}^1}$, or $J_E$.

Figure 6.3 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 6.1 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 6.4, 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 $\mathrm{L}^1$-norm, which corresponds to the $\mathrm{L}^2$-norm on eigenvectors, the $\mathrm{H}^1$-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 $\mathrm{H}^{-1}$-norm of $\nabla_a V_a$ (see *e.g.* [GK2]), and the distance

$$\|\nabla\sqrt{\rho_1} - \nabla\sqrt{\rho_2}\|_{\mathrm{L}^2}$$

(recall that the von-Weizsäcker kinetic energy reads $\frac{1}{2}\int_{\mathbb{R}} |\nabla\sqrt{\rho}|^2$). We observe similar behaviours 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 6.1 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 $J_{\mathrm{L}^2}$ and $J_{\mathrm{H}^1}$ 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 $J_{\mathrm{L}^2}$ and $J_{\mathrm{H}^1}$ give similar results for any number of basis functions $N_b$, so that the $\mathrm{L}^2$ and $\mathrm{H}^1$ norm optimizations seem equivalent.

In terms of computational time, first note that criterion $J_{\mathrm{H}^1}$ is always more expensive to compute than $J_{\mathrm{L}^2}$ 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 $J_{\mathrm{L}^2}$ in our implementation.

Finally, for the sake of completeness, we plot in Figure 6.5 the different basis functions built with each

[1] https://github.com/gkemlin/1D_basis_optimization

criterion for different values of $N_\mathrm{b}$, confirming again the previous observations that the optimal basis functions are quite close to the standard Hermite basis functions.

The main conclusion of these observations is that, for $N_\mathrm{b}$ 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 6.5), building optimal bases allows to increase accuracy on the quantities of interest we focused on by one order of magnitude in average.

Value of $J_{\mathrm{L}^2}$ for the different basis sets

| Basis | $N_\mathrm{b}=1$ | $N_\mathrm{b}=2$ | $N_\mathrm{b}=3$ | $N_\mathrm{b}=4$ |
|---|---|---|---|---|
| HBS | -7.40829 | -7.70051 | -7.74312 | -7.77138 |
| $\mathrm{L}^2$-OBS | -7.43954 | -7.76479 | -7.77725 | -7.77773 |
| $\mathrm{H}^1$-OBS | -7.43928 | -7.76466 | -7.77724 | -7.77772 |
| E-OBS | -7.39410 | -7.76425 | -7.77720 | -7.77772 |

Value of $J_{\mathrm{H}^1}$ for the different basis sets

| Basis | $N_\mathrm{b}=1$ | $N_\mathrm{b}=2$ | $N_\mathrm{b}=3$ | $N_\mathrm{b}=4$ |
|---|---|---|---|---|
| HBS | -10.5613 | -11.0566 | -11.1451 | -11.2402 |
| $\mathrm{L}^2$-OBS | -10.6256 | -11.2338 | -11.2630 | -11.2650 |
| $\mathrm{H}^1$-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 | $N_\mathrm{b}=1$ | $N_\mathrm{b}=2$ | $N_\mathrm{b}=3$ | $N_\mathrm{b}=4$ |
|---|---|---|---|---|
| HBS | $3.77956\times10^{-2}$ | $3.98301\times10^{-3}$ | $1.86537\times10^{-3}$ | $1.35309\times10^{-4}$ |
| $\mathrm{L}^2$-OBS | $6.52016\times10^{-2}$ | $2.18282\times10^{-4}$ | $1.01365\times10^{-6}$ | $3.22260\times10^{-8}$ |
| $\mathrm{H}^1$-OBS | $6.83537\times10^{-2}$ | $2.40548\times10^{-4}$ | $1.27251\times10^{-6}$ | $3.91885\times10^{-8}$ |
| E-OBS | $3.69610\times10^{-2}$ | $1.92087\times10^{-4}$ | $6.93394\times10^{-7}$ | $2.54014\times10^{-8}$ |

L–BFGS iterations

| Basis | $N_\mathrm{b}=1$ | $N_\mathrm{b}=2$ | $N_\mathrm{b}=3$ | $N_\mathrm{b}=4$ |
|---|---|---|---|---|
| $\mathrm{L}^2$-OBS | 4 | 13 | 48 | 219 |
| $\mathrm{H}^1$-OBS | 7 | 17 | 235 | not converged after 500 it |
| E-OBS | 6 | 19 | 52 | 134 |

TABLE 6.1 – (Top & Middle) Values of the different criteria for the HBS and optimal bases, for increasing values of $N_\mathrm{b}$. (Bottom) Number of iterations of L–BFGS required for each criterion to achieve convergence up to requested tolerance ($10^{-7}$ on the $\ell^2$-norm of the gradient).

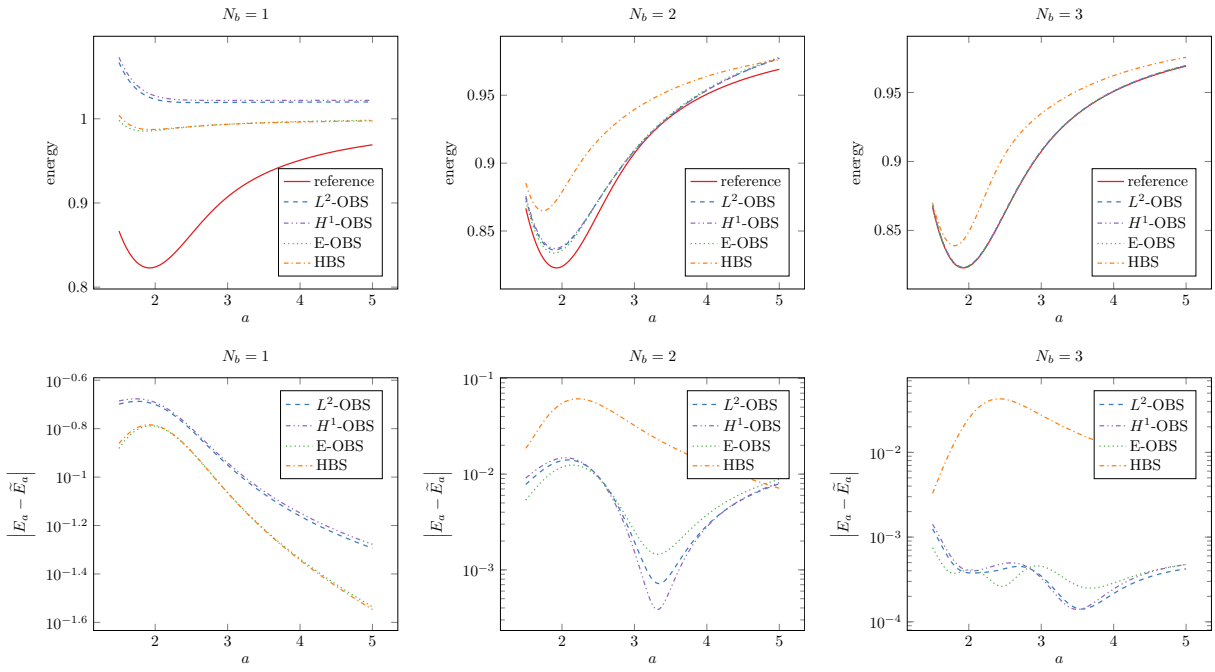



FIGURE 6.3 – Energies for the optimal bases obtained with the different criteria. (Top) Dissociation curve. (Bottom) Errors on the energy on the range of configuration $\mathcal{I} = [1.5, 5]$.

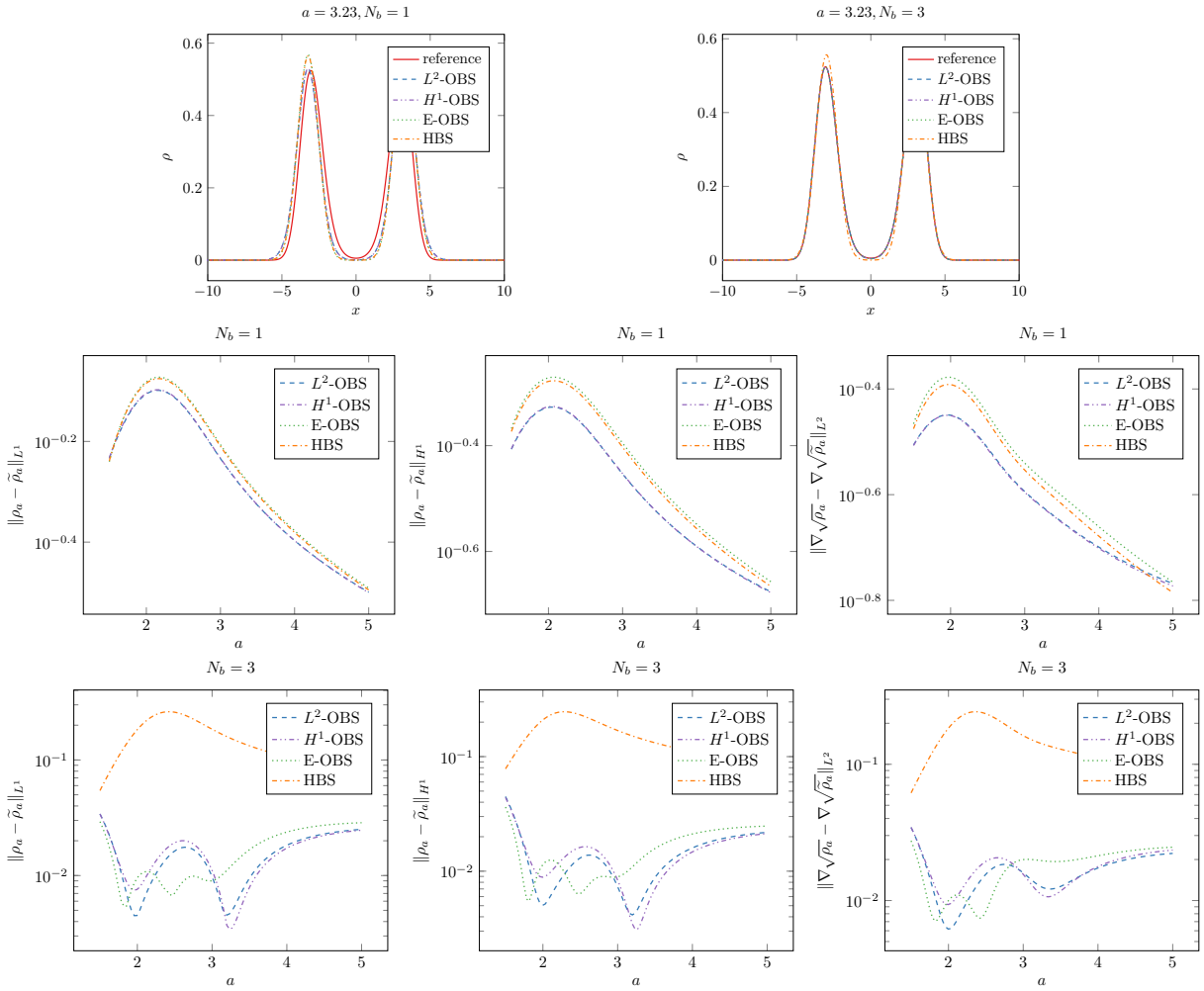



FIGURE 6.4 – (Top) Densities for the optimal bases obtained with the different criteria. (Middle) Errors on the density for different norms with $N_\mathrm{b} = 1$. (Bottom) Error on the density for different norms with $N_\mathrm{b} = 3$.

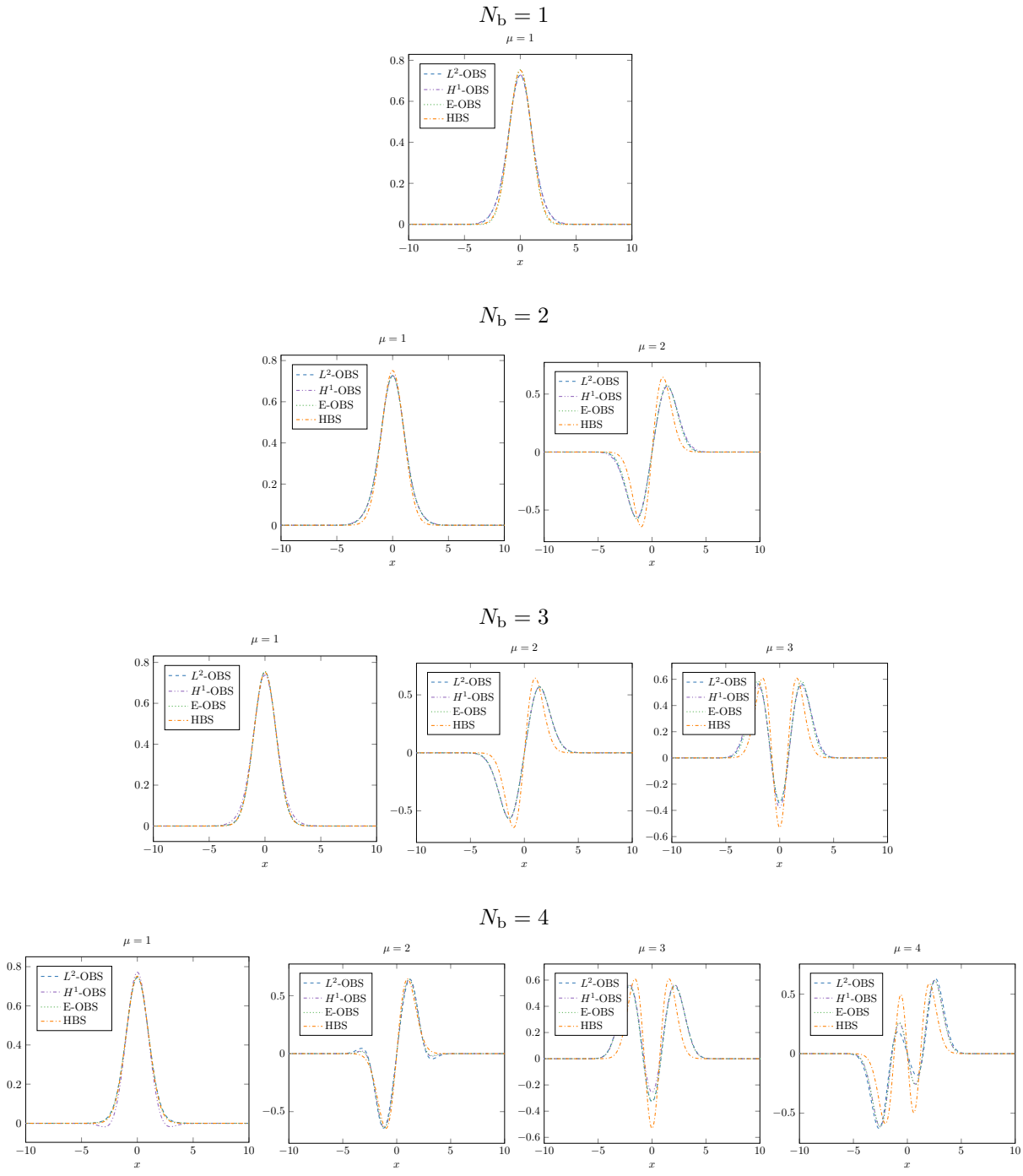



FIGURE 6.5 – Optimal basis functions for different criteria, each of them being optimized for different values of $N_\mathrm{b}$.

### 6.4.2 Influence of numerical parameters

#### Random starting points

In Section 6.4.1, we used the first $N_\mathrm{b}$ 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 6.5, but still span the same space as the variational solutions are equal.

#### Extrapolating the parameter space $\mathcal{I}$

In Section 6.4.1, 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 $N_\mathrm{b}$ increases. This phenomenon was observed already for $N_\mathrm{b} = 3$ or $N_\mathrm{b} = 4$ 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 $\chi_a$ (*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 (6.3.9) 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 6.6. 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.

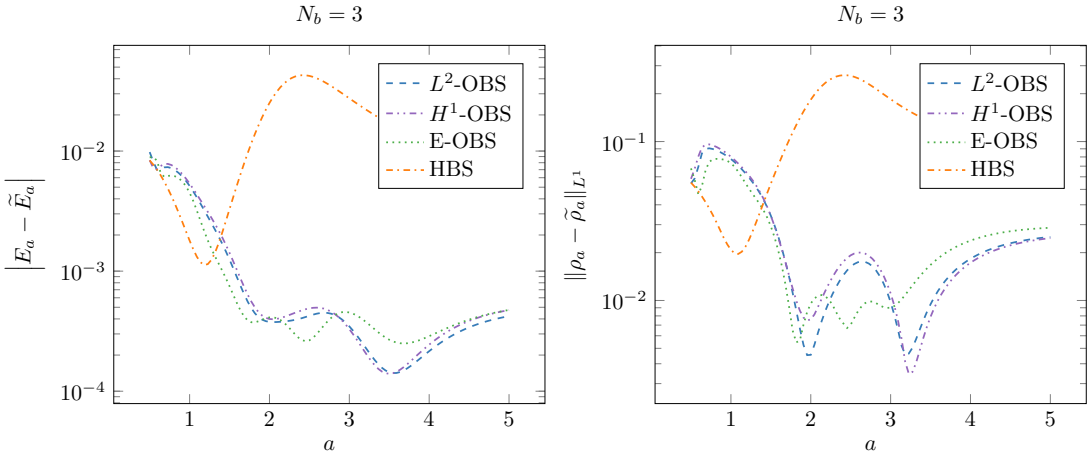



FIGURE 6.6 – Energy and densities error with extrapolation up to $a = 0.5$, with basis functions optimized on $\mathcal{I} = [1.5, 5]$

#### 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 $N_\mathrm{c} = 10$ points used in Section 6.4.1 (see Eq. (6.4.1)) 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 6.7 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 $N_\mathrm{b} = 3$.

It appears that the latter sampling already provides satisfactory accuracy. The criteria $J_{\mathrm{L}^2}$ and $J_{\mathrm{H}^1}$ are equal to $-5 \times 10^{-6}$ for optimized basis to be compared with $-1.8 \times 10^{-3}$ for standard HBS. Hence they provide a gain of accuracy in energy of three orders of magnitude over the whole dissociation curve.

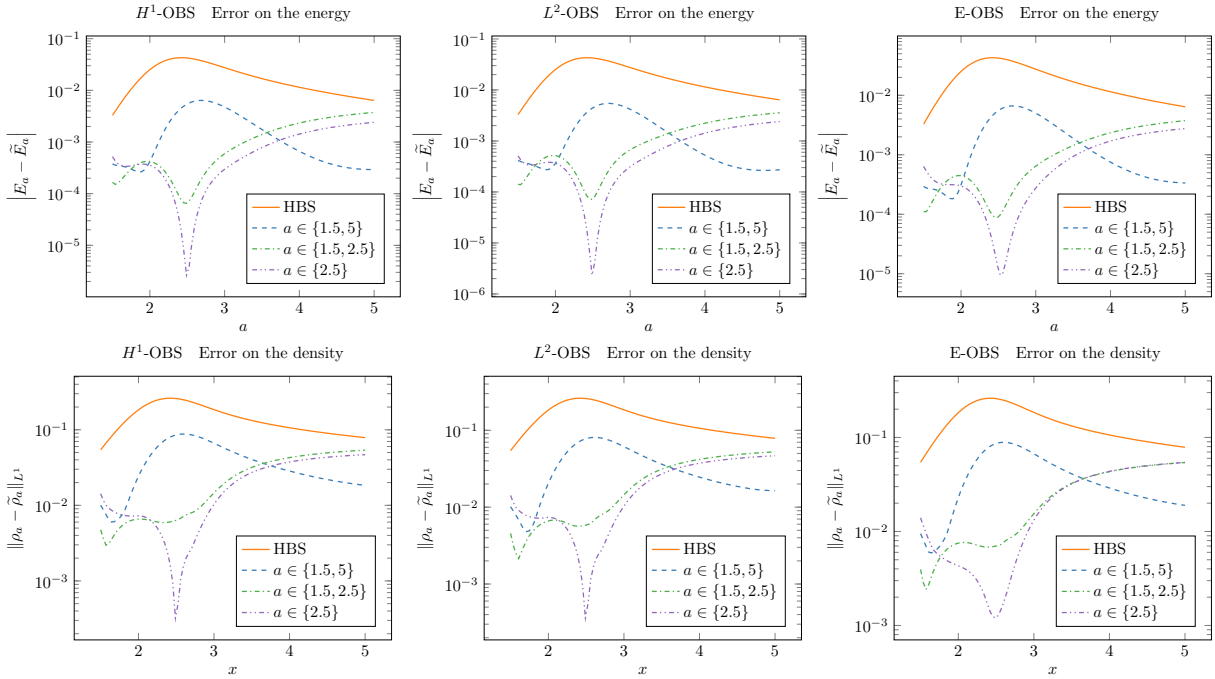



FIGURE 6.7 – Error plots for probability measures $\mathbb{P}$ corresponding to very sparse samplings of the interval $\mathcal{I} = [1.5, 5]$: i) the two endpoints of $\mathcal{I}$ ii) two points near the equilibrium distance and iii) one point near the equilibrium distance. (Top line) Error on energy. (Bottom line) Error on density in $\mathrm{L}^1$ norm. (Left) OBS for $J_{\mathrm{H}^1}$. (Middle) OBS for $J_{\mathrm{L}^2}$. (Right) OBS for $J_E$. The "$a$" in legends are the sampled configurations $a$.

### Number of Hilbert basis functions

We now take the same setting as in Section 6.4.1, except that we set $\mathcal{N} = 5$ instead of $\mathcal{N} = 10$. This provides similar results as those collected in Table 6.1, see Table 6.2. However, the values of the criteria $J_A$ and $J_E$ are higher than for $\mathcal{N} = 10$, in particular for $N_\mathrm{b} = 4$, 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 $N_\mathrm{b} = 1, 2, 3$, the criteria are slightly improved but for $N_\mathrm{b} = 4$, 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.

Value of $J_{\mathrm{L}^2}$ for the different basis sets

| Basis | $N_\mathrm{b} = 1$ | $N_\mathrm{b} = 2$ | $N_\mathrm{b} = 3$ | $N_\mathrm{b} = 4$ |
|---|---|---|---|---|
| HBS | -7.40829 | -7.70051 | -7.74312 | -7.77138 |
| $\mathrm{L}^2$-OBS | -7.43933 | -7.76304 | -7.77554 | -7.77618 |
| $\mathrm{H}^1$-OBS | -7.43923 | -7.76258 | -7.77525 | -7.77612 |
| E-OBS | -7.39401 | -7.76259 | -7.77545 | -7.77615 |

Value of $J_{\mathrm{H}^1}$ for the different basis sets

| Basis | $N_\mathrm{b} = 1$ | $N_\mathrm{b} = 2$ | $N_\mathrm{b} = 3$ | $N_\mathrm{b} = 4$ |
|---|---|---|---|---|
| HBS | -10.5613 | -11.0566 | -11.1451 | -11.2402 |
| $\mathrm{L}^2$-OBS | -10.6237 | -11.2225 | -11.2541 | -11.2577 |
| $\mathrm{H}^1$-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 | $N_\mathrm{b} = 1$ | $N_\mathrm{b} = 2$ | $N_\mathrm{b} = 3$ | $N_\mathrm{b} = 4$ |
|---|---|---|---|---|
| HBS | $3.77956\times10^{-2}$ | $3.98301\times10^{-3}$ | $1.86537\times10^{-3}$ | $1.35309\times10^{-4}$ |
| $\mathrm{L}^2$-OBS | $6.43832\times10^{-2}$ | $2.46466\times10^{-4}$ | $1.58667\times10^{-5}$ | $1.01128\times10^{-5}$ |
| $\mathrm{H}^1$-OBS | $6.13025\times10^{-2}$ | $2.45930\times10^{-4}$ | $1.62235\times10^{-5}$ | $1.00611\times10^{-5}$ |
| E-OBS | $3.69681\times10^{-2}$ | $1.30365\times10^{-4}$ | $1.41935\times10^{-5}$ | $9.74560\times10^{-6}$ |

TABLE 6.2 – Value of the different criteria for the different local (optimized and Hermite) bases, with $\mathcal{N} = 5$ and increasing values of $N_\mathrm{b}$.

# 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 $R, H \in \mathbb{R}^{\mathcal{N}\times N_\mathrm{b}}$ and define $I_H = \begin{pmatrix} H & 0 \\ 0 & H \end{pmatrix}$. One has

$$\begin{aligned} J_A(R+H) - J_A(R) = &-\int_\Omega \mathrm{Tr}\big(M_A^{\text{offline}}(a)\big(2I_R[S^A(B_aI_R)]^{-1}I_H^T + I_R\left[\mathrm{d}[S^A]^{-1}(B_aI_R)\cdot(B_aI_H)\right]I_R^T\big)\big)\,\mathrm{d}\mathbb{P}(a) \\ &+ O(\|H\|^2) \end{aligned} \tag{6.4.2}$$

Considering that

$(M+H)^{-1}-M^{-1} = -M^{-1}HM^{-1}+O(\|H\|^2)$ and $S^A(BI_{R+H})-S^A(B_aI_R) = I_H^TS^A(B)I_R+I_R^TS^A(B)I_H+O(\|H\|^2)$,

it follows from the chain rule that

$$\mathrm{d}[S^A]^{-1}(B_aI_R)\cdot(B_aI_H) = -[S^A(B_aI_R)]^{-1}\big(I_H^TS^A(B_a)I_R + I_R^TS^A(B_a)I_H\big)[S^A(B_aI_R)]^{-1}.$$

From this computation, we obtain that the integrand in expression (6.4.2) writes for all $a$

$$\begin{aligned} &2\mathrm{Tr}\big(M_A^{\text{offline}}(a)\big[I_R[S^A(B_aI_R)]^{-1}I_H^T - I_R[S^A(B_aI_R)]^{-1}I_H^TS^A(B_a)I_R[S^A(B_aI_R)]^{-1}I_R^T\big]\big) \\ &= 2\mathrm{Tr}\big(M_A^{\text{offline}}(a)I_R[S^A(B_aI_R)]^{-1}I_H^T - I_H^TS^A(B_a)I_R[S^A(B_aI_R)]^{-1}I_R^TM_A^{\text{offline}}(a)I_R[S^A(B_aI_R)]^{-1}\big). \end{aligned} \tag{6.4.3}$$

The idea is now to write the expression (6.4.3) as the inner product of $H$ with a given matrix of $\mathbb{R}^{\mathcal{N}\times N_\mathrm{b}}$, 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 (6.4.3) 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

$$M = \begin{pmatrix} M^{++} & M^{+-} \\ M^{-+} & M^{--} \end{pmatrix}. \tag{6.4.4}$$

Expression (6.4.3) therefore immediately reads

$$\begin{aligned} &2\mathrm{Tr}\Big(I_H^T\underbrace{\big[M_A^{\text{offline}}(a)I_R[S^A(B_aI_R)]^{-1} - S^A(B_a)I_R[S^A(B_aI_R)]^{-1}I_R^TM_A^{\text{offline}}(a)I_R[S^A(B_aI_R)]^{-1}\big]}_{M_A(a,R)}\Big) \\ &\qquad = 2\mathrm{Tr}\big(H^T\big(M_A(a,R)^{++} + M_A(a,R)^{--}\big)\big). \end{aligned} \tag{6.4.5}$$

One can verify that $M_A(a,R)^{++} + M_A(a,R)^{--}$ is in $\mathbb{R}^{\mathcal{N}\times N_\mathrm{b}}$ and we conclude by identification that

$$\nabla J_A(R) = -2\int_\Omega \big(M_A(a,R)^{++} + M_A(a,R)^{--}\big)\mathrm{d}\mathbb{P}(a). \tag{6.4.6}$$

### Computation of the gradient of $J_E$

Let $R, H \in \mathbb{R}^{\mathcal{N}\times N_\mathrm{b}}$ and define $I_H = \begin{pmatrix} H & 0 \\ 0 & H \end{pmatrix}$. We immediately have that

$$\nabla J_E(R) = -2\int_\Omega \nabla E_a(R)(E_a - E_a(R))\mathrm{d}\mathbb{P}(a), \tag{6.4.7}$$

where

$$E_a(R) = \mathrm{Tr}\big(C_a(R)(C_a(R))^T \times \mathcal{H}_a(R)\big), \tag{6.4.8}$$

with $C_a(R)$ defined in Section 6.3.2 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 (6.4.4), we introduce

$$M_a \coloneqq M_E^{\text{offline}}(a) \text{ and } \Sigma(H) \coloneqq I_H^T M_a I_R = \begin{pmatrix} H^T M_a^{++} R & H^T M_a^{+-} R \\ H^T M_a^{-+} R & H^T M_a^{--} R \end{pmatrix} \in \mathbb{R}^{(2N_\mathrm{b}) \times (2N_\mathrm{b})},$$

so that, with $P = CC^T$,

$$\begin{aligned} \mathrm{Tr}(P[\mathrm{d}\mathcal{H}_a(R) \cdot H]) &= \mathrm{Tr}\big(P[\Sigma(H) + \Sigma(H)^T]\big) = 2\mathrm{Tr}(P\Sigma(H)) \\ &= 2\mathrm{Tr}\big(H^T\big(M_a^{++} R P^{++} + M_a^{-+} R P^{+-} + M_a^{+-} R P^{-+} + M_a^{--} R P^{--}\big)\big). \end{aligned}$$

In the end,

$$\nabla_R \mathcal{E}_a(R, C) = 2\big(M_a^{++} R (CC^T)^{++} + M_a^{+-} R (CC^T)^{-+} + M_a^{-+} R (CC^T)^{+-} + M_a^{--} R (CC^T)^{--}\big) \in \mathbb{R}^{\mathcal{N} \times N_\mathrm{b}}.$$

**Computation of the second gradient $\nabla_C \mathcal{E}_a$** The Euler–Lagrange equation of the minimization problem (6.3.21) 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

$$\begin{aligned} \nabla_C \mathcal{E}_a(R, C_a(R)) \cdot (\mathrm{d}C_a(R) \cdot H) &= 2\mathrm{Tr}\big((S(B_a I_R) C_a(R) \Lambda_a(R))^T (\mathrm{d}C_a(R) \cdot H)\big) \\ &= -\mathrm{Tr}\big((\mathrm{d}S(B_a I_R) \cdot H) C_a(R) \Lambda_a(R) C_a(R)^T\big). \end{aligned}$$

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

$$\begin{aligned} \nabla_R E_a(R) = {} & 2\big(M_a^{++} R P_a(R)^{++} + M_a^{+-} R P_a(R)^{-+} + M_a^{-+} R P_a(R)^{+-} + M_a^{--} R P_a(R)^{--}\big) \\ & - 2\big(S_a^{++} R Q_a(R)^{++} + S_a^{+-} R Q_a(R)^{-+} + S_a^{-+} R Q_a(R)^{+-} + S_a^{--} R Q_a(R)^{--}\big) \end{aligned} \tag{6.4.9}$$

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 (6.4.7).

# Bibliography


[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

[2] J. Almlöf and P. R. Taylor. General contraction of Gaussian basis sets. I. Atomic natural orbitals for first- and second-row atoms. *Journal of Chemical Physics*, 86(7):4070–4077, 1987.

[3] F. Alouges and C. Audouze. Preconditioned gradient flows for nonlinear eigenvalue problems and application to the Hartree–Fock functional. *Numerical Methods for Partial Differential Equations. An International Journal*, 25(2):380–400, 2009.

[4] R. Altmann, D. Peterseim, and T. Stykel. Energy-adaptive Riemannian optimization on the Stiefel manifold. *ESAIM: Mathematical Modelling and Numerical Analysis*, 56(5):1629–1653, 2022.

[5] A. Anantharaman and E. Cancès. Existence of minimizers for Kohn–Sham models in quantum chemistry. *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, 26(6):2425–2455, 2009.

[6] X. Antoine, A. Levitt, and Q. Tang. Efficient spectral computation of the stationary states of rotating Bose–Einstein condensates by preconditioned nonlinear conjugate gradient methods. *Journal of Computational Physics*, 343:92–109, 2017.

[7] F. Aryasetiawan and O. Gunnarsson. The GW method. *Reports on Progress in Physics*, 61(3):237–312, 1998.

[8] I. Babuška and J. Osborn. Eigenvalue problems. In *Handbook of Numerical Analysis*, volume 2 of *Finite Element Methods (Part 1)*, pages 641–787. Elsevier, 1991.

[9] V. Bach, E. H. Lieb, M. Loss, and J. Solovej. There are no unfilled shells in unrestricted Hartree–Fock theory. *Physical Review Letters*, 72(19):2981–2983, 1994.

[10] M. Bachmayr, H. Chen, and R. Schneider. Error estimates for Hermite and even-tempered Gaussian approximations in quantum chemistry. *Numerische Mathematik*, 128(1):137–165, 2014.

[11] G. Bacskay. A quadratically convergent Hartree–Fock (QC-SCF) method. Application to closed shell systems. *Chemical Physics*, 61(3):385–404, 1981.

[12] W. Bao and Y. Cai. Mathematical theory and numerical methods for Bose–Einstein condensation. *Kinetic and Related Models*, 6:1, 2013.

[13] W. Bao and Q. Du. Computing the Ground State Solution of Bose–Einstein Condensates by a Normalized Gradient Flow. *SIAM Journal on Scientific Computing*, 25(5):1674–1697, 2004.

[14] S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi. Phonons and related crystal properties from density-functional perturbation theory. *Reviews of Modern Physics*, 73(2):515–562, 2001.

[15] S. Baroni, P. Giannozzi, and A. Testa. Green's-function approach to linear response in solids. *Physical Review Letters*, 58(18):1861–1864, 1987.

[16] N. W. Bazley and D. W. Fox. Lower Bounds for Eigenvalues of Schrödinger's Equation. *Physical Review*, 124(2):483–492, 1961.

[17] R. Benguria, H. Brezis, and E. H. Lieb. The Thomas–Fermi–von Weizsäcker theory of atoms and molecules. *Communications in Mathematical Physics*, 79(2):167–180, 1981.

[18] S. Bernstein. Sur la nature analytique des solutions des équations aux dérivées partielles du second ordre. *Mathematische Annalen*, 59(1-2):20–76, 1904.

[19] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

[20] 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. *Journal of the American Chemical Society*, 1980.

[21] S. Blatt. On the analyticity of solutions to non-linear elliptic partial differential systems. *arXiv:2009.08762 [math.AP]*, 2020.

[22] S. M. Blinder. Eigenvalues for a Pure Quartic Oscillator. *arXiv:1903.07471 [quant-ph]*, 2019.

[23] P. E. Blöchl. Projector augmented-wave method. *Physical Review B*, 50(24):17953–17979, 1994.

[24] D. Boffi, R. G. Durán, F. Gardini, and L. Gastaldi. A posteriori error analysis for nonconforming approximation of multiple eigenvalues. *Mathematical Methods in the Applied Sciences*, 40(2):350–369, 2017.

[25] S. F. Boys and A. C. Egerton. Electronic wave functions - I. A general method of calculation for the stationary states of any molecular system. *Proceedings of the Royal Society of London A*, 200(1063):542–554, 1950.

[26] D. Braess, V. Pillwein, and J. Schöberl. Equilibrated residual error estimates are p-robust. *Computer Methods in Applied Mechanics and Engineering*, 198(13-14):1189–1197, 2009.

[27] K. Burke, J. Werschnik, and E. K. U. Gross. Time-dependent density functional theory: Past, present, and future. *Journal of Chemical Physics*, 123(6):062206, 2005.

[28] E. Cancès. Self-consistent field algorithms for Kohn–Sham models with fractional occupation numbers. *Journal of Chemical Physics*, 114(24):10616–10622, 2001.

[29] E. Cancès. Introduction to First-Principle Simulation of Molecular Systems. In M. Mateos and P. Alonso, editors, *Computational Mathematics, Numerical Analysis and Applications*, volume 13, pages 61–106. Springer International Publishing, Cham, 2017.

[30] E. Cancès, R. Chakir, and Y. Maday. Numerical Analysis of Nonlinear Eigenvalue Problems. *Journal of Scientific Computing*, 45(1):90–117, 2010.

[31] E. Cancès, R. Chakir, and Y. Maday. Numerical analysis of the planewave discretization of some orbital-free and Kohn–Sham models. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46(2):341–388, 2012.

[32] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday. Computational quantum chemistry: A primer. In *Handbook of Numerical Analysis*, volume 10 of *Special Volume, Computational Chemistry*, pages 3–270. Elsevier, 2003.

[33] E. Cancès, A. Deleurence, and M. Lewin. A New Approach to the Modeling of Local Defects in Crystals: The Reduced Hartree–Fock Case. *Communications in Mathematical Physics*, 281(1):129–177, 2008.

[34] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík. A perturbation-method-based a posteriori estimator for the planewave discretization of nonlinear Schrödinger equations. *Comptes Rendus Mathematique*, 352(11):941–946, 2014.

[35] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík. Guaranteed and Robust a Posteriori Bounds for Laplace Eigenvalues and Eigenvectors: Conforming Approximations. *SIAM Journal on Numerical Analysis*, 55(5):2228–2254, 2017.

[36] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík. Guaranteed and robust a posteriori bounds for Laplace eigenvalues and eigenvectors: A unified framework. *Numerische Mathematik*, 140(4):1033–1079, 2018.

[37] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík. Guaranteed a posteriori bounds for eigenvalues and eigenvectors: Multiplicities and clusters. *Mathematics of Computation*, 89(326):2563–2611, 2020.

[38] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík. Post-processing of the planewave approximation of Schrödinger equations. Part I: Linear operators. *IMA Journal of Numerical Analysis*, 41(4):2423–2455, 2021.

[39] E. Cancès, V. Ehrlacher, D. Gontier, A. Levitt, and D. Lombardi. Numerical quadrature in the Brillouin zone for periodic Schrödinger operators. *Numerische Mathematik*, 144(3):479–526, 2020.

[40] E. Cancès and C. Le Bris. Can we outperform the DIIS approach for electronic structure calculations? *International Journal of Quantum Chemistry*, 79(2):82–90, 2000.

[41] E. Cancès and C. Le Bris. On the convergence of SCF algorithms for the Hartree–Fock equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 34(4):749–774, 2000.

[42] E. Cancès, C. Le Bris, and Y. Maday. *Méthodes Mathématiques En Chimie Quantique. Une Introduction*, volume 53 of *Mathématiques & Applications*. Springer Berlin Heidelberg, 2006.

[43] E. Cancès, A. Levitt, Y. Maday, and C. Yang. Numerical methods for Kohn–Sham models: Discretization, algorithms, and error analysis. In E. Cancès and G. Friesecke, editors, *Density Functional Theory*, chapter 7. Springer, 2021.

[44] E. Cancès and M. Lewin. The Dielectric Permittivity of Crystals in the Reduced Hartree–Fock Approximation. *Archive for Rational Mechanics and Analysis*, 197(1):139–177, 2010.

[45] E. Cancès and N. Mourad. A mathematical perspective on density functional perturbation theory. *Nonlinearity*, 27(9):1999–2033, 2014.

[46] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral Methods: Fundamentals in Single Domains*. Springer Science & Business Media, 2007.

[47] C. Carstensen and J. Gedicke. Guaranteed lower bounds for eigenvalues. *Mathematics of Computation*, 83(290):2605–2629, 2014.

[48] I. Catto, C. Le Bris, and P. L. Lions. *Mathematical Theory of Thermodynamic Limits : Thomas–Fermi Type Models*. Oxford University Press, 1998.

[49] G. Chaban, M. Schmidt, and M. Gordon. Approximate second order method for orbital optimization of SCF and MCSCF wavefunctions. *Theoretical Chemistry Accounts*, 97(1):88–95, 1997.

[50] I. Charpentier, F. De Vuyst, and Y. Maday. A component mode synthesis method of infinite order of accuracy using subdomain overlapping: Numerical analysis and experiments. *Publication du laboratoire d'Analyse Numerique*, 96002:55–65, 1996.

[51] I. Charpentier, F. De Vuyst, and Y. Maday. Méthode de synthèse modale avec une décomposition de domaine par recouvrement. *Comptes rendus de l'Académie des sciences. Série 1, Mathématique*, 322(9):881–888, 1996.

[52] H. Chen, X. Dai, X. Gong, L. He, and A. Zhou. Adaptive Finite Element Approximations for Kohn–Sham Models. *Multiscale Modeling & Simulation*, 12(4):1828–1869, 2014.

[53] H. Chen, L. He, and A. Zhou. Finite element approximations of nonlinear eigenvalue problems in quantum physics. *Computer Methods in Applied Mechanics and Engineering*, 200(21-22):1846–1865, 2011.

[54] M. Chupin, M.-S. Dupuy, G. Legendre, and E. Séré. Convergence analysis of adaptive DIIS algorithms with application to electronic ground state calculations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 55(6):2785–2825, 2021.

[55] L. E. Daga, B. Civalleri, and L. Maschio. Gaussian Basis Sets for Crystalline Solids: All-Purpose Basis Set Libraries vs System-Specific Optimizations. *Journal of Chemical Theory and Computation*, 16(4):2192–2201, 2020.

[56] X. Dai, S. de Gironcoli, B. Yang, and A. Zhou. Mathematical Analysis and Numerical Approximations of Density Functional Theory Models for Metallic Systems. *arXiv:2201.07035 [math.NA]*, 2022.

[57] X. Dai, Z. Liu, L. Zhang, and A. Zhou. A Conjugate Gradient Method for Electronic Structure Calculations. *SIAM Journal on Scientific Computing*, 39(6):A2702–A2740, 2017.

[58] X. Dai, Y. Pan, B. Yang, and A. Zhou. Convergence and Optimal Complexity of the Adaptive Planewave Method for Eigenvalue Computations. *arXiv:2106.01008 [math.NA]*, 2021.

[59] X. Dai, J. Xu, and A. Zhou. Convergence and optimal complexity of adaptive finite element eigenvalue computations. *Numerische Mathematik*, 110(3):313–355, 2008.

[60] I. Danaila and B. Protas. Computation of Ground States of the Gross–Pitaevskii Functional via Riemannian Optimization. *SIAM Journal on Scientific Computing*, 39(6):B1102–B1129, 2017.

[61] E. B. Davies. *Spectral Theory and Differential Operators*. Cambridge University Press, first edition, 1995.

[62] J. H. de Boer and E. J. W. Verwey. Semi-conductors with partially and with completely filled 3d-lattice bands. *Proceedings of the Physical Society*, 49(4S):59–71, 1937.

[63] P. Dederichs and R. Zeller. Self-consistency iterations in electronic-structure calculations. *Physical Review B*, 28(10):5462, 1983.

[64] P. Destuynder and B. Métivet. Explicit error bounds in a conforming finite element method. *Mathematics of Computation*, 68(228):1379–1396, 1999.

[65] P. A. M. Dirac. Quantum Mechanics of Many-Electron Systems. *Proceedings of the Royal Society of London A*, 123(792):714–733, 1929.

[66] T. H. Dunning. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *Journal of Chemical Physics*, 90(2):1007–1023, 1989.

[67] M.-S. Dupuy. *Analysis of the Projector Augmented-Wave Method for Electronic Structure Calculations in Periodic Settings*. Thèse de doctorat, Sorbonne Paris Cité, 2018.

[68] R. G. Durán, C. Padra, and R. Rodríguez. A Posteriori Error Estimates for the Finite Element Approximation of Eigenvalue Problems. *Mathematical Models and Methods in Applied Sciences*, 13(08):1219–1229, 2003.

[69] G. Dusson. Post-processing of the plane-wave approximation of Schrödinger equations. Part II: Kohn–Sham models. *IMA Journal of Numerical Analysis*, 41(4):2456–2487, 2021.

[70] G. Dusson and Y. Maday. A posteriori analysis of a nonlinear Gross–Pitaevskii-type eigenvalue problem. *IMA Journal of Numerical Analysis*, 37(1):94–137, 2017.

[71] G. Dusson, I. Sigal, and B. Stamm. Analysis of the Feshbach-Schur method for the planewave discretizations of Schrödinger operators. *arXiv:2008.10871 [cs, math]*, 2020.

[72] A. Edelman, T. A. Arias, and S. T. Smith. The Geometry of Algorithms with Orthogonality Constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[73] A. Ern and M. Vohralík. Polynomial-Degree-Robust A Posteriori Estimates in a Unified Setting for Conforming, Nonconforming, Discontinuous Galerkin, and Mixed Discretizations. *SIAM Journal on Numerical Analysis*, 53(2):1058–1081, 2015.

[74] G. E. Forsythe. Asymptotic lower bounds for the frequencies of certain polygonal membranes. *Pacific Journal of Mathematics*, 4(3):467–480, 1954.

[75] C. Freysoldt, S. Boeck, and J. Neugebauer. Direct minimization technique for metals in density functional theory. *Physical Review B*, 79(24):241103, 2009.

[76] A. Friedman. On the Regularity of the Solutions of Non-Linear Elliptic and Parabolic Systems of Partial Differential Equations. *Indiana University Mathematics Journal*, 7(1):43–59, 1958.

[77] S. Giani and E. J. C. Hall. An a posteriori error estimator for hp-adaptive discontinuous Galerkin methods for elliptic eigenvalue problems. *Mathematical Models and Methods in Applied Sciences*, 22(10):1250030, 2012.

[78] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch. QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials. *Journal of Physics. Condensed Matter: An Institute of Physics Journal*, 21(39):395502, 2009.

[79] S. Goedecker, M. Teter, and J. Hutter. Separable dual-space Gaussian pseudopotentials. *Physical Review B*, 54(3):1703, 1996.

[80] D. Gontier and S. Lahbabi. Convergence rates of supercell calculations in the reduced Hartree–Fock model. *ESAIM: Mathematical Modelling and Numerical Analysis*, 50(5):1403–1424, 2016.

[81] X. Gonze. Adiabatic density-functional perturbation theory. *Physical Review A*, 52(2):1096–1114, 1995.

[82] X. Gonze. Perturbation expansion of variational principles at arbitrary order. *Physical Review A*, 52(2):1086–1095, 1995.

[83] X. Gonze. Towards a potential-based conjugate gradient algorithm for order-N self-consistent total energy calculations. *Physical Review B*, 54(7):4383, 1996.

[84] X. Gonze, B. Amadon, G. Antonius, F. Arnardi, L. Baguet, J.-M. Beuken, J. Bieder, F. Bottin, J. Bouchet, E. Bousquet, N. Brouwer, F. Bruneval, G. Brunin, T. Cavignac, J.-B. Charraud, W. Chen, M. Côté, S. Cottenier, J. Denier, G. Geneste, P. Ghosez, M. Giantomassi, Y. Gillet, O. Gingras, D. R. Hamann, G. Hautier, X. He, N. Helbig, N. Holzwarth, Y. Jia, F. Jollet, W. Lafargue-Dit-Hauret, K. Lejaeghere, M. A. L. Marques, A. Martin, C. Martins, H. P. C. Miranda, F. Naccarato, K. Persson, G. Petretto, V. Planes, Y. Pouillon, S. Prokhorenko, F. Ricci, G.-M. Rignanese, A. H. Romero, M. M. Schmitt, M. Torrent, M. J. van Setten, B. Van Troeye, M. J. Verstraete, G. Zérah, and J. W. Zwanziger. The Abinit project: Impact, environment and recent developments. *Computer Physics Communications*, 248:107042, 2020.

[85] X. Gonze and J.-P. Vigneron. Density-functional approach to nonlinear-response coefficients of solids. *Physical Review B*, 39(18):13120–13128, 1989.

[86] A. Griewank and A. Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, Second Edition.* Society for Industrial and Applied Mathematics, second edition, 2008.

[87] D. R. Hamann, M. Schlüter, and C. Chiang. Norm-Conserving Pseudopotentials. *Physical Review Letters*, 43(20):1494–1497, 1979.

[88] B. L. Hammond, W. A. Lester, and P. J. Reynolds. *Monte Carlo Methods in* Ab Initio *Quantum Chemistry*, volume 1 of *World Scientific Lecture and Course Notes in Chemistry.* World Scientific, 1994.

[89] D. R. Hartree. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(1):89–110, 1928.

[90] D. R. Hartree. The wave mechanics of an atom with a non-Coulomb central field. Part II. Some results and discussion. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(1):111–132, 1928.

[91] C. Hartwigsen, S. Goedecker, and J. Hutter. Relativistic separable dual-space Gaussian pseudopotentials from H to Rn. *Physical Review B*, 58(7):3641–3662, 1998.

[92] Y. Hashimoto. A Remark on the Analyticity of the Solutions for Non-Linear Elliptic Partial Differential Equations. *Tokyo Journal of Mathematics*, 29(2):271–281, 2006.

[93] W. J. Hehre, R. F. Stewart, and J. A. Pople. Self-Consistent Molecular-Orbital methods. I. Use of gaussian expansions of Slater-Type atomic orbitals. *Journal of Chemical Physics*, 51(6):2657–2664, 1969.

[94] P. Heid, B. Stamm, and T. P. Wihler. Gradient Flow Finite Element Discretizations with Energy-Based Adaptivity for the Gross–Pitaevskii Equation. *arXiv:1906.06954 [cs, math]*, 2019.

[95] T. Helgaker, P. Jørgensen, and J. Olsen. *Molecular Electronic-Structure Theory*. John Wiley & Sons, Ltd, Chichester, UK, 2000.

[96] H. Hellmann. *Einführung in die Quantenchemie*. J.W. Edwards, 1944.

[97] P. Henning and D. Peterseim. Sobolev Gradient Flow for the Gross–Pitaevskii Eigenvalue Problem: Global Convergence and Computational Efficiency. *SIAM Journal on Numerical Analysis*, 58(3):1744–1772, 2020.

[98] M. F. Herbst and A. Levitt. Black-box inhomogeneous preconditioning for self-consistent field iterations in density functional theory. *Journal of Physics: Condensed Matter*, 33(8):085503, 2020.

[99] M. F. Herbst and A. Levitt. A robust and efficient line search for self-consistent field iterations. *Journal of Computational Physics*, 459(C):111–127, 2022.

[100] M. F. Herbst, A. Levitt, and E. Cancès. A posteriori error estimation for the non-self-consistent Kohn–Sham equations. *Faraday Discussions*, 224:227–246, 2020.

[101] M. F. Herbst, A. Levitt, and E. Cancès. DFTK: A Julian approach for simulating electrons in solids. *Proceedings of the JuliaCon Conferences*, 3(26):69, 2021.

[102] N. J. Higham. *Functions of Matrices*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 2008.

[103] A. Hjorth Larsen, J. Jørgen Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. Bjerre Jensen, J. Kermode, J. R. Kitchin, E. Leonhard Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. Bergmann Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017.

[104] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Physical Review*, 136(3B):B864–B871, 1964.

[105] R. Holmes. A formula for the spectral radius of an operator. *The American Mathematical Monthly*, 75(2):163–166, 1968.

[106] J. Hu, Y. Huang, and Q. Lin. Lower Bounds for Eigenvalues of Elliptic Operators: By Nonconforming Finite Element Methods. *Journal of Scientific Computing*, 61(1):196–221, 2014.

[107] J. Hu, Y. Huang, and Q. Shen. The Lower/Upper Bound Property of Approximate Eigenvalues by Nonconforming Finite Element Methods for Elliptic Operators. *Journal of Scientific Computing*, 58(3):574–591, 2014.

[108] D. Johnson. Modified Broyden's method for accelerating convergence in self-consistent calculations. *Physical Review B*, 38(18):12807–12813, 1988.

[109] M. F. Kasim and S. M. Vinko. Learning the Exchange-Correlation Functional from Nature with Fully Differentiable Density Functional Theory. *Physical Review Letters*, 127(12):126403, 2021.

[110] T. Kato. On the Upper and Lower Bounds of Eigenvalues. *Journal of the Physical Society of Japan*, 4(4-6):334–339, 1949.

[111] T. Kato. *Perturbation Theory for Linear Operators*. Classics in Mathematics. Springer-Verlag, Berlin Heidelberg, second edition, 1995.

[112] G. P. Kerker. Efficient iteration scheme for self-consistent pseudopotential calculations. *Physical Review B*, 23(6):3082–3084, 1981.

[113] J. Kirkpatrick, B. McMorrow, D. H. P. Turban, A. L. Gaunt, J. S. Spencer, A. G. D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, L. R. Castellanos, S. Petersen, A. W. R. Nelson, P. Kohli, P. Mori-Sánchez, D. Hassabis, and A. J. Cohen. Pushing the frontiers of density functionals by solving the fractional electron problem. *Science*, 374(6573):1385–1389, 2021.

[114] L. Kleinman and D. M. Bylander. Efficacious Form for Model Pseudopotentials. *Physical Review Letters*, 48(20):1425–1428, 1982.

[115] A. V. Knyazev. Toward the Optimal Preconditioned Eigensolver: Locally Optimal Block Preconditioned Conjugate Gradient Method. *SIAM Journal on Scientific Computing*, 23(2):517–541, 2001.

[116] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, 1965.

[117] G. Kresse. VASP - Vienna Ab initio Simulation Package.

[118] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 54(16):11169–11186, 1996.

[119] K. N. Kudin, G. E. Scuseria, and E. Cancès. A black-box self-consistent field convergence algorithm: One step closer. *Journal of Chemical Physics*, 116(19):8255–8261, 2002.

[120] W. Kutzelnigg. Theory of the expansion of wave functions in a Gaussian basis. *International Journal of Quantum Chemistry*, 1994.

[121] P. Ladeveze and D. Leguillon. Error Estimate Procedure in the Finite Element Method and Applications. *SIAM Journal on Numerical Analysis*, 20(3):485–509, 1983.

[122] M. G. Larson. A Posteriori and a Priori Error Analysis for Finite Element Approximations of Self-Adjoint Elliptic Eigenvalue Problems. *SIAM Journal on Numerical Analysis*, 38(2):608–625, 2000.

[123] S. Lehtola. Curing basis set overcompleteness with pivoted Cholesky decompositions. *The Journal of Chemical Physics*, 151(24):241102, 2019.

[124] A. Levitt. Convergence of gradient-based algorithms for the Hartree–Fock equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46(6):1321–1336, 2012.

[125] A. Levitt. Screening in the Finite-Temperature Reduced Hartree–Fock Model. *Archive for Rational Mechanics and Analysis*, 238(2):901–927, 2020.

[126] M. Levy. Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the $v$-representability problem. *Proceedings of the National Academy of Sciences*, 76(12):6062–6065, 1979.

[127] M. Levy. Electron densities in search of Hamiltonians. *Physical Review A*, 26(3):1200–1208, 1982.

[128] M. Lewin. *Théorie Spectrale Et Mécanique Quantique*. Springer, 2022.

[129] L. Li, S. Hoyer, R. Pederson, R. Sun, E. D. Cubuk, P. Riley, and K. Burke. Kohn–Sham Equations as Regularizer: Building Prior Knowledge into Machine-Learned Physics. *Physical Review Letters*, 126(3):036401, 2021.

[130] E. H. Lieb. Thomas–Fermi and related theories of atoms and molecules. *Reviews of Modern Physics*, 53(4):603–641, 1981.

[131] E. H. Lieb. Density functionals for Coulomb systems. *International Journal of Quantum Chemistry*, 24(3):243–277, 1983.

[132] E. H. Lieb and B. Simon. The Hartree–Fock theory for Coulomb systems. *Communications In Mathematical Physics*, 53(3):185–194, 1977.

[133] L. Lin and J. Lu. *A Mathematical Introduction to Electronic Structure Theory*. SIAM Spotlights. Society for Industrial and Applied Mathematics, 2019.

[134] L. Lin, J. Lu, and L. Ying. Numerical methods for Kohn–Sham density functional theory. *Acta Numerica*, 28:405–539, 2019.

[135] P. L. Lions. Solutions of Hartree–Fock equations for Coulomb systems. *Communications in Mathematical Physics*, 109(1):33–97, 1987.

[136] B. Liu, H. Chen, G. Dusson, J. Fang, and X. Gao. An Adaptive Planewave Method for Electronic Structure Calculations. *Multiscale Modeling & Simulation*, 20(1):524–550, 2022.

[137] X. Liu. A framework of verified eigenvalue bounds for self-adjoint differential operators. *Applied Mathematics and Computation*, 267:341–355, 2015.

[138] X. Liu, X. Wang, Z. Wen, and Y. Yuan. On the Convergence of the Self-Consistent Field Iteration in Kohn–Sham Density Functional Theory. *SIAM Journal on Matrix Analysis and Applications*, 35(2):546–558, 2014.

[139] X. Liu, Z. Wen, X. Wang, M. Ulbrich, and Y. Yuan. On the Analysis of the Discretized Kohn–Sham Density Functional Theory. *SIAM Journal on Numerical Analysis*, 53(4):1758–1785, 2015.

[140] P.-O. Löwdin. On the nonorthogonality problem. In P.-O. Löwdin, editor, *Advances in Quantum Chemistry*, volume 5, pages 185–199. Academic Press, 1970.

[141] F. Luo, Q. Lin, and H. Xie. Computing the lower and upper bounds of Laplace eigenvalue problem: By combining conforming and nonconforming finite element methods. *Science China Mathematics*, 55(5):1069–1082, 2012.

[142] D. Mao, L. Shen, and A. Zhou. Adaptive finite element algorithms for eigenvalue problems based on local averaging type a posteriori error estimates. *Advances in Computational Mathematics*, 25(1-3):135–160, 2006.

[143] L. Marks and D. Luke. Robust mixing for ab initio quantum mechanical calculations. *Physical Review B*, 78(7):075114, 2008.

[144] R. M. Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, first edition, 2004.

[145] N. Marzari, D. Vanderbilt, and M. C. Payne. Ensemble density-functional theory for ab initio molecular dynamics of metals and finite-temperature insulators. *Physical review letters*, 79(7):1337, 1997.

[146] R. McWeeny. The density matrix in self-consistent field theory. I. Iterative construction of the density matrix. *Proceedings of the Royal Society of London A*, 235:496–509, 1956.

[147] Q. Mérigot, F. Santambrogio, and C. Sarrazin. Non-asymptotic convergence bounds for Wasserstein approximation using point clouds. *Advances in Neural Information Processing Systems*, 34:12810–12821, 2021.

[148] J. W. Milnor. *Topology from the Differentiable Viewpoint*. 1997.

[149] P. K. Mogensen and A. N. Riseth. Optim: A mathematical optimization package for Julia. *Journal of Open Source Software*, 3(24):615, 2018.

[150] H. J. Monkhorst and J. D. Pack. Special points for Brillouin-zone integrations. *Physical Review B*, 13(12):5188–5192, 1976.

[151] A. Mostofi, P. Haynes, C.-K. Skylaris, and M. Payne. Preconditioned iterative minimization for linear-scaling electronic structure calculations. *Journal of Chemical Physics*, 119(17):8842–8848, 2003.

[152] N. F. Mott and R. Peierls. Discussion of the paper by de Boer and Verwey. *Proceedings of the Physical Society*, 49(4S):72–73, 1937.

[153] M. T. Nakao, M. Plum, and Y. Watanabe. *Numerical Verification Methods and Computer-Assisted Proofs for Partial Differential Equations*. Number 53 in Springer Series in Computational Mathematics. Springer, Singapore, 2019.

[154] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2006.

[155] P. Norman, K. Ruud, and T. Saue. *Principles and Practices of Molecular Properties: Theory, Modeling and Simulations*. John Wiley & Sons, Ltd, Chichester, UK, 2018.

[156] J. Olsen. An introduction and overview of basis sets for molecular and Solid-State calculations. In E. Perlt, editor, *Basis Sets in Computational Chemistry*, pages 1–16. Springer International Publishing, Cham, 2021.

[157] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

[158] G. Pagès. Introduction to vector quantization and its applications for numerics. *ESAIM: Proceedings and Surveys*, 48:29–79, 2015.

[159] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos. Iterative minimization techniques for ab initio total-energy calculations: Molecular dynamics and conjugate gradients. *Reviews of Modern Physics*, 64(4):1045–1097, 1992.

[160] J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized Gradient Approximation Made Simple. *Physical Review Letters*, 77(18):3865–3868, 1996.

[161] E. Perlt, editor. *Basis Sets in Computational Chemistry*, volume 107 of *Lecture Notes in Chemistry*. Springer International Publishing, Cham, 2021.

[162] K. Persson. Materials data on TiO2 (SG:136) by materials project, 2014.

[163] I. G. Petrovskii. Sur l'analyticité des solutions des systèmes d'équations différentielles. *Matematiceskij sbornik*, 47(1):3–70, 1939.

[164] D. H. Pham. *Galerkin Method Using Optimized Wavelet-Gaussian Mixed Bases for Electronic Structure Calculations in Quantum Chemistry*. Thèse de doctorat, Université Grenoble Alpes, 2017.

[165] L. P. Pitaevskii and S. Stringari. *Bose–Einstein Condensation*. International Series of Monographs on Physics. Oxford University Press, Oxford, New York, 2003.

[166] W. Prager and J. L. Synge. Approximations in elasticity based on the concept of function space. *Quarterly of Applied Mathematics*, 5(3):241–269, 1947.

[167] P. Pulay. Convergence acceleration of iterative sequences. the case of SCF iteration. *Chemical Physics Letters*, 73(2):393–398, 1980.

[168] P. Pulay. Improved SCF convergence acceleration. *Journal of Computational Chemistry*, 3(4):556–560, 1982.

[169] D. Raczkowski, A. Canning, and L. W. Wang. Thomas–Fermi charge mixing for obtaining self-consistency in density functional calculations. *Physical Review B*, 64(12):121101, 2001.

[170] L. E. Ratcliff, W. Dawson, G. Fisicaro, D. Caliste, S. Mohr, A. Degomme, B. Videau, V. Cristiglio, M. Stella, M. D'Alessandro, S. Goedecker, T. Nakajima, T. Deutsch, and L. Genovese. Flexibilities of wavelets as a computational basis set for large-scale electronic structure calculations. *Journal of Chemical Physics*, 152(19):194110, 2020.

[171] M. Reed and B. Simon. *Analysis of Operators*. Number 4 in Methods of Modern Mathematical Physics. Academic Press, 1978.

[172] M. Reed and B. Simon. *Functional Analysis*. Number 1 in Methods of Modern Mathematical Physics. Academic Press, 1980.

[173] J. Revels, M. Lubin, and T. Papamarkou. Forward-Mode Automatic Differentiation in Julia. *arXiv:1607.07892 [cs.MS]*, 2016.

[174] T. Rohwedder and R. Schneider. An analysis for the DIIS acceleration method used in quantum chemistry calculations. *Journal of Mathematical Chemistry*, 49(9):1889, 2011.

[175] A. H. Romero, D. C. Allan, B. Amadon, G. Antonius, T. Applencourt, L. Baguet, J. Bieder, F. Bottin, J. Bouchet, E. Bousquet, F. Bruneval, G. Brunin, D. Caliste, M. Côté, J. Denier, C. Dreyer, P. Ghosez, M. Giantomassi, Y. Gillet, O. Gingras, D. R. Hamann, G. Hautier, F. Jollet, G. Jomard, A. Martin, H. P. C. Miranda, F. Naccarato, G. Petretto, N. A. Pike, V. Planes, S. Prokhorenko, T. Rangel, F. Ricci, G.-M. Rignanese, M. Royo, M. Stengel, M. Torrent, M. J. van Setten, B. Van Troeye, M. J. Verstraete, J. Wiktor, J. W. Zwanziger, and X. Gonze. ABINIT: Overview and focus on selected capabilities. *Journal of Chemical Physics*, 152(12):124102, 2020.

[176] C. Roothaan. New developments in molecular orbital theory. *Reviews of Modern Physics*, 23(2):69–89, 1951.

[177] E. Rudberg. Difficulties in applying pure Kohn–Sham density functional theory electronic structure methods to protein molecules. *Journal of Physics: Condensed Matter*, 24(7):072202, 2012.

[178] Y. Saad. *Numerical Methods for Large Eigenvalue Problems: Revised Edition.* Society for Industrial and Applied Mathematics, 2011.

[179] Y. Saad, J. R. Chelikowsky, and S. M. Shontz. Numerical Methods for Electronic Structure Calculations of Materials. *SIAM Review*, 52(1):3–54, 2010.

[180] D. Sánchez-Portal, E. Artacho, and J. M. Soler. Analysis of atomic orbital basis sets from the projection of plane-wave results. *Journal of Physics: Condensed Matter*, 8(21):3859–3880, 1996.

[181] A. Schmidt, D. Wittwar, and B. Haasdonk. Rigorous and effective a-posteriori error bounds for nonlinear problems—application to RB methods. *Advances in Computational Mathematics*, 46(2):32, 2020.

[182] S. Scholz and H. Yserentant. On the approximation of electronic wavefunctions by anisotropic Gauss and Gauss–Hermite functions. *Numerische Mathematik*, 136(3):841–874, 2017.

[183] R. A. Shaw. The completeness properties of Gaussian-type orbitals in quantum chemistry. *International Journal of Quantum Chemistry*, 120(17):93, 2020.

[184] J. R. Shewchuk. An Introduction to the Conjugate Gradient Method Without the Agonizing Pain. 1994.

[185] J. C. Slater. Atomic Shielding Constants. *Physical Review*, 36(1):57–64, 1930.

[186] J. C. Slater. A Simplification of the Hartree–Fock Method. *Physical Review*, 81(3):385–390, 1951.

[187] G. Srivastava. Broyden's method for self-consistent field convergence acceleration. *Journal of Physics A*, 17(6):L317–L321, 1984.

[188] R. M. Sternheimer. Electronic Polarizabilities of Ions from the Hartree–Fock Wave Functions. *Physical Review*, 96(4):951–968, 1954.

[189] R. F. Stewart. Small gaussian expansions of atomic orbitals. *Journal of Chemical Physics*, 50(6):2485–2495, 1969.

[190] S. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering.* CRC press, 2001.

[191] W. C. Topp and J. J. Hopfield. Chemically Motivated Pseudopotential for Sodium. *Physical Review B*, 7(4):1295–1303, 1973.

[192] J. Toulouse. Review of approximations for the exchange-correlation energy in density-functional theory. *arXiv:2103.02645 [cond-mat, physics:physics]*, 2021.

[193] N. Troullier and J. L. Martins. Efficient pseudopotentials for plane-wave calculations. *Physical Review B*, 43(3):1993–2006, 1991.

[194] P. Upadhyaya, E. Jarlebring, and E. H. Rubensson. A density matrix approach to the convergence of the self-consistent field iteration. *Numerical Algebra, Control & Optimization*, 11(1):99, 2021.

[195] R. Van Noorden, B. Maher, and R. Nuzzo. The top 100 papers. *Nature News*, 514(7524):550, 2014.

[196] E. Vecharynski, C. Yang, and J. E. Pask. A projected preconditioned conjugate gradient algorithm for computing a large invariant subspace of a Hermitian matrix. *Journal of Computational Physics*, 290:73–89, 2015.

[197] R. Verfürth. A posteriori error estimation and adaptive mesh-refinement techniques. *Journal of Computational and Applied Mathematics*, 50(1-3):67–83, 1994.

[198] R. Vrabel, P. Tanuska, P. Vazan, P. Schreiber, and V. Liska. Duffing-Type Oscillator with a Bounded from above Potential in the Presence of Saddle-Center Bifurcation and Singular Perturbation: Frequency Control. *Abstract and Applied Analysis*, 2013:1–7, 2013.

[199] W. Walter. *Ordinary Differential Equations*. Number 182 in Graduate Texts in Mathematics ; Readings in Mathematics. Springer, New York, 1998.

[200] T. Wazewski. Systèmes des équations et des inégualités différentielles ordinaires aux deuxièmes membres monotones et leurs applications. *Annales de la Société polonaise de mathématique*, 23:112–166, 1950.

[201] H. F. Weinberger. Upper and lower bounds for eigenvalues by finite difference methods. *Communications on Pure and Applied Mathematics*, 9(3):613–623, 1956.

[202] N. Woods. *On the Nature of Self-Consistency in Density Functional Theory*. Master thesis, University of Cambridge, 2018.

[203] N. Woods, M. Payne, and P. Hasnip. Computing the self-consistent field in Kohn–Sham density functional theory. *Journal of Physics: Condensed Matter*, 31(45):453001, 2019.

[204] N. Yan and A. Zhou. Gradient recovery type a posteriori error estimates for finite element approximations on irregular meshes. *Computer Methods in Applied Mechanics and Engineering*, 190(32-33):4289–4299, 2001.

[205] B. Yang and A. Zhou. Eigenfunction behavior and adaptive finite element approximations of nonlinear eigenvalue problems in quantum physics. *ESAIM: Mathematical Modelling and Numerical Analysis*, 55(1):209–227, 2021.

[206] C. Yang, W. Gao, and J. C. Meza. On the Convergence of the Self-Consistent Field Iteration for a Class of Nonlinear Eigenvalue Problems. *SIAM Journal on Matrix Analysis and Applications*, 30:1773–1788, 2008.

[207] X. Zhang, J. Zhu, Z. Wen, and A. Zhou. Gradient type optimization methods for electronic structure calculations. *SIAM Journal on Scientific Computing*, 36:265–289, 2014.

[208] Z. Zhang. Exponential convergence of Sobolev gradient descent for a class of nonlinear eigenproblems. *arXiv:1912.02135 [math.NA]*, 2019.

[209] Z. Zhao, Z. Bai, and X. Jin. A Riemannian Newton algorithm for nonlinear eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 36:752–774, 2015.

[210] A. Zhou. Finite dimensional approximations for the electronic ground state solution of a molecular system. *Mathematical Methods in the Applied Sciences*, 30(4):429–447, 2007.