Error estimation for linear and nonlinear eigenvalue problems arising from electronic structure calculation

Geneviève Dusson

▶ **To cite this version:**

Geneviève Dusson. Error estimation for linear and nonlinear eigenvalue problems arising from electronic structure calculation. Mathematical Physics [math-ph]. Université Pierre et Marie Curie - Paris VI, 2017. English. NNT : 2017PA066238 . tel-01689793

HAL Id: tel-01689793
https://theses.hal.science/tel-01689793

Submitted on 22 Jan 2018

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.


Université Pierre et Marie Curie - Paris VI

Thèse de doctorat

Discipline: Mathématiques Appliquées

présentée par

Geneviève Dusson

Estimation d'erreur pour des problèmes aux valeurs propres linéaires et non-linéaires issus du calcul de structure électronique

Error estimation for linear and nonlinear eigenvalue problems arising from electronic structure calculation

---

dirigée par Yvon Maday et Jean-Philip Piquemal

Soutenue publiquement le 23 octobre 2017 devant le jury composé de :

M.	Eric Cancès	Examinateur
M.	Thierry Deutsch	Examinateur
Mme	Maria Esteban	Examinatrice
M.	Pascal Frey	Examinateur
M.	Lin Lin	Rapporteur
M.	Yvon Maday	Directeur de thèse
M.	Jean-Philip Piquemal	Co-directeur de thèse
M.	Reinhold Schneider	Rapporteur

Remerciements

J'ai eu la chance de rencontrer Yvon Maday (presque) par hasard alors que je n'étais pas encore étudiante en master. Il m'a donné goût à la recherche en y accompagnant mes premiers pas avec un enthousiasme communicatif. Faire ma thèse sous sa direction a été une expérience extrêmement enrichissante, et je le remercie profondément.

Cette thèse a été particulièrement passionnante grâce à la collaboration étroite de Eric Cancès, Benjamin Stamm, et Martin Vohralík, à qui ce travail doit beaucoup. Eric s'est impliqué sans compter ni son temps ni ses conseils, toujours très avisés, et je lui adresse mes sincères remerciements. Un grand merci à Benjamin pour tous les conseils de code, la transformation du café en tea-orems ainsi que les déjeuners maths-chimie. Martin, mille mercis pour les séances de travail à l'Inria, pour ta patience et tes encouragements.

Je souhaite ensuite remercier Jean-Philip Piquemal, qui m'a permis de mettre un pied du côté de la chimie, et m'a aidée à comprendre les enjeux de ma thèse dans ce domaine.

Je remercie vivement Pascal Frey pour la confiance qu'il m'a accordée et son accueil chaleureux à l'ISCD. Je suis ravie qu'il ait accepté de faire partie de mon jury.

I would like to thank the members of my PhD committee, starting with Lin Lin and Reinhold Schneider, who accepted to review my work. Je suis également très reconnaissante à Thierry Deutsch et Maria Esteban d'avoir accepté de faire partie de mon jury de thèse.

Au cours de cette thèse, j'ai eu la chance de beaucoup voyager, et j'aimerais remercier les personnes rencontrées au cours de cette aventure. I would like to sincerely thank Gero Friesecke for hosting me in Munich and for a fruitful collaboration, which is hopefully only beginning. Many thanks to Huajie Chen for the great invitation in Beijing, and also to Xiaoying Dai for the kind invitation at the Chinese Academy of Sciences. J'aimerais aussi remercier Ionut Danaila pour son invitation à l'institut Fields, et Guillaume Vergez pour les discussions sur les condensats de Bose–Einstein. Merci à Antoine Levitt pour les temps partagés en France et à l'étranger.

J'aimerais ensuite remercier Andreas Savin pour de nombreuses conversations intéressantes, ainsi que Filippo Lipparini et Elisa Rebolini qui m'ont introduit au monde de la chimie quantique avec beaucoup de patience et de bienveillance. Merci également aux participants des déjeuners maths-chimie, en particulier Carlo Marcati, Chaoyu Quan et Etienne Polack. Merci aussi à Athmane Bakhta; l'organisation du GTT maths-chimie a été un plaisir.

Le laboratoire Jacques-Louis Lions constitue un environnement formidable pour préparer une thèse. J'en profite pour remercier l'équipe du secrétariat, Catherine, Malika et Salima, toujours très efficaces et sympathiques, ainsi que l'équipe informatique, et notamment Kashayar et Stephan pour leur aide précieuse.

Merci à tous les doctorant-e-s (passés et présents) qui font du laboratoire Jacques-Louis Lions un lieu vivant et sympathique. Merci en particulier à toutes celles et ceux qui ont partagé mon bureau et qui ont contribué à la vie étudiante du labo, pour les déjeuners et les soirées partagées. Merci également aux doctorants de l'ISCD.

Enfin, merci infiniment à mes parents, mes frères et sœur, et Antoine, pour leur soutien indéfectible.

Résumé

L'objectif de cette thèse est de fournir des bornes d'erreur pour des problèmes aux valeurs propres linéaires et non linéaires issus du calcul de structure électronique. Nous nous intéressons en particulier au calcul de l'état fondamental basé sur la théorie de la fonctionnelle de la densité, comprenant le modèle de Kohn–Sham. Nos estimations reposent principalement sur des analyses d'erreur *a posteriori*. Plus précisément, nous commençons par étudier un phénomène de compensation d'erreur de discrétisation pour un problème simple, linéaire aux valeurs propres, pour lequel les solutions analytiques sont disponibles. L'étude mathématique est basée sur une analyse *a priori* de l'erreur sur l'énergie. Ensuite, nous présentons une analyse *a posteriori* pour le problème du laplacien aux valeurs propres discrétisé avec des éléments finis conformes et non conformes. Nous fournissons des bornes d'erreur garanties, calculables et efficaces, pour les valeurs propres simples de l'opérateur de Laplace et leurs vecteurs propres associés. Par la suite, nous nous concentrons sur des problèmes aux valeurs propres non linéaires. Tout d'abord, nous proposons une analyse *a posteriori* pour l'équation de Gross-Pitaevskii. Les bornes d'erreur obtenues sont valables sous des hypothèses qui peuvent être vérifiées numériquement. Elles peuvent être séparées en deux composantes venant respectivement de la discrétisation et de l'algorithme itératif utilisé pour résoudre le problème non linéaire aux valeurs propres. L'équilibrage de ces composantes d'erreur permet d'optimiser les ressources numériques. Deuxièmement, nous présentons une méthode de post-traitement pour le problème de Kohn–Sham, qui améliore la précision des orbitales de l'état fondamental, et qui a un faible coût de calcul pour des simulations en ondes planes. Les solutions post-traitées peuvent être utilisées soit comme solutions plus précises du problème, soit pour calculer une estimation de l'erreur de discrétisation, qui n'est alors pas garantie, mais qui est en pratique proche de l'erreur réelle.

Abstract

The objective of this thesis is to provide error bounds for linear and nonlinear eigenvalue problems arising from electronic structure calculation. We focus on ground-state calculations based on Density Functional Theory, including Kohn–Sham models. Our estimations mostly rely on *a posteriori* error analysis. More precisely, we start by studying a phenomenon of discretization error cancellation for a simple linear eigenvalue problem, for which analytical solutions are available. The mathematical study is based on an *a priori* analysis for the energy error. Then, we present an *a posteriori* analysis for the Laplace eigenvalue problem discretized with conforming and nonconforming finite elements. We provide guaranteed, fully computable and efficient error bounds, for simple eigenvalues of the Laplace operator and their corresponding eigenvectors. Thereafter, we focus on nonlinear eigenvalue problems. First, we provide an *a posteriori* analysis for the Gross–Pitaevskii equation. The error bounds are valid under assumptions that can be numerically checked, and can be separated in two components coming respectively from the discretization and the iterative algorithm used to solve the nonlinear eigenvalue problem. Balancing these error components allows to optimize the computational resources. Second, we present a post-processing method for the Kohn–Sham problem, which improves the accuracy of planewave computations of ground state orbitals at a low computational cost. The post-processed solutions can be used either as a more precise solution of the problem, or used for computing an estimation of the discretization error, which is not guaranteed, but in practice close to the real error.

Préambule

À l'échelle atomique, un système moléculaire composé de M noyaux et de N électrons peut être très bien décrit par sa fonction d'onde, une fonction à valeurs complexes dépendant de $4(M+N)$ variables, trois variables d'espace et une variable de spin par particule (et de $3(M+N)$ variables lorsque le spin est négligé). Connaissant la fonction d'onde, il est possible de calculer de nombreuses propriétés physiques et chimiques du système, par exemple l'énergie de dissociation, la conductivité ou la conformation spatiale. Le calcul précis et efficace de ces propriétés est un défi important dans de nombreux domaines tels que la chimie, la physique de la matière condensée ou la science des matériaux, surtout lorsque les expériences sont très coûteuses ou impossibles à réaliser.

Pour un système donné, la fonction d'onde est solution d'une équation de Schrödinger dépendante du temps, une équation aux dérivées partielles linéaire posée sur un espace de dimension $3^{(N+M)}$. Malgré le caractère linéaire de l'équation, la grande dimension de l'espace sur lequel l'équation est posée empêche de résoudre cette équation numériquement. En effet, la résolution de cette équation est trop coûteuse lorsque le nombre de particules dépasse un ou deux, même avec les ordinateurs et les supercalculateurs d'aujourd'hui.

Un problème plus simple mais toujours représentatif dans ce domaine est le calcul de l'état fondamental électronique du système, qui est l'état stationnaire de plus basse énergie, pour une configuration donnée des noyaux. Ce problème prend la forme d'une équation de Schrödinger indépendante du temps, qui est encore impossible à résoudre en pratique même pour les systèmes de petite taille. Dans ce cadre, l'affirmation suivante annoncée par Dirac en 1929 [90] est encore pertinente aujourd'hui:

Les lois physiques sous-jacentes nécessaires à la théorie mathématique d'une grande partie de la physique et de l'ensemble de la chimie sont donc complètement connues, et la difficulté est que l'application exacte de ces lois conduit à des équations trop compliquées pour être solubles. Il est donc souhaitable que des méthodes d'approximation pratiques de l'application de la mécanique quantique soient développées, ce qui peut conduire à expliquer les principales caractéristiques des systèmes atomiques complexes sans trop de calcul.

Pour résoudre le problème linéaire très complexe de détermination de l'état fondamental électronique du système, différentes approximations sont utilisées. Tout d'abord, le modèle est simplifié, afin d'obtenir des équations résolubles numériquement. Cela se fait généralement par une réduction drastique de la dimension du modèle, souvent contrebalancée par l'introduction d'une non-linéarité. Deuxièmement, les équations du modèle simplifié sont discrétisées, réduisant ainsi le problème à un problème en dimension finie. Troisièmement, les équations discrétisées sont résolues, éventuellement en utilisant des algorithmes itératifs avec un nombre fini d'itérations. Une question naturelle qui se pose à ce stade est la mesure de ces approximations et l'impact qu'elles ont sur le résultat final.

Pour répondre à cette question, il faut quantifier les erreurs introduites à chaque approximation. De telles estimations permettent d'obtenir une certification des simulations numériques, et donc des résultats fiables. Mais ils servent également à optimiser les paramètres utilisés pour exécuter les simulations. En effet, en quantifiant les erreurs provenant des différentes sources d'approximation (simplifications du modèle, discrétisation, algorithmes), il est possible d'évaluer les approximations ayant un plus grand impact sur le résultat et donc d'affiner uniquement les paramètres appropriés. Ainsi, équilibrer les composantes d'erreur permet d'optimiser le rapport entre le coût et la précision des calculs.

Cette thèse se concentre sur l'estimation d'erreur pour des problèmes provenant du calcul de structure électronique. Les problèmes considérés sont des problèmes aux valeurs propres linéaires et non linéaires. L'objectif principal est de fournir des bornes d'erreur calculables et

garanties pour l'erreur de discrétisation de différents problèmes, ainsi qu'une séparation et un équilibrage de différentes composantes d'erreur. Par exemple, nous présentons une séparation des erreurs de discrétisation et d'algorithme pour un problème aux valeurs propres non linéaire. Nous avons également développé une méthode de post-traitement améliorant la précision des simulations numériques à un faible coût de calcul.

Cette thèse est composée de neuf chapitres, regroupés en cinq parties.

La première partie, composée de deux chapitres, est une partie introductive. Le chapitre 1 donne un aperçu des différents modèles utilisés en chimie quantique pour le calcul de structure électronique. Différentes méthodes de discrétisation ainsi que des algorithmes itératifs pour résoudre de tels problèmes sont également présentés. Ces approximations successives constituent les sources d'erreurs que nous souhaiterions contrôler. Le chapitre 2 fournit une introduction à l'estimation d'erreur, à la fois *a priori* et *a posteriori*, ainsi qu'une description des différentes contributions de cette thèse. Pour chaque contribution, les principaux résultats sont présentés.

Les quatre autres parties présentent les différentes contributions de cette thèse et peuvent être lues indépendamment. L'ordre des parties suit la complexité croissante des modèles considérés.

Dans la partie II (chapitre 3), le modèle considéré est unidimensionnel. Nous étudions une équation linéaire aux valeurs propres découlant d'un problème de minimisation, pour lequel les solutions analytiques sont disponibles. L'équation est paramétrée par une variable spatiale, modélisant la position des noyaux dans une molécule. Nous présentons une estimation d'erreur *a priori* pour l'énergie, à partir de laquelle nous expliquons un phénomène de compensation d'erreur de discrétisation entre différents systèmes, observé à travers des simulations numériques pour deux systèmes moléculaires simples. Il est montré que l'amélioration de convergence menant à la compensation d'erreur réside dans le préfacteur et non dans le taux de convergence de l'erreur.

La partie III se concentre sur le problème du laplacien aux valeurs propres. Dans le chapitre 4, une estimation d'erreur *a posteriori* pour des méthodes d'éléments finis conformes est présentée. Les bornes d'erreur sont complètement calculables et valides dans des conditions vérifiables *a posteriori*. L'analyse est effectuée non seulement pour la valeur propre la plus basse du laplacien, mais aussi pour d'autres valeurs propres simples. En outre, l'erreur due au solveur algébrique inexact est prise en compte, ce qui rend possible un équilibrage d'erreur entre l'erreur de discrétisation et l'erreur algébrique. Dans le chapitre 5, une extension de cette analyse au cas des éléments finis non conformes est présentée.

La partie IV (chapitre 6) expose une estimation d'erreur *a posteriori* pour un problème non linéaire aux valeurs propres : l'équation de Gross-Pitaevskii, qui est mathématiquement proche du modèle de Thomas–Fermi. L'équation, posée dans un cadre périodique unidimensionnel, est discrétisée en ondes planes (séries de Fourier) et résolue à l'aide d'un algorithme de point fixe (appelé SCF pour self-consistent field). Une première borne d'erreur grossière mais garantie est fournie, ce qui permet de justifier de manière garantie l'utilisation d'une deuxième borne *a posteriori* plus précise. L'estimateur d'erreur est ensuite décomposé en deux parties : l'une d'elles dépend principalement de la discrétisation, l'autre dépend du nombre d'itérations effectuées dans l'algorithme de point fixe. Cela permet de raffiner de manière adaptative la dimension de l'espace discret et le nombre d'itérations dans l'algorithme afin d'obtenir un bon compromis entre précision et coût de calcul.

La partie V présente une méthode de post-traitement basée sur la théorie des perturbations de Rayleigh-Schrödinger, pour les problèmes de valeurs propres linéaires et non linéaires. L'idée clé est d'effectuer un calcul complet dans une grille grossière, puis un calcul peu coûteux

dans une grille plus fine, de sorte que la précision de la solution post-traitée est comparable à la précision de la solution sur la grille fine. La méthode est présentée pour la première fois pour le modèle Kohn–Sham dans le chapitre 7, avec des simulations numériques montrant l'amélioration de la précision sur l'énergie apportée par le post-traitement. Un taux de convergence théorique amélioré pour l'énergie est également annoncé. Les démonstrations de cette estimation sur l'énergie sont ensuite présentées dans le cas des opérateurs de Schrödinger linéaires dans le cadre des matrices densité dans le chapitre 8. La réduction de l'erreur sur les matrices densités est elle aussi estimée. Enfin, les démonstrations sont étendues au cas non linéaire, comprenant le modèle Kohn–Sham, dans le chapitre 9.

Preamble

At the atomic scale, a molecular system composed of M nuclei and N electrons can be very well described by its wave-function, a complex-valued function depending on $4(M+N)$ variables, three space variables and one spin variable per particle (and $3(M+N)$ variables when the spin is neglected). From the knowledge of the wave-function, many physical and chemical properties of the system can be computed, e.g. dissociation energies, conductivity, or spatial conformations. The accurate and efficient computation of these properties is an important challenge in many fields such as chemistry, condensed matter physics, or materials science, especially when experiments are highly expensive or practically impossible.

For a given system, the wave-function is solution to a time-dependent many-body Schrödinger equation, which is a linear partial differential equation posed on a $3^{(N+M)}$-dimensional space. Unlike its linear property, the huge space on which the equation is posed prevents from solving this equation numerically. Indeed, the resolution of this equation is way too costly when the number of particles exceeds one or two, even with today's computers and supercomputers.

A simpler but still representative problem in this domain is the computation of the electronic ground state of the system, which is the steady state of lowest energy, for a given configuration of the nuclei. This problem takes the form of a time-independent N-body Schrödinger equation, which is still impossible to solve in practice even for small size systems. In this framework, the following assertion announced by Dirac in 1929 [90] is still relevant today:

> *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.*

To handle this very complex linear problem of finding the electronic ground-state of the system, different approximations are used. First, the model is simplified, in order to obtain tractable equations. This is usually done by a drastic dimension reduction of the model, often counterbalanced by the introduction of a nonlinearity. Second, the equations of the simplified models are discretized, reducing the problem to a finite-dimensional one. Third, the discretized equations are solved, possibly using iterative algorithms with a finite number of iterations. A natural question arising at this point is how large these approximations are and what impact they have on the final result.

To answer this question, the errors introduced at each approximation step need to be quantified. Such estimations allow to get certified, and hence reliable results in the numerical simulations. But they also serve to optimize the parameters used to run the simulations. Indeed, by quantifying the errors coming from the different sources (model simplification, discretization, algorithms), one can evaluate which approximations lead to the largest errors, and hence refine only the appropriate parameters. Thus, balancing the error components allows to optimize the ratio between the computational cost and the accuracy of the computations.

This thesis focuses on the error analysis for problems arising from electronic structure calculation. The problems under consideration are linear and nonlinear eigenvalue problems. The main goal is to provide computable and guaranteed error bounds for the discretization error of different problems, as well as a separation and balancing of different error components. For example, we provide a separation of the discretization and algorithm errors for a nonlinear eigenvalue problem. We also developed a post-processing method improving the accuracy of numerical simulations at a low computational cost.

This thesis is composed of nine chapters, grouped into five parts.

The first part, which is composed of two chapters, is an introductory part. Chapter 1 gives an overview of different models used in quantum chemistry for the calculation of electronic structures. Different discretization methods as well as iterative algorithms for solving such problems are also presented. These successive approximations constitute the sources of errors we aim at controlling. Chapter 2 provides an introduction to error analysis, both *a priori* and *a posteriori* together with a description of the different contributions of this thesis. For each contribution, the main results are presented.

The four other parts present the different contributions and can be read independently. The order of the parts follows the increasing complexity of the models under consideration.

In Part II (Chapter 3), the model under consideration is one-dimensional. We study a linear eigenvalue equation arising from a minimization problem, for which analytic solutions are available. The equation is parametrized by a spatial parameter, modeling the position of the nuclei in a molecule. We provide an *a priori* error estimation for the energy, from which we explain a phenomenon of discretization error cancellation between different systems, observed through numerical simulations for two simple molecular systems. This assesses that the convergence improvement leading to the error cancellation lies in the prefactor and not in the convergence rate of the error.

Part III focuses on the Laplace eigenvalue problem. In Chapter 4, an *a posteriori* error estimation for conforming finite element methods is presented. The error bounds are fully computable, and valid under *a posteriori* verifiable conditions. The analysis is performed not only for the lowest eigenvalue of the Laplace operator, but also for other simple eigenvalues. Moreover, inexact algebraic solvers are taken into account, leading to a possible error balance between discretization error and algebraic error. In Chapter 5, an extension to the case of nonconforming finite element methods is presented.

Part IV (Chapter 6) provides an *a posteriori* error estimation for a nonlinear eigenvalue problem: the Gross–Pitaevskii equation, which is mathematically close to the Thomas–Fermi model. The equation posed in a one-dimensional periodic setting is discretized using planewaves (Fourier series) and solved using a self-consistent field (fixed point) algorithm. A first coarse but guaranteed bound is provided, which allows to justify in a guaranteed manner the use of a second, more precise *a posteriori* bound. The error estimator is then decomposed into two parts: one of them depends mainly on the discretization, the other one on the number of iterations performed in the fixed point algorithm. This allows to refine adaptively the dimension of the discrete space and the number of iterations in the algorithm to get a good compromise between accuracy and computational cost.

Part V presents a post-processing method based on Rayleigh–Schrödinger perturbation theory, for linear and nonlinear eigenvalue problems. The key idea is to perform a full computation in a coarse grid, and then a cheap computation in a finer grid, so that the accuracy of the post-processed solution is comparable to the accuracy of the solution on the fine grid. The method is first presented for the Kohn–Sham model in Chapter 7, with numerical simulations showing the improvement of the energy accuracy brought by the post-processing. A theoretical improved convergence rate for the energy is also announced. The proofs for this estimate are then presented in the linear case of Schrödinger operators in the framework of density matrices in Chapter 8, together with the improvement on the density matrix. Finally, the proofs are extended to the nonlinear case, including the Kohn–Sham model in Chapter 9.

List of publications

Here is a list of articles (accepted, submitted or in preparation) that were written during this thesis:

- [49] Eric Cancès, Geneviève Dusson, Yvon Maday, Benjamin Stamm, Martin Vohralík, *A perturbation-method-based a posteriori estimator for the planewave discretization of nonlinear Schrödinger equations*, Comptes Rendus Mathematique, 352 (2014), pp. 941–946.
- [50] Eric Cancès, Geneviève Dusson, Yvon Maday, Benjamin Stamm, Martin Vohralík, *A perturbation-method-based post-processing for the planewave discretization of Kohn–Sham models*, Journal of Computational Physics 307 (2016) 446–459.
- [98] Geneviève Dusson, Yvon Maday, *A Posteriori Analysis of a Non-Linear Gross–Pitaevskii type Eigenvalue Problem*, IMA Journal of Numerical Analysis (2016), drw001.
- [47] *Discretization error cancellation in electronic structure calculation: toward a quantitative study*, ESAIM: Mathematical Modelling and Numerical Analysis, 51 (2017), pp. 1617–1636.
- [52] Eric Cancès, Geneviève Dusson, Yvon Maday, Benjamin Stamm, Martin Vohralík, *Guaranteed and robust a posteriori bounds for Laplace eigenvalues and eigenvectors: conforming approximations*, SIAM Journal on Numerical Analysis, 55 (2017), pp. 2228–2254.
- [51] Eric Cancès, Geneviève Dusson, Yvon Maday, Benjamin Stamm, Martin Vohralík, *Guaranteed and robust a posteriori bounds for Laplace eigenvalues and eigenvectors: a unified framework,* submitted.
- [48] Eric Cancès, Geneviève Dusson, Yvon Maday, Benjamin Stamm, Martin Vohralík, *Post-processing of the planewave approximation of Schrödinger equations. Part I: linear operators,* in preparation.
- [97] Geneviève Dusson, *Post-processing of the planewave approximation of Schrödinger equations. Part II: Kohn–Sham models,* in preparation.

Contents

I Introduction **19**

1 Electronic structure calculation **21**
1.1 Introduction . . . 21
1.2 The Schrödinger model . . . 22
1.2.1 Description of a molecular system . . . 22
1.2.2 Electronic time-independent Schrödinger equation . . . 23
1.2.3 Limitations of the Schrödinger model . . . 24
1.3 Approximations for the electronic Schrödinger equation . . . 25
1.3.1 Hartree–Fock and post Hartree–Fock methods . . . 25
1.3.2 Multi-configuration methods . . . 28
1.3.3 Density Functional Theory . . . 28
1.3.4 Quantum Monte Carlo . . . 32
1.4 Other standard approximations . . . 33
1.4.1 Supercell model . . . 33
1.4.2 Pseudopotentials . . . 34
1.5 Discretization methods . . . 35
1.5.1 Different codes for different discretizations . . . 35
1.5.2 Molecular orbitals . . . 35
1.5.3 Planewaves . . . 38
1.5.4 Wavelets . . . 39
1.5.5 Finite elements . . . 40
1.5.6 Finite differences . . . 41
1.6 Algorithms for mean-field models . . . 41
1.6.1 Self-Consistent Field algorithms . . . 41
1.6.2 Direct minimization methods . . . 44

2 Error analysis for linear and nonlinear eigenvalue problems **45**
2.1 Introduction . . . 45
2.1.1 Importance of error control . . . 45
2.1.2 An error arising from different sources . . . 46
2.1.3 *A priori* analysis for eigenvalue problems . . . 48
2.1.4 *A posteriori* analysis for eigenvalue problems . . . 49
2.2 Error cancellation for the discretization error . . . 51
2.3 A posteriori error estimation for the Laplace eigenvalue problem . . . 53
2.3.1 Presentation of the problem . . . 53
2.3.2 Results in the conforming case . . . 54
2.3.3 Results in the nonconforming case . . . 58
2.4 A posteriori error estimation for the Gross–Pitaevskii equation . . . 58
2.5 Post-processing for the Kohn–Sham model . . . 62

2.5.1 Perturbation theory . . . 62
2.5.2 Application to linear Schrödinger operators . . . 63
2.5.3 Application to the Kohn–Sham model . . . 66
2.6 A posteriori estimation and post-processing methods in a unified framework . . 69
2.6.1 Generic linear and nonlinear eigenvalue problems . . . 69
2.6.2 Some convergence results . . . 70
2.6.3 A posteriori estimation for the Gross–Pitaevskii equation . . . 71
2.6.4 A posteriori estimation for the Hartree–Fock problem . . . 72
2.6.5 Two-grid methods . . . 73
2.6.6 Post-processing methods based on perturbation theory . . . 74
2.7 Conclusion and perspectives . . . 77

II Discretization error cancellation for a linear eigenvalue problem 79

3 Discretization error cancellation 81
3.1 Introduction . . . 82
3.2 Discretization error cancellation in planewave calculations . . . 84
3.2.1 Ground state potential energy surface of the H_2 molecule . . . 85
3.2.2 Energy of a simple chemical reaction . . . 86
3.3 Mathematical analysis of a toy model . . . 87
3.4 Appendix: proof of Theorem 3.3.2 . . . 93

III A posteriori error estimation for the Laplace eigenvalue problem 101

4 A posteriori error estimation for the Laplace eigenvalue problem: conforming case 103
4.1 Introduction . . . 104
4.2 Setting . . . 106
4.2.1 The Laplace eigenvalue problem . . . 106
4.2.2 Residual and its dual norm . . . 106
4.3 Generic equivalences . . . 107
4.3.1 i-th eigenvalue error equivalences . . . 109
4.3.2 i-th eigenvector error equivalences . . . 110
4.4 Dual norm of the residual equivalences . . . 111
4.4.1 Meshes and discrete spaces . . . 111
4.4.2 Equilibrated flux reconstruction for inexact solvers . . . 112
4.4.3 Conforming local residual liftings . . . 112
4.4.4 Dual norm of the residual equivalences . . . 113
4.5 Guaranteed and fully computable upper and lower bounds . . . 114
4.5.1 Eigenvalues . . . 114
4.5.2 Eigenvectors . . . 117
4.5.3 Comments . . . 117
4.6 Application to conforming finite elements . . . 118
4.7 Numerical experiments . . . 118
4.7.1 First eigenvalue on the unit square . . . 119
4.7.2 First eigenvalue on an L-shaped domain: mesh adaptivity . . . 120
4.7.3 First eigenvalue on a domain with a hole: mesh adaptivity . . . 121
4.7.4 Higher eigenvalues . . . 121

4.7.5 Inexact algebraic eigenvalue solvers . . . 122
4.7.6 Comparison with existing results . . . 123

5 A posteriori error estimation for the Laplace eigenvalue problem: nonconforming case 129
5.1 Introduction . . . 130
5.2 Setting . . . 132
5.2.1 The Laplace eigenvalue problem . . . 132
5.2.2 Meshes and generic piecewise polynomial spaces . . . 133
5.2.3 Broken and discrete gradients . . . 133
5.2.4 Residual and its dual norm . . . 134
5.3 Eigenvector and equilibrated flux reconstructions . . . 134
5.3.1 Orthogonality of the residual . . . 135
5.3.2 Reconstruction spaces . . . 135
5.3.3 Equilibrated flux reconstruction . . . 135
5.3.4 Eigenvector reconstruction . . . 136
5.4 Dual norm of the residual and nonconformity bounds . . . 136
5.4.1 Some additional notation and useful inequalities . . . 136
5.4.2 Stability of the equilibrated flux and eigenvector reconstructions . . . 137
5.4.3 Dual norm of the residual and nonconformity bounds . . . 137
5.5 Elliptic regularity bounds on the Riesz representation of the residual . . . 139
5.6 Guaranteed and computable upper and lower bounds in a unified framework . . . 141
5.6.1 Eigenvalues . . . 141
5.6.2 Eigenvectors . . . 144
5.6.3 Comments . . . 147
5.7 Application to common nonconforming numerical methods . . . 149
5.7.1 Nonconforming finite elements . . . 149
5.7.2 Discontinuous Galerkin finite elements . . . 149
5.7.3 Mixed finite elements . . . 150
5.8 Numerical experiments . . . 151
5.8.1 Nonconforming finite element method . . . 151
5.8.2 Discontinuous Galerkin finite element method . . . 154
5.9 Concluding remarks . . . 156
5.10 Appendix . . . 157
5.10.1 Extension to a generic operator . . . 157
5.10.2 Further improvement of the first eigenvalue upper bound . . . 159

IV A posteriori error estimation for a nonlinear eigenvalue problem 161

6 A posteriori analysis of a nonlinear Gross–Pitaevskii-type eigenvalue problem 163
6.1 Introduction . . . 164
6.2 A priori analysis . . . 170
6.3 A posteriori analysis . . . 178
6.3.1 Preliminaries . . . 178
6.3.2 More accurate *a posteriori* estimate . . . 187
6.4 Numerical results . . . 194
6.4.1 General framework . . . 195
6.4.2 With a large number of iterations . . . 196

6.4.3 In large dimension for the discretization space . . . 196
6.4.4 Error balance . . . 197
6.5 Conclusions . . . 198

V Post-processing for the planewave approximation of linear and nonlinear Schrödinger operators 201

7 A perturbation-based post-processing method for the Kohn–Sham models 203
7.1 Introduction . . . 204
7.2 DFT Kohn–Sham models . . . 205
7.2.1 Introduction to Kohn–Sham models . . . 205
7.2.2 Periodic Kohn–Sham models . . . 206
7.2.3 Pseudopotentials . . . 207
7.3 Discretization and resolution of the Kohn–Sham model . . . 209
7.3.1 Planewave discretization . . . 209
7.3.2 SCF-iterations . . . 211
7.3.3 Smoothness assumptions and *a priori* results . . . 212
7.4 A post-processing based on perturbation theory . . . 213
7.5 Numerical results . . . 216
7.5.1 Simulations with a constant E_{c} . . . 216
7.5.2 Simulations with a constant $E_{\mathrm{c,res}}$ . . . 217
7.5.3 Simulations with a constant λ . . . 218

8 Post-processing of the plane-wave approximation of linear Schrödinger equations 221
8.1 Introduction . . . 222
8.2 Post-processing for the Kohn–Sham linear subproblem . . . 224
8.2.1 Problem setting . . . 224
8.2.2 Functional setting . . . 226
8.2.3 Discretization . . . 229
8.2.4 *A priori* results on the density matrices . . . 229
8.3 Post-processing of the plane-wave approximation . . . 231
8.3.1 A key remark . . . 231
8.3.2 Corrections computation . . . 231
8.4 Convergence improvement on the density matrix and the energy . . . 233
8.4.1 Main results . . . 233
8.4.2 Proofs . . . 233
8.5 Numerical results . . . 241
8.5.1 Convergence of the density matrix and the energy . . . 242
8.5.2 Comparison between different eigenvalue clusters . . . 243
8.5.3 Comparison of different regularities . . . 244

9 Post-processing of the plane-wave approximation of nonlinear Schrödinger equations 247
9.1 Introduction . . . 248
9.2 Periodic Kohn–Sham models with pseudopotentials . . . 249
9.2.1 Problem setting . . . 249
9.2.2 Functional setting . . . 251
9.3 Discretization and resolution of the Kohn–Sham model . . . 252

9.3.1 Planewave discretization 252
9.3.2 Smoothness assumptions and *a priori* results 253
9.4 Post-processing of the planewave approximation 256
9.5 Convergence improvement on the density matrix and the energy 258
9.5.1 Theorem 258
9.5.2 Proof 258
9.6 Appendix 268

Part I

Introduction

Chapter 1

Electronic structure calculation

1.1 Introduction

In this thesis, we focus on *ab initio* methods in quantum chemistry. The main advantage of these methods is that the models under consideration contain no empirical parameters, except a few fondamental constants of physics: the reduced Planck constant $\hbar$, the electron mass m_e, the elementary charge e, the dielectric permittivity of the vacuum ε_0, the Boltzmann constant k_B, as well as the masses and atomic numbers of the nuclei contained in the system under consideration. In the following, we will work in atomic units, which implies that

$$\hbar = 1, e = 1, m_e = 1, 4\pi\varepsilon_0 = 1.$$

In this setting, the elementary length is the Bohr radius and is equal $5.29.10^{-11}$ m. The mass unit is $9.11.10^{-31}$ kg. The energy unit is the Hartree and is equal to $4.36.10^{-18}$ J. Thus, the distance nucleus-electron is close to 1 for the hydrogen atom. And in general, the numbers under consideration are of order 1 or of order close to 1.

More precisely, we concentrate on the computation of the ground state electronic structure of a molecular system, i.e. the state of lowest energy of the system. In electronic structure calculation, the determination of the ground state is a standard, however challenging problem. It is needed for the computation of physical and chemical properties of the system, and is often used as an input for other problems, such as the determination of excited states.

The aim of this chapter is to give an overview of the different models used to compute the ground state of a molecular system, as well as practical resolution issues, such as discretization methods and algorithms. For a more detailed description of these models, the reader is referred to [230] and [135]. Besides, a mathematical presentation of the models can be found in [56] (in french), see also [45] (in english).

In this chapter, we start by presenting the Schrödinger model in Section 1.2. This model gives rise to a way too high-dimensional problem. There exist therefore many simplifications for this model, some of which are presented in Section 1.3. We focus particularly on Hartree–Fock and post Hartree–Fock methods, Density Functional Theory (DFT), and we mention Quantum Monte Carlo (QMC). In Section 1.4, we present the main discretization methods used in the field, which particularly apply for mean-field models, i.e. Hartree–Fock equations and Density Functional Theory models. Two other standard approximations used in the context of molecular simulation are described in Section 1.5: the supercell model, which is used to simulate molecules in a periodic framework, and pseudopotentials, which are used for different reasons, including model reduction and regularity issues. Finally, we present the algorithms used to solve numerically the equations of mean-field models in Section 1.6.

1.2 The Schrödinger model

1.2.1 Description of a molecular system

At the molecular scale, the theory of classical mechanics is no longer valid, and one has to use a quantum description of matter. A molecular system is hence not described by the collection of the positions and velocities of the particles, as it is the case in classical mechanics, but by its wave-function: a complex-valued function depending on the positions of all the particles of the system.

A priori, in a molecular system composed of M nuclei and N electrons, all particles should be treated as quantum particles. The wave-function should therefore depend on $3(N+M)$ variables. However, the mass of a proton being about 1838 times the mass of an electron, the nuclei are way heavier than the electrons, and in the vast majority of the simulations, the nuclei are considered as classical particles. This approximation is called Born–Oppenheimer approximation [31], and comes from two successive approximations. First, the adiabatic theorem [30] allows to write the wave-function as a product of a nuclear wave-function and an electronic wave-function. Second, the semi-classical limit of the nuclei leads to a classical time-evolution of the nuclei, as the electrons instantaneously relax in their ground-state. For more details about this approximation, the reader is referred to [195], [126], [5].

Therefore, we consider a physical system composed of

- M nuclei, considered as classical particles, with electric charges $z_1, z_2, \ldots, z_M \in \mathbb{N}^*$, and positions $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M \in \mathbb{R}^3$,
- N electrons, considered as quantum particles, described by a wave-function

$$\Psi \in \bigotimes_{i=1}^{N} L^2(\mathbb{R}^3, \mathbb{C}),$$

  i.e. a square integrable function depending on the spatial positions $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) \in \mathbb{R}^{3N}$ of the electrons. Here $\bigotimes$ denotes the tensor product. Note that any function in $L^2(\mathbb{R}^{3N}, \mathbb{C})$ can be approximated by functions in $\bigotimes_{i=1}^{N} L^2(\mathbb{R}^3, \mathbb{C})$. Therefore, the wave-function Ψ can also be seen as an element of $L^2(\mathbb{R}^{3N}, \mathbb{C})$.

According to the Pauli principle, the electronic wave-function is an antisymmetric function with respect to the inversion of the variables of two particles. This expresses the fact that electrons are indiscernible particles, which is a postulate of quantum mechanics, and that they are fermions, a property related to the spin of these particles. Therefore, the wave-function belongs in fact to the space

$$\mathcal{H}_e = \bigwedge_{i=1}^{N} L^2(\mathbb{R}^3, \mathbb{C}),$$

which is the antisymmetrized product space of $L^2(\mathbb{R}^3, \mathbb{C})$. The antisymmetric property of the wave-function reads

$$\forall i, j = 1, 2, \ldots, N, \quad \Psi(\ldots, \mathbf{r}_i, \ldots, \mathbf{r}_j, \ldots) = -\Psi(\ldots, \mathbf{r}_j, \ldots, \mathbf{r}_i, \ldots).$$

As $\Psi(\ldots, \mathbf{r}_i, \ldots, \mathbf{r}_i, \ldots) = 0$, the probability of having two electrons at the same position in space is equal to zero.

The space $\mathcal{H}_e$ is a Hilbert space and is endowed with the inner product

$$\forall \Psi, \Phi \in \mathcal{H}_e, \quad \langle \Psi | \Phi \rangle = \int_{\mathbb{R}^{3N}} \Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)^\star \Phi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) d\mathbf{r}_1 d\mathbf{r}_2 \ldots d\mathbf{r}_N,$$

where $\Psi^\star$ denotes the complex conjugate of Ψ. The corresponding norm is denoted by $\|\cdot\|$.

Physically, the square modulus of the wave-function $|\Psi(\mathbf{r}_1,\ldots,\mathbf{r}_N)|^2$ is the probability density of finding the N electrons at the positions $\mathbf{r}_1, \mathbf{r}_2$ to $\mathbf{r}_N$. Hence, the norm of the wave-function in $\mathcal{H}_e$ is equal to 1:

$$\|\Psi\|^2 = \int_{\mathbb{R}^{3N}} |\Psi(\mathbf{r}_1,\ldots,\mathbf{r}_N)|^2 \, d\mathbf{r}_1 \ldots d\mathbf{r}_N = 1.$$

An important notion on which rely several methods in electronic structure calculation is the electronic density. It is a real-valued function on $\mathbb{R}^3$ whose integral over the whole space is equal to the number of electrons and which represents the distribution of the N electrons in $\mathbb{R}^3$. It is defined as

$$\rho(\mathbf{r}) := N \int_{\mathbb{R}^{3(N-1)}} |\Psi(\mathbf{r},\mathbf{r}_2,\ldots,\mathbf{r}_N)|^2 \, d\mathbf{r}_2 \ldots d\mathbf{r}_N. \tag{1.2.1}$$

Note that in the following presentation of the models, we always omit the spin variables, even though the wave-function normally also depends on the spins of all the electrons, and should be an element of $\bigwedge_{i=1}^N L^2(\mathbb{R}^3 \times \{|+\rangle, |-\rangle\}, \mathbb{C})$, $|+\rangle$ and $|-\rangle$ representing respectively the spins up and down. This simplification of the notations in the equations can physically be viewed as considering only spinless or closed-shell systems. However, the numerical simulations presented hereafter for molecular systems are performed using closed-shells, hence taking into account the spin variables. This is in particular the case in Chapter 7.

1.2.2 Electronic time-independent Schrödinger equation

The time-independent ground-state problem corresponds to the energy minimization of the system under a specific configuration of the nuclei $(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M)$. It reads

$$E(\mathbf{R}_1,\mathbf{R}_2,\ldots,\mathbf{R}_M) = \inf_{\substack{\Psi\in\mathcal{H}_e,\\ \|\Psi\|=1}} \left\langle \Psi, \mathcal{H}_{(\mathbf{R}_1,\mathbf{R}_2,\ldots,\mathbf{R}_M)}\Psi \right\rangle, \tag{1.2.2}$$

where the Hamiltonian $\mathcal{H}_{(\mathbf{R}_1,\mathbf{R}_2,\ldots,\mathbf{R}_M)}$ of the system reads

$$\mathcal{H}_{(\mathbf{R}_1,\mathbf{R}_2,\ldots,\mathbf{R}_M)} = -\sum_{j=1}^N \left(\frac{1}{2}\Delta_{\mathbf{r}_j} + \sum_{\alpha=1}^M \frac{z_\alpha}{|\mathbf{R}_\alpha - \mathbf{r}_j|} \right) + \sum_{\substack{i,j=1\\ i<j}}^N \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}. \tag{1.2.3}$$

The Hamiltonian is composed of three parts:

- The first part
$$-\sum_{j=1}^N \frac{1}{2}\Delta_{\mathbf{r}_j}$$
models the kinetic energy contribution of the particles.

- The second part
$$-\sum_{j=1}^N \sum_{\alpha=1}^M \frac{z_\alpha}{|\mathbf{R}_\alpha - \mathbf{r}_j|}$$
models the Coulomb interaction between the nuclei and the electrons.

- The third part

$$+\sum_{\substack{i,j=1\\ i<j}}^{N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}$$

  models the Coulomb interaction between the electrons.

Writing the first-order optimality conditions of this energy minimization problem, one obtains the ground state eigenvalue problem proposed by Schrödinger in 1926 [223]: Find the lowest eigenvalue $E(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M)$ and corresponding eigenfunction $\Psi(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M)$ verifying

$$\mathcal{H}_{(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M)} \Psi(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M) = E(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M) \Psi(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M),$$

where $\Psi(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M)$ is called the ground-state wave-function, and $E(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M)$ is called the ground state energy. Here, the coefficients in $\mathcal{H}_{(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M)}$ are real hence the complex and real parts of the wave-function $\Psi(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M)$ are not mixed by the Hamiltonian and are both solutions to the ground state problem. We shall therefore consider only real-valued wave-functions in the following. Remark that in case of external interactions, further contributions may be added to the Hamiltonian. This is for example the case in the presence of an external magnetic field.

To obtain the molecular configuration of minimal energy, one needs to solve the so-called geometry optimization problem which corresponds to the minimization problem

$$\inf_{(\mathbf{R}_1, \ldots, \mathbf{R}_M) \in \mathbb{R}^{3M}} E(\mathbf{R}_1, \ldots, \mathbf{R}_M),$$

where $E(\mathbf{R}_1, \ldots, \mathbf{R}_M)$ is the solution to the electronic Schrödinger equation depending on the parameters $(\mathbf{R}_1, \ldots, \mathbf{R}_M)$.

In the sequel, we focus only on the resolution of the electronic problem, and we forget the dependency on the nuclei. The electronic Hamiltonian will be denoted by $\mathcal{H}$, and the wave-function by Ψ.

1.2.3 Limitations of the Schrödinger model

The electronic Schrödinger equation gives a very good approximation of the ground-state electronic structure for a large variety of systems, in which case the Schrödinger model can be considered as a reference model. For many systems, the theoretical results are indeed in very good agreement with experiments.

However, this model neglects all relativistic effects, which appear in particular in the case of heavy atoms, whose core electrons have a velocity close to the light velocity. The effect of these core electrons is to screen the nuclear charge for the valence electrons. This explains for example the color difference between gold and silver, which is not predicted without taking relativistic effects into account [209]. In 1928, Dirac proposed an equation to include these effects [89], and Breit complemented it four years later. This was further developed in a theory called quantum electrodynamics (QED).

Note that the Born–Oppenheimer approximation can also be a limitation of the Schrödinger model. Indeed, when ground and excited electronic surfaces get very close, the adiabatic theorem does not hold any more, and therefore the Born–Oppenheimer approximation is no longer valid. In practice, it is very rarely considered as being the case, and the major part of numerical simulations are performed within this approximation.

1.3 Approximations for the electronic Schrödinger equation

For a system with N electrons, the electronic wave-function depends on $3N$ variables. A naive discretization with only 10 points per dimension would lead to a discretization space with 10^{3N} points, which grows exponentially with N and becomes untractable when N grows, even for supercomputers. In order to lower the dimensionality of the problem, different approximate models have been developed. We present here several of them.

1.3.1 Hartree–Fock and post Hartree–Fock methods

Hartree–Fock model

The Hartree–Fock model consists of assuming a specific form for the wave-function, and thus minimizing the energy over a smaller set of antisymmetric functions. Instead of the antisymmetrized tensor product $\bigwedge_{i=1}^N L^2(\mathbb{R}^3,\mathbb{C})$, only determinants of three-dimensional functions are considered. These functions are called Slater determinants [226], and write

$$\Psi(\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_N)=\frac{1}{\sqrt{N!}}\begin{vmatrix}\psi_1(\mathbf{x}_1) & \psi_2(\mathbf{x}_1) & \cdots & \psi_N(\mathbf{x}_1)\\ \psi_1(\mathbf{x}_2) & \psi_2(\mathbf{x}_2) & \cdots & \psi_N(\mathbf{x}_2)\\ \vdots & \vdots & \ddots & \vdots\\ \psi_1(\mathbf{x}_N) & \psi_2(\mathbf{x}_N) & \cdots & \psi_N(\mathbf{x}_N)\end{vmatrix}. \tag{1.3.1}$$

The functions $\psi_1,\dots,\psi_N\in L^2(\mathbb{R}^3,\mathbb{R})$ are called orbitals, and are usually chosen orthonormal in $L^2(\mathbb{R}^3,\mathbb{R})$. The factor $\frac{1}{\sqrt{N!}}$ is a normalization factor. A Slater determinant with orbitals $\psi_1,\dots,\psi_N$ is denoted by

$$\Psi=|\psi_1,\dots,\psi_N\rangle.$$

The energy minimization problem can then be written as

$$E_{HF}^0=\inf_{\Psi\in V_{HF}}\langle\Psi|\mathcal{H}\Psi\rangle=\inf_{\Psi\in V_{HF}}E_{HF}(\Psi), \tag{1.3.2}$$

where $\mathcal{H}$ is the electronic Hamiltonian defined in (1.2.3), V_{HF} is the set of Slater determinants

$$V_{HF}=\big\{\Psi=|\psi_1,\dots,\psi_N\rangle\,|\,\psi_i\in H^1(\mathbb{R}^3),\ \langle\psi_j|\psi_i\rangle=\delta_{ij}\big\}, \tag{1.3.3}$$

and $E_{HF}(\Psi)$ is defined below. Since $V_{HF}\subset\bigwedge_{i=1}^N L^2(\mathbb{R}^3,\mathbb{C})$, the Hartree–Fock energy is always larger or equal than the exact ground state energy. This is called the variational principle.

Still denoting by $\Psi=\{\psi_i\}$ the wave-function built on the orbitals $\psi_1,\dots,\psi_N$, the Hartree–Fock energy of a Slater determinant Ψ can be written in term of the orbitals $\psi_1,\dots,\psi_N$ as

$$\begin{aligned}E_{HF}(\Psi)=&\sum_{j=1}^N\frac{1}{2}\int_{\mathbb{R}^3}|\nabla\psi_j|^2+\int_{\mathbb{R}^3}\rho_\Psi V+\frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\rho_\Psi(\mathbf{x})\rho_\Psi(y)}{|\mathbf{x}-\mathbf{y}|}d\mathbf{x}d\mathbf{y}\\&-\frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{|\sum_{j=1}^N\psi_j(\mathbf{x})\psi_j(\mathbf{y})|^2}{|\mathbf{x}-\mathbf{y}|}d\mathbf{x}d\mathbf{y},\end{aligned}$$

where the electronic density is

$$\rho_\Psi(\mathbf{x})=\sum_{j=1}^N|\psi_j(\mathbf{x})|^2,$$

and the nuclear potential is

$$V(\mathbf{x}) = \sum_{\alpha=1}^{M} \frac{z_\alpha}{|\mathbf{R}_\alpha - \mathbf{x}|}.$$

A proof of existence for the solutions of the Hartree–Fock problem can be found in [171], see also [175].

The Euler–Lagrange equations of the problem read in a strong form: Find the set of orbitals $\Psi_0 = \{\psi_1, \dots, \psi_N\}$ building the Slater determinant $|\psi_1, \dots, \psi_N\rangle$ and the Lagrange-multipliers $(\varepsilon_{i,j})_{1\leq i,j\leq N}$ such that for all $j = 1, \dots, N$

$$\begin{aligned} &-\frac{1}{2}\Delta_{\mathbf{x}}\psi_j(\mathbf{x}) + V(\mathbf{x})\psi_j(\mathbf{x}) + \sum_{i=1}^{N}\int_{\mathbb{R}^3}\frac{|\psi_i(\mathbf{y})|^2}{|\mathbf{x}-\mathbf{y}|}d\mathbf{y}\,\psi_j(\mathbf{x}) \\ &-\sum_{i=1}^{N}\int_{\mathbb{R}^3}\frac{\psi_i(\mathbf{y})\psi_j(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}d\mathbf{y}\,\psi_i(\mathbf{x}) = \sum_{i=1}^{N}\varepsilon_{ij}\psi_i(\mathbf{x}) \end{aligned}$$

and for all $i, j = 1, \dots, N$,

$$\langle\psi_j|\psi_i\rangle = \delta_{ij}.$$

The Hartree-Fock energy being invariant by a unitary transformation on the orbitals ψ_1 to ψ_N, the problem can be rewritten as

$$\begin{aligned} &\text{Find } \Psi_0 = \{\psi_1, \dots, \psi_N\} \text{ and } (\varepsilon_1, \dots, \varepsilon_N) \in \mathbb{R}^N \text{ such that:} \\ &\qquad \mathcal{F}_{\Psi_0}\psi_j = \varepsilon_j\,\psi_j, \qquad \forall j = 1, \dots, N \\ &\qquad \langle\psi_j|\psi_i\rangle = \delta_{ij}, \qquad \forall i, j = 1, \dots, N, \end{aligned} \tag{1.3.4}$$

where the action of the Fock operator $\mathcal{F}_\Psi$ with $\Psi = \{\psi_1, \dots, \psi_N\}$, for any arbitrary set of orbitals $\Phi = \{\varphi_1, \dots, \varphi_N\}$ is given by

$$\mathcal{F}_\Psi\varphi_j = -\frac{1}{2}\Delta_{\mathbf{x}}\varphi_j(\mathbf{x}) + V(\mathbf{x})\varphi_j(\mathbf{x}) + \sum_{i=1}^{N}\int_{\mathbb{R}^3}\frac{|\psi_i(\mathbf{y})|^2}{|\mathbf{x}-\mathbf{y}|}d\mathbf{y}\,\varphi_j(\mathbf{x}) - \sum_{i=1}^{N}\int_{\mathbb{R}^3}\frac{\psi_i(\mathbf{y})\varphi_j(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}d\mathbf{y}\,\psi_i(\mathbf{x}).$$

Due to the dependency of the Fock operator on the orbitals $\psi_1, \dots, \psi_N$, this problem is a nonlinear eigenvalue problem. Hence, the linearity of the full Schrödinger problem is lost, but the dimensionality of the problem is way smaller, as one only needs to determine N three-dimensional functions.

The Hartree–Fock model offers a very good compromise between simplicity and accuracy. It predicts around 90% of the total energy in general and up to 99% for some systems. However, the remaining part of the energy can hide interesting physical phenomena that cannot be predicted with the Hartree–Fock model only. For this reason, different methods have been designed to capture the last 1 to 10%. Based on a Hartree–Fock computation as an input, they are called post Hartree–Fock methods.

Post Hartree–Fock Configuration Interaction method

In the Configuration Interaction (CI) method, the wave-function is searched as a linear combination of Slater determinants

$$\Psi = \sum_i \alpha_i|\psi_1^i, \dots, \psi_N^i\rangle.$$

In this sum, the Slater determinants are built from well-chosen eigenfunctions of the Fock operator $\mathcal{F}_{\Psi_0}$ (1.3.4). Note that the Hartree–Fock determinant is built from the eigenfunctions

of the Fock operator which correspond to the N lowest eigenvalues, while Slater determinants built from orbitals containing eigenfunctions whose eigenvalues are not in the N lowest are called excited determinants.

In a discrete basis (the discretization methods will be presented in Section 1.5), if all Slater determinants are considered, the method is called full configuration interaction (FCI), and is equivalent to a discrete version of the Schrödinger equation. But the number of determinants to consider is huge, as it is equal to $\binom{\mathcal{N}!}{N}$, with $\mathcal{N}$ the number of basis functions for the discretization of one orbital.

Therefore, in practice, only a small subset of Slater determinants is considered. In general, it corresponds to all Slater determinants differing from the Hartree–Fock determinant by at most two, three or four orbitals. The methods are in this case respectively called CISD (configuration interaction single double), CISDT (configuration interaction single double triple), CISDTQ (configuration interaction single double triple quadruple). The more determinants considered in the linear combination, the more precise the method, but also the more costly. Indeed, mathematically, the problem is a linear eigenvalue problem arising from the problem of minimizing the energy $\langle \Psi | \mathcal{H} \Psi \rangle$ over the set of considered determinants, so the cost is cubic with respect to the number of determinants.

This method gives particularly good results when the error of the Hartree–Fock method is already very small. When several determinants have an important contribution in the linear combination, the configuration interaction method is not very accurate. In this case, one has to use a multi-configuration method (see Section 1.3.2).

An important drawback of the CI method is that it is not size-consistent: the energy of two separated systems is not necessarily equal to the sum of the energies of the two systems. For this reason, other methods such as Coupled Cluster (CC) are often preferred over the CI method.

Post Hartree–Fock Coupled Cluster

The Coupled Cluster (CC) method, which guarantees size-consistency, uses a different ansatz for the wave-function compared to the CI method. The wave-function is written as an exponential operator called cluster operator applied to the Hartree–Fock determinant. The cluster operator, which has to be determined, couples the different orbitals of the Hartree–Fock problem. Loosely speaking, the wave-function is based on a linear combination of many determinants, with coefficients related by nonlinear relations. This allows, for a fixed number of coefficients, to include more determinants than in the CI method. The resulting equations are nonlinear, but can be solved using an iterative algorithm.

This is one of the reference methods to get accurate results for small to medium size molecules. A presentation of the method from a mathematical viewpoint can be found in [221]. An error analysis for this method has been provided in [216].

Møller–Plesset perturbation theory

The Hartree–Fock solution can also be improved by means of Rayleigh–Schrödinger perturbation theory, as in Møller–Plesset (MP) perturbation theory. In this method, the exact Hamiltonian is considered as a perturbation of the Fock operator. A perturbation parameter $\lambda \in [0, 1]$ is introduced as well as an intermediate Hamiltonian

$$\mathcal{H}(\lambda) = \mathcal{F}_{\Psi_0} + \lambda(\mathcal{H} - \mathcal{F}_{\Psi_0}),$$

so that $\mathcal{H}(0)$ is the Fock operator and $\mathcal{H}(1)$ is the exact Hamiltonian. Then the wave-function and the energy are written as power series depending on the perturbation parameter λ. The

exact equation is then expanded up to a certain order in λ. Note that the perturbative series may not be convergent. Then, taking $\lambda = 1$ provides corrections for the wave-function and the energy. An example of perturbation theory will be presented in Section 2.5.2. The perturbative development at different orders gives the different names for the method: MP2, MP3, MP4, and so on.

1.3.2 Multi-configuration methods

For some molecular systems, e.g. when the ground state is quasi-degenerate, which corresponds to a case of static correlation, Hartree–Fock and post Hartree–Fock methods fail to give an accurate description of the molecule. Multi-configuration methods have been particularly developed to overcome this problem.

Instead of considering a single determinant as in the Hartree–Fock model, the wave-function is written as a linear combination of K determinants and the Schrödinger energy is directly minimized on the space

$$V_{MC} = \left\{ \Psi = \sum_{i=1}^{K} \alpha_i |\psi_1^i, \dots, \psi_N^i\rangle,\ (\alpha_i)_{1\leq i\leq K} \in \mathbb{R}^K, \psi_j^i \in H^1(\mathbb{R}^3), \langle \psi_k^i | \psi_j^i \rangle = \delta_{kj}, 1 \leq j \leq N \right\}.$$

This means that the orbitals are optimized at the same time as the coefficients in the linear combination. Of course, compared to the Hartree–Fock problem, the resolution of this problem is computationally more demanding. One of these methods is called MCSCF for Multi-Configuration Self Consistent Field, and was proposed in [137], see also [249].

1.3.3 Density Functional Theory

In Density Functional Theory (DFT), the main object of interest is not the wave-function, but the electronic density defined in (1.2.1). It was first developed by Hohenberg and Kohn in [139], and later on by Levy and Lieb [169]. Considering the electronic density which depends only on three space variables reduces the dimension of the original problem. But it requires to express the Schrödinger problem in terms of the density, which is not trivial in general.

A key idea is to reformulate the electronic Schrödinger minimization problem by grouping the wave-functions having the same electronic density, and writing the problem as a minimization over the set of densities. Formally,

$$\min_{\Psi} \langle \Psi | \mathcal{H} | \Psi \rangle = \min_{\rho} \left[\min_{\Psi, \Psi \to \rho} \langle \Psi | \mathcal{H} | \Psi \rangle \right].$$

Then defining the functional of the density

$$E(\rho) = \min_{\Psi, \Psi \to \rho} \langle \Psi | \mathcal{H} | \Psi \rangle,$$

the original problem can be written only in terms of the density as

$$\min_{\rho} E(\rho).$$

Of course, finding $E(\rho)$ is as difficult as finding the exact wave-function. Therefore, in practice, the functional $E(\rho)$ is modeled. But once this functional is given, the problem is reduced to a minimization problem over the space of densities. Moreover, the space of densities arising from admissible wave-functions, called space of N-representable densities, is explicitly known. It is:

$$I_N = \left\{ \rho,\ |\ \rho \geq 0, \quad \int_{\mathbb{R}^3} \rho(\mathbf{r}) d\mathbf{r} = N, \quad \int_{\mathbb{R}^3} |\rho^{1/2}(\mathbf{r})|^2 d\mathbf{r} < +\infty \right\}.$$

Different expressions for the energy functional $E(\rho)$ have been proposed in the past. They can be grouped into two main categories: the ones in which the energy functional depends only on the density, called orbital-free models, and the Kohn–Sham models. We present some of them in the following. For more details, an interesting review article on the subject can be found in [39]. Therafter, we also introduce the Gross–Pitaevskii equation, which has a physical origin different from density functional theory, but is mathematically very close to the orbital-free models.

Orbital-free models

The first orbital-free models were inspired from the behavior of the electron gas, and proposed in 1927 independently by Thomas [233] and Fermi [107]. In the setting described in Section 1.2.1, the electrostatic potential generated by the nuclei and felt by the electrons is

$$V^{\mathrm{nuc}}(\mathbf{r}) = -\sum_{k=1}^{M} \frac{z_k}{|\mathbf{r}-\mathbf{R}_k|}. \tag{1.3.5}$$

In all orbital-free models, the energy functional is explicit in terms of the electronic density, and the minimization problem reads

$$\inf\left\{F(\rho) + \int_{\mathbb{R}^3} \rho V^{\mathrm{nuc}},\ \rho \geq 0,\ \sqrt{\rho} \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3} \rho = N\right\}. \tag{1.3.6}$$

The expression of the functional F depends on the model under consideration. In the Thomas–Fermi model, the functional reads

$$F(\rho) := C_{\mathrm{TF}} \int_{\mathbb{R}^3} \rho^{5/3} + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(\mathbf{x})\rho(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|} d\mathbf{x}d\mathbf{y},$$

where the first term models the kinetic energy of the electrons, and the second terms models the interaction between electrons.

In the Thomas–Fermi–von Weizsäcker model originally introduced in [243], the functional reads

$$F(\rho) := \frac{C_{\mathrm{W}}}{2} \int_{\mathbb{R}^3} |\nabla\sqrt{\rho}|^2 + C_{\mathrm{TF}} \int_{\mathbb{R}^3} \rho^{5/3} + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(\mathbf{x})\rho(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|} d\mathbf{x}d\mathbf{y}.$$

In this expression, the first two terms model the kinetic energy of the electrons, and the third term models the electron-electron interactions.

In the Thomas–Fermi–Dirac–von Weizsäcker, the functional reads

$$F(\rho) := \frac{C_{\mathrm{W}}}{2} \int_{\mathbb{R}^3} |\nabla\sqrt{\rho}|^2 + C_{\mathrm{TF}} \int_{\mathbb{R}^3} \rho^{5/3} - C_D \int_{\mathbb{R}^3} \rho^{4/3} + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(\mathbf{x})\rho(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|} d\mathbf{x}d\mathbf{y},$$

where the third term is called Dirac exchange functional. In atomic units, the constants for the three models are $C_{TF} = \left(\frac{3^{5/3}\pi^{4/3}}{10}\right)$, $C_{\mathrm{W}} = 1$, $1/5$ or $1/9$ depending on the context [93], and $C_D = \frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}$.

Despite their simplicity, these models are not much used any more for real calculations as they lack of accuracy. However, they are interesting from a mathematical point of view, because they are scalar models, hence reasonably simple, while they keep some of the properties present in more complicated models, still in use nowadays, such as Kohn–Sham models.

Kohn–Sham models

In Kohn–Sham models [152], the ground-state minimization problem can still be written as

$$\inf\left\{F(\rho)+\int_{\mathbb{R}^3}\rho V^{\mathrm{nuc}},\ \rho\geq 0,\ \sqrt{\rho}\in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3}\rho=N\right\},$$

but in this case, the reference system is a system of non-interacting electrons, and not the electron gas as in Thomas–Fermi type models. The functional $F(\rho)$ is decomposed as

$$F(\rho)=T_{KS}(\rho)+J(\rho)+E_{\mathrm{xc}}(\rho),$$

where the kinetic part is

$$T_{KS}(\rho)=\inf\left\{\frac{1}{2}\sum_{i=1}^{N}\int_{\mathbb{R}^3}|\nabla\psi_i|^2,\quad \Psi=\{\psi_i\}_{1\leq i\leq N}\in(H^1(\mathbb{R}^3))^N,\quad \rho_\Psi=\rho\right\},$$

the electronic density being defined as

$$\rho_\Psi(\mathbf{x}):=2\sum_{i=1}^{N}|\psi_i(\mathbf{x})|^2,$$

and

$$J(\rho)=\frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\rho(\mathbf{x})\rho(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}d\mathbf{x}d\mathbf{y}.$$

The last part is $E_{\mathrm{xc}}(\rho)=F(\rho)-T_{KS}(\rho)-J(\rho)$ called exchange-correlation functional is a correction term, which is essential to describe quantitatively, and sometimes even qualitatively, the physics and chemistry of the system.

Thus, the Kohn–Sham minimization problem can be written as

$$\inf\left\{E^{\mathrm{KS}}(\Psi),\ \Psi=\{\psi_i\}_{1\leq i\leq N}\in(H^1(\mathbb{R}^3))^N\right\},\tag{1.3.7}$$

with

$$E^{\mathrm{KS}}(\Psi):=\sum_{i=1}^{N}\int_{\mathbb{R}^3}|\nabla\psi_i|^2+\int_{\mathbb{R}^3}V^{\mathrm{nuc}}\rho_\Psi+\frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\rho_\Psi(\mathbf{x})\rho_\Psi(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}d\mathbf{x}d\mathbf{y}+E_{\mathrm{xc}}(\rho_\Psi).$$

In this energy functional, the first term models the kinetic energy, the second term the interactions between nuclei and electrons, and the third term the interaction between electrons. The fourth term corrects the error introduced by the first three terms. Theoretically, there exists an *exact* exchange-correlation functional (see e.g. [139, 167, 169, 237]), i.e. a functional depending only on the electronic density ρ_Ψ such that the solution of the Kohn–Sham minimization problem is the exact ground state density of the N-body Schrödinger equation. Unfortunately, such exact exchange-correlation functional is not known, and different approximations are commonly used.

A classical exchange-correlation functional is called local density approximation (LDA) and consists in approximating the exchange-correlation functional by

$$\int_{\mathbb{R}^3}e_{\mathrm{xc}}^{\mathrm{LDA}}(\rho(\mathbf{x}))\,d\mathbf{x},$$

where $e_{\mathrm{xc}}^{\mathrm{LDA}}(\overline{\rho})$ is an approximation of the exchange-correlation energy per unit volume in a uniform electron gas with density $\overline{\rho}$. Among these functionals can be found the Perdew–Zunger [201] and Perdew–Wang [200] functionals. This Kohn–Sham LDA model is properly understood from a mathematical viewpoint [6].

For this model, the Euler–Lagrange equations of the model read in a strong form: Find the set of orbitals $\Psi_0 = \{\psi_1, \dots, \psi_N\}$ building the Slater determinant $|\psi_1, \dots, \psi_N\rangle$ such that for all $j = 1, \dots, N$

$$-\frac{1}{2}\Delta_{\mathbf{x}}\psi_j(\mathbf{x}) + V^{\mathrm{nuc}}(\mathbf{x})\psi_j(\mathbf{x}) + \sum_{i=1}^{N}\int_{\mathbb{R}^3}\frac{|\psi_i(\mathbf{y})|^2}{|\mathbf{x}-\mathbf{y}|}d\mathbf{y}\;\psi_j(\mathbf{x}) + \frac{de_{\mathrm{xc}}^{\mathrm{LDA}}}{d\rho}(\rho_{\Psi_0}(\mathbf{x}))\psi_j(\mathbf{x}) = \sum_{i=1}^{N}\lambda_{ij}\psi_i(\mathbf{x})$$

and for all $i, j = 1, \dots, N$,

$$\langle\psi_j|\psi_i\rangle = \delta_{ij}.$$

The Kohn–Sham energy being invariant by a unitary transformation on the orbitals ψ_1 to ψ_N, exactly as the Hartree–Fock energy, the problem can be rewritten as

$$\begin{aligned}&\text{Find } \Psi_0 = \{\psi_1, \dots, \psi_N\} \text{ such that:}\\ &\qquad \mathcal{F}_{\Psi_0}\psi_j = \varepsilon_j\,\psi_j, \qquad \forall j = 1, \dots, N\\ &\qquad \langle\psi_j|\psi_i\rangle = \delta_{ij}, \qquad \forall i, j = 1, \dots, N,\end{aligned} \tag{1.3.8}$$

where the action of the operator $\mathcal{F}_\Psi$ with $\Psi = \{\psi_1, \dots, \psi_N\}$, for any arbitrary set of orbitals $\Phi = \{\varphi_1, \dots, \varphi_N\}$ is given by

$$\mathcal{F}_\Psi\varphi_j = -\frac{1}{2}\Delta_{\mathbf{x}}\varphi_j(\mathbf{x}) + V^{\mathrm{nuc}}(\mathbf{x})\varphi_j(\mathbf{x}) + \frac{1}{2}\int_{\mathbb{R}^3}\frac{\rho_\Psi(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}d\mathbf{y}\,\varphi_j(\mathbf{x}) + \frac{de_{\mathrm{xc}}^{\mathrm{LDA}}}{d\rho}(\rho_{\Psi_0}(\mathbf{x}))\varphi_j(\mathbf{x}).$$

Note that this formulation of the problem is very close to the Hartree–Fock problem. The main difference is the expression of the operator $\mathcal{F}_\Psi$ and in particular the presence of the exchange-correlation functional. The meaning of the orbitals is also different: in the Kohn–Sham model, the orbitals do not have a physical meaning, only the electronic density is relevant, whereas in the Hartree–Fock model, the orbitals can be used to build the wave-function of the system. In both case, the problem is a nonlinear eigenvalue problem.

More accurate and complicated exchange-correlation functionals have been developed, for example generalized gradient approximations (GGA) [199], which rely not only on the electronic density but on its gradient. Among these hybrid functionals can be found Becke's functional introduced in 1993 [20]. The most popular hybrid functional nowadays in chemistry is called B3LYP [163]. From a mathematical viewpoint, these problems are less understood, and many questions remain open. In any case, the Kohn–Sham models are among the most widely used nowadays, because they provide a good compromise between accuracy and computational cost.

Gross–Pitaevskii equation

Let us now mention another mean-field model arising in a different field: the Gross–Pitaevskii equation [204], for which a derivation is presented in [170]. In Chapter 6, we present some results on this equation, which can be seen as a toy model for density functional theory, and in particular orbital-free models, even though it does not find its origin in the quantum chemistry domain. This equation indeed arises from modeling bosons at very low temperature. Due to the bosonic nature of the particles, the particles all fall into the same quantum state, which is the state of lowest energy of the system. This phenomenon is called Bose–Einstein condensation. For a system of N bosons, the wave-function Ψ of the system can therefore be decomposed as a product, hence there exists a complex-valued function ψ on $\mathbb{R}^3$ such that

$$\Psi(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) = \psi(\mathbf{r}_1)\psi(\mathbf{r}_2)\dots\psi(\mathbf{r}_N).$$

Besides, the problem of minimizing the energy of the system can be expressed only in terms of ψ. This is a constrained minimization problem as in the case of orbital-free models,

$$\inf\left\{E(\psi), \quad \psi : \mathbb{R}^3 \to \mathbb{C}, \quad \int_{\mathbb{R}^3} |\psi(\mathbf{r})|^2 d\mathbf{r} = 1\right\},$$

where the energy functional E reads

$$E(\psi) = \frac{1}{2}\int_{\mathbb{R}^3} |\nabla\psi(\mathbf{r})|^2 \, d\mathbf{r} + \frac{1}{2}\int_{\mathbb{R}^3} V(\mathbf{r})|\psi(\mathbf{r})|^2 + g\int_{\mathbb{R}^3} |\psi(\mathbf{r})|^4.$$

The function V is an external potential, called confining potential and $g \in \mathbb{R}$ is a coupling constant. Writing the Euler–Lagrange equation of this minimization problem, one obtains the time-independent Gross–Pitaevskii equation

$$-\Delta\psi + V\psi + 4g|\psi|^2\psi = \mu\psi, \tag{1.3.9}$$

where $\mu \in \mathbb{R}$, called chemical potential, is the Lagrange multiplier of the constraint $\int_{\mathbb{R}^3} |\psi(\mathbf{r})|^2 d\mathbf{r} = 1$.

This equation is a nonlinear eigenvalue problem, hence encloses two of the difficulties that will be treated hereafter: the nonlinearity and the eigenvalue issues. However, this problem is not too difficult for at least three reasons: the energy functional is convex, the nonlinearity is explicit and can be well handled, and only the smallest eigenvalue of the operator is sought. Therefore, one does not need to consider several eigefunctions as in the case of Kohn–Sham models. When a magnetic field is applied to the system, the equation becomes more complex, but this goes beyond the scope of this thesis.

1.3.4 Quantum Monte Carlo

Beyond Hartree–Fock, post Hartree–Fock, and DFT models, a third class of models used in electronic structure calculation is based on stochastic calculations, namely Quantum Monte Carlo (QMC) methods [127, 166, 165]. There exist several types of QMC methods, among which the variational Monte Carlo (VMC) and the diffusion Monte Carlo (DMC). Let us briefly mention these two methods now.

In variational Monte Carlo, one considers a set of wave-functions $\Psi \in V$. The ground-state problem (1.2.2) can be written as

$$\inf_{\Psi\in V} E(\Psi) \;=\; \inf_{\Psi\in V} \frac{\langle\Psi, \mathcal{H}\Psi\rangle}{\langle\Psi,\Psi\rangle}.$$

In fact, the energy can be expressed as

$$E(\Psi) = \frac{\displaystyle\int_{\mathbb{R}^{3N}} |\Psi(\mathbf{r}_1,\dots,\mathbf{r}_N)|^2 \frac{\mathcal{H}\Psi(\mathbf{r}_1,\dots,\mathbf{r}_N)}{\Psi(\mathbf{r}_1,\dots,\mathbf{r}_N)} \, d\mathbf{r}_1\dots d\mathbf{r}_N}{\displaystyle\int_{\mathbb{R}^{3N}} |\Psi(\mathbf{r}_1,\dots,\mathbf{r}_N)|^2 \, d\mathbf{r}_1\dots d\mathbf{r}_N}.$$

As the function $\dfrac{|\Psi(\mathbf{r}_1,\dots,\mathbf{r}_N)|^2}{\int_{\mathbb{R}^{3N}} |\Psi(\mathbf{r}_1,\dots,\mathbf{r}_N)|^2 \, d\mathbf{r}_1\dots d\mathbf{r}_N}$ can be interpreted as a probability distribution, a Monte Carlo method can be used for evaluating the integral, and compute the energy for a given wave-function Ψ. Then, a minimization procedure can be applied to minimize the energy and find the ground state.

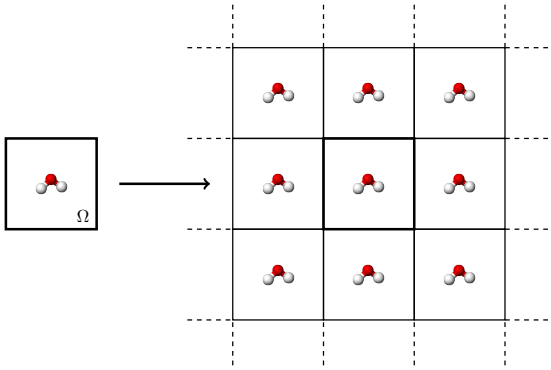

Figure 1.1 – Example of a periodic lattice for a molecular system

Diffusion Monte Carlo (DMC) is based on the representation of the solution to the Schrödinger equation as an expected value of a specific process. The random variable has then to be sampled, and variance reduction methods are used in order to obtain accurate results.

The main advantage of the Monte Carlo methods comes from the ability of sampling a high-dimensional space, which is more feasible than solving a partial differential equation on the same space.

1.4 Other standard approximations

In addition to the necessary model simplifications, two approximations are very standard in the simulation of molecular systems: the supercell model, which allows to simulate molecular systems within a periodic framework, and pseudopotentials, which reduce the number of degrees of freedom in the models as well as increase regularity properties. We present these two approximations in the following.

1.4.1 Supercell model

The supercell model is essentially used in two different contexts: for the simulation of molecular systems with periodic boundary conditions, and for the simulation of crystals, with or without defects.

In the framework of condensed matter, the density of the system under consideration is defined on $\mathbb{R}^3$. It is therefore impossible to simulate the whole domain. In practice, the domain is truncated, or said differently, the molecular system is placed in a very large cubic box, for which side effects are negligible. In this context, the use of periodic boundary conditions corresponds to replicate the box in the three space directions, as in Figure 1.1. Generally, the computational domain, also called the supercell, is denoted by Ω and is the unit cell of a periodic lattice $\mathcal{R}$. Of course, the domain Ω must be large enough so that the interactions between two consecutive unit cells are negligible, but if the domain Ω is too large, the computations become extremely expensive. For this reason, there is a trade-off between the accuracy and the computational cost of the simulations. Note that for periodic boundary conditions, the standard discretization method is planewaves (Fourier series), and is presented in Section 1.5.3.

When the system under consideration is charged, it is not possible to simulate the molecule as such, because the periodic system has infinite positive or negative charges. To overcome this problem, a uniform electron gas called jellium is added to the system, so that the total charge of the uniform gas compensates exactly the charge of the system. To compute the electronic

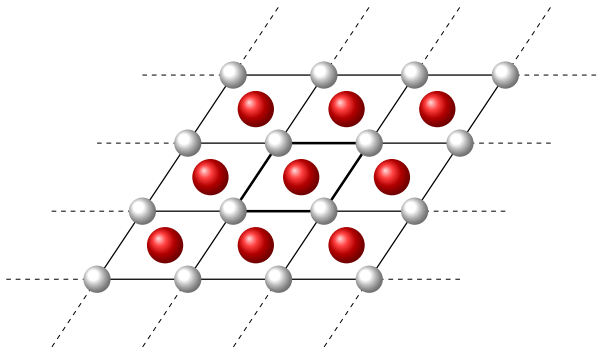

Figure 1.2 – Example of a periodic lattice of a crystal

density of the charged system, one only needs to remove the electron density of the uniform electron gas in the background to the density of the neutral system.

The supercell method is also used in the context of crystals, which is very different, the simulated systems being of infinite size. Since crystals without defects are periodic, the periodic framework fits particularly well for such systems. The box used for the simulations depends on the geometry of the crystal, and usually corresponds to the unit cell of the crystal, see Figure 1.2. When the crystal has defects, it is possible to choose a larger supercell, consisting of a given (possibly large) number of unit cells. However, the electronic structure problems posed in the context of crystal are different and rely on Bloch theory [24]. They require to solve equations parametrized by wavevectors $\mathbf{k}$ called $\mathbf{k}$-points, belonging to the reciprocal space of the periodic lattice. In this thesis, we focus only on the simulation of condensed matter systems, therefore we do not describe this theory in more detail.

In this thesis, we consider the choice of the supercell as a model error, whereas it could alternatively be seen as a discretization error, as emphasized in the introduction of Chapter 3.

1.4.2 Pseudopotentials

Pseudopotentials model the effect of core electrons. They are very much used in solid state physics, materials science, and also for the simulation of molecular systems with heavy atoms, which have many core electrons. This approach, already proposed by Hellmann in 1935, is based on the observation that the core electrons of an atom are not much affected by the chemical environment of this atom. Therefore, the use of pseudopotentials reduces the electronic problem to a problem on the valence electrons only, modeling by a pseudopotential the effect of the core electrons. Instead of the neutral core composed with the nucleus, one considers a ionic core composed of the nucleus and the core electrons. Since the problem has to be solved only for the valence electrons, the use of pseudopotentials brings an important dimensionality reduction, and hence lowers the computational cost.

The use of pseudopotentials also increases the regularity of the wave-functions under consideration. Indeed, as the Coulombic potential is replaced by a smoother potential, the valence orbitals are more regular than in all-electron calculations. This means in particular that, for a planewave discretization (presented in Section 1.5.3), the orbitals can be represented with fewer basis functions. Moreover, some pseudopotentials incorporate relativistic effects, without using a relativistic treatment of the equations.

There exists a large variety of pseudo-potentials, for example the Kerker's pseudopotentials [148], Troullier–Martins [235], Goedecker pseudo-potential [117], norm-conserving pseudopotentials, or Vanderbilt ultrasoft pseudopotentials [238]. We refer to [58] for a clarification of

the mathematical framework on norm-conserving pseudo-potentials.

1.5 Discretization methods

1.5.1 Different codes for different discretizations

The mean-field models presented in Sections 1.3.1 and 1.3.3 take the form of partial differential equations posed on continuous spaces. Numerically, only finite dimensional problems can be handled, so the equations are discretized, using in the vast majority of the cases a Galerkin approximation in a given basis set, or Taylor expansions (finite differences, see Section 1.5.6). In the following, we present the main discretization methods: molecular orbitals, planewaves, wavelets, finite elements and finite differences.

Depending on the physical system under consideration, different methods or basis sets are better suited. For example, periodic systems such as crystals are commonly discretized with periodic basis sets. Other methods are used with Dirichlet boundary conditions. Also, the number of basis functions necessary to represent the solutions varies a lot between the different discretization methods, and plays an important role in the computational time of the simulations. Indeed, a priori, a computation done with few basis functions will be less expensive than a computation with many basis functions. However, other factors have a large influence on the computational cost, for example the time needed for computing bielectronic integrals, which are integrals depending on the product of four basis functions.

There exists a large variety of codes based on different basis sets to solve electronic structure problems numerically. These codes are based on different languages or softwares (C, C++, Fortran, Java, Matlab, Python), and they handle different problems (Hartree–Fock, post Hartree–Fock, DFT, etc.). We concentrate now on codes which can perform (but are not necessarily limited to) DFT. Many other parameters are important when considering the use of these codes, e.g. the user-friendliness, the ability to add new features, the license type, the cost and the possibility to access and modify the source code. In Table 1.1, we present different codes with their corresponding discretization method, together with the language and the license type. Note that this list is by no means exhaustive.

Beyond the computational cost of the different codes, a crucial issue is the reproducibility and the comparison of the results obtained with different codes on the same systems. In this framework, a comparison between the quantum chemistry codes Abinit and Quantum Espresso has been proposed in [206]. More recently, a study involving many more codes has been presented in [164].

1.5.2 Molecular orbitals

A very popular variational method in quantum chemistry relies on atomic orbitals. These basis functions have specific radial and angular dependencies and are centered at the nuclei of each atom of a molecule. Different atomic orbitals are defined for each element of the periodic table. The intuition behind such an approach is that the orbitals can be approximated locally around each nuclei as linear combinations of nuclei-centered functions. Thus, an orbital φ is sought as linear combination of atomic orbitals, in short LCAO, or sometimes LCAO-MO

$$\forall \mathbf{x} \in \mathbb{R}^3, \quad \varphi(\mathbf{x}) = \sum_{\alpha=1}^{M} \sum_{\mu=1}^{N_\alpha} C_\mu^\alpha \, \xi_\mu^\alpha(\mathbf{x} - \mathbf{R}_\alpha), \tag{1.5.1}$$

where the first sum runs over the nuclei, and the second sum runs over the basis functions for each nucleus. Here, the parameters $\mathbf{R}_\alpha$ denote the positions of the nuclei, the C_μ^α are real

Discretization method	Name of the code	Language	License type
Molecular orbitals (GTO)	**Dalton** [1]	Fortran	academic
Molecular orbitals (GTO)	**Gaussian** [112]	Fortran	commercial
Molecular orbitals (GTO)	**MOLPRO** [248]	Fortran	commercial
Molecular orbitals (NAO)	**FHI-Aims** [27]	Fortran	academic and commercial
Planewaves	**Abinit** [121]	Fortran	GPL
Planewaves	**CASTEP** [78]	Fortran	academic and commercial
Planewaves	**Quantum ESPRESSO** [114]	Fortran	GPL
Planewaves	**VASP** [153]	Fortran	academic and commercial
Planewaves	**KSSOLV** [252]	Matlab	free under Matlab license
Wavelets	**BigDFT** [191]	Fortran	GPL
Wavelets	**Madness** [130]	C++	GPL
Finite differences	**Octopus** [188]	Fortran, C	GPL
Finite differences	**PARSEC** [219]	Fortran	GPL
Finite differences	**RMG** [38]	C, C++	GPL
Finite elements	**RealSPACES** [70]		
Finite elements	**DGDFT** [142]		

Table 1.1 – Different quantum chemistry codes using different discretization methods.

coefficients, and the ξ_μ^α are basis functions called atomic orbitals, and will be detailed in the following.

Slater-type orbitals (STO) A first natural idea that came up to design atomic orbitals was to use hydrogen-like orbitals, which are analytically computable. A simplified version thereof is called Slater type orbitals and the basis functions are given by

$$\xi(r,\theta,\varphi) = r^\ell e^{-\frac{Zr}{n}} Y_\ell^m(\theta,\varphi) \qquad \text{(up to normalization)}. \tag{1.5.2}$$

These functions are written in terms of spherical coordinates (r,θ,φ) and have a specific dependency in the radial part, in $r^\ell e^{-\frac{Zr}{n}}$, where $\ell \in \mathbb{N}$, and $\frac{Z}{n}$ is a constant related to the effective charge of the nucleus. The angular dependency is described by means of spherical harmonics which are denoted by Y_ℓ^m, $m \in \mathbb{Z}$, $-\ell \leq m \leq \ell$. Figure 1.3 illustrates some examples of the dependency on the radial part of such basis functions.

The main disadvantage of this Ansatz is that the bielectronic 6-dimensional integrals needed in the resolution of the Hartree–Fock (1.3.4) and Kohn–Sham (1.3.8) problems are expensive to compute without further simplifications. Besides, the number of bielectronic integrals scales as N_c^4, where N_c is the number of basis functions. Therefore, they are not much used in practice in quantum chemistry computer codes.

Gaussian type orbitals (GTO) Gaussian-type basis functions were introduced to circumvent this complexity problem. Indeed, integrals of Gaussians, products thereof and polynomials times Gaussians can be computed analytically, which is not possible for general basis functions. The computational complexity of the 6-dimensional bielectronic integrals is thus reduced to the computation of a one-dimensional integral. Note that the total number of bielectronic integrals still scales as $\mathcal{O}(N_c^4)$, with N_c the number of basis functions, but can be reduced in practice to $\mathcal{O}(N_c^{2.7})$ since the overlap of gaussians corresponding to far-away atoms is negligible.

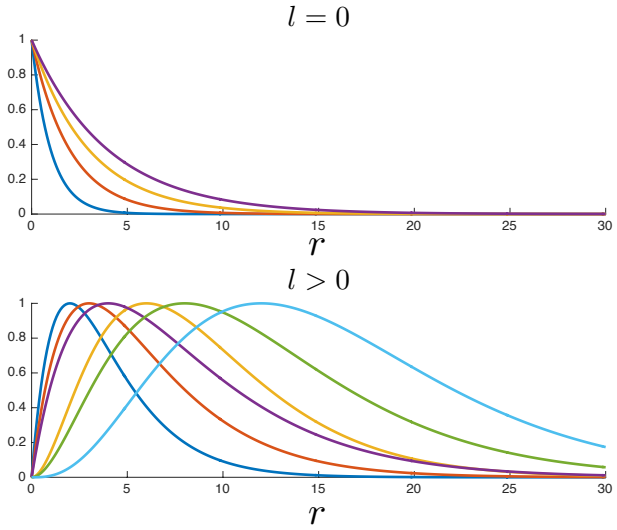

Figure 1.3 – Radial dependency of Slater-type orbitals (STO).

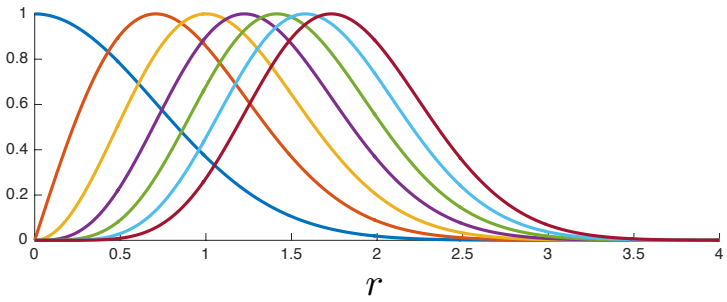

Figure 1.4 – Radial dependency of Gaussian-type orbitals (GTO).

These basis functions are called Gaussian Type Orbitals (GTO), and write in spherical coordinates:

$$\xi(r, \theta, \varphi) = r^{\ell} e^{-\beta r^2} Y_{\ell}^{m}(\theta, \varphi) \qquad \text{(up to normalization)},$$

β being a positive parameter, $\ell \in \mathbb{N}$, $m \in \mathbb{Z}$, $-\ell \leq m \leq \ell$. As for STO, spherical harmonics Y_{ℓ}^{m} handle the angular dependency. The radial part is simply a polynomial times a Gaussian function. Figure 1.4 illustrates some examples of the dependency on the radial part of these Gaussian-type orbitals.

The main disadvantage of these basis functions is that their asymptotic behavior around the origin (called cusp) and towards infinity (called fall-off) is not correct, in contrast to Slater type orbitals. In consequence, many basis functions are needed to get good approximations.

Contracted Gaussians In practice, it was however observed that only a small number of basis functions were needed. Indeed, in general, only a very few particular linear combinations of gaussian-type orbitals are sufficient to obtain good approximations. These linear combinations are called contracted gaussians and can be predefined and computed once and for all. The basis functions are given in cartesian coordinates by:

$$\xi(x, y, z) = \sum_{k=1}^{N_k} C_k \, x^{n_x^k} \, y^{n_y^k} \, z^{n_z^k} e^{-\alpha_k r^2}, \qquad \text{(up to normalization)} \tag{1.5.3}$$

where $r^2 = x^2 + y^2 + z^2$, C_k are real coefficients, $n_x^k, n_y^k, n_z^k \in \mathbb{N}$, and $\alpha_k \in \mathbb{R}^+$. As one can see one basis function is already a linear combination of Gaussian type orbitals.

Contracted Gaussian type orbitals are used in many quantum chemistry codes because of the computational simplicity of bielectronic integrals, combined with a small number of basis functions. They are used in Gaussian [112], and Molpro [248], two of the main commercial codes in computational chemistry. They are also used in academic codes, for example in Dalton [1].

Numeric atom-centered orbitals (NAO) Of course, it is possible to define other atom-centered sets of orbitals, by freely choosing the basis functions, and then compute the orbitals following (1.5.1). This allows to design orbitals with specific properties. A NAO basis set is used for example in the code FHI-Aims [27]. In any case, the basis functions are precomputed, and once the basis set is chosen, the computations are all performed within this basis set.

As the LCAO discretization method is variational, the numerical results can be systematically improved by increasing the basis set size. However, a main drawback is that there is no straightforward way of increasing the basis to reach convergence. One can increase the number of basis functions related to each atom, but it is a priori difficult to know in advance which function to add to the basis. Numerically, this leads to convergence problems, and the accuracy of the computations is limited. This also makes the mathematical analysis of this problem very hard, and up to now, only few publications on the error analysis for atom-centered basis functions have been published [74].

1.5.3 Planewaves

The planewave discretization method is a state-of-the-art method in solid state physics and materials science, and is particularly suited for periodic systems. It can also be used for molecular systems, in the context of the supercell method (see Section 1.4.1). This method is used in many codes, among which Abinit [121], Vasp [153], Quantum Espresso [114], and CASTEP [78].

The simulation domain is denoted by Ω and is equipped with periodic boundary conditions. It can be a cubic box, but it is in general the unit cell of a periodic lattice $\mathcal{R} \subset \mathbb{R}^3$. The orbitals are written as linear combinations of the basis functions $e_{\mathbf{k}}$ defined by

$$\forall \mathbf{r} \in \Omega, \quad e_{\mathbf{k}}(\mathbf{r}) := |\Omega|^{-1/2} e^{i\mathbf{k}\cdot\mathbf{r}},$$

which are the Fourier modes with wavevectors $\mathbf{k} \in \mathcal{R}^*$, where $\mathcal{R}^*$ denotes the dual lattice of $\mathcal{R}$. In this expression, $|\Omega| = \int_\Omega d\mathbf{x}$ is the volume of the unit cell.

The kinetic energy of a basis function $e_{\mathbf{k}}$ is given by $\frac{1}{2}|\mathbf{k}|^2$, where $|\cdot|$ denotes the Euclidean norm. Therefore, in the simulations, an energy cutoff $E_c > 0$ is chosen and the basis set is composed of basis functions with kinetic energy smaller than E_c, i.e. $|\mathbf{k}| \leq \sqrt{2E_c}$. That is, for each cutoff E_c, we set $N_c = \sqrt{\frac{E_c}{2}}\frac{L}{\pi}$ and consider the finite-dimensional discretization space

$$\left\{ \sum_{\mathbf{k}\in\mathcal{R}^*, |\mathbf{k}| \leq \frac{2\pi}{L} N_c} \widehat{v}_{\mathbf{k}}\, e_{\mathbf{k}} \;\middle|\; \forall \mathbf{k}, \quad \widehat{v}_{-\mathbf{k}} = \widehat{v}_{\mathbf{k}}^* \right\}.$$

Then, a Galerkin approximation of the problem is used. The resolution is hence variational, and the approximate ground-state energy is always above the exact energy.

One of the main advantages of this method is that it is systematically refinable. Indeed, the family $(e_{\mathbf{k}})_{\mathbf{k}\in\mathcal{R}^*}$ forms an orthonormal basis of $L^2_\#(\Omega, \mathbb{C})$. To get closer to the exact solution, one only needs to increase the energy cut-off E_c. From a mathematical point of view, it is

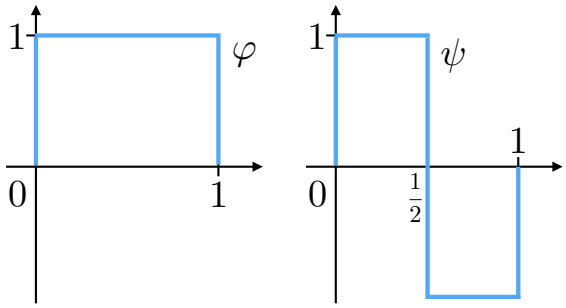

Figure 1.5 – Wavelet example: Haar mother scaling function φ and Haar wavelet ψ.

therefore easier to analyze planewave methods than methods based on molecular orbitals. For example, convergence analysis and error estimations have been proposed for density functional theory inspired problems in [43] and [44]. On top of that, the Laplace operator, present in many electronic structure equations, is diagonal in planewaves. This is very useful, and particularly exploited in the perturbation method proposed in Part V.

As a drawback, the number of basis functions needed to represent correctly the orbitals is way larger than for molecular orbitals. Indeed, it is very difficult to reproduce the nuclear cusps with planewaves, as we need to consider very high kinetic energy cutoffs to get an accurate representation of cusps.

1.5.4 Wavelets

The wavelet method is also a variational method, which was developed more recently. Very popular for signal processing, it is presented in detail in [185]. In the context of partial differential equations, an introduction to the theory of wavelets can be found in [116], and a review article on the use of wavelets in DFT is [22].

Let us start by introducing a wavelet basis on $\mathbb{R}$. It is based on two particular functions: a mother scaling function denoted by ϕ and a mother wavelet denoted by ψ. A simple example of these two functions is the Haar wavelet, shown in Figure 1.5. In the literature, there exist different classes of wavelets having different properties. The electronic structure code BigDFT based on wavelets [191] uses for example Daubechies wavelets [133]. The corresponding mother scaling functions and wavelet are presented in Figure 1.6. The code Madness [130] is also based on wavelets, and uses a multi-wavelet basis, i.e. a wavelet basis based on multiple scaling functions.

The basis functions are obtained by integer translations and dilations of the mother scaling function and the wavelet. At a given resolution level k, the basis functions ϕ_i^k and ψ_i^k are defined by

$$\phi_i^k(x) \propto \phi(2^k x - i),$$
$$\psi_i^k(x) \propto \psi(2^k x - i).$$

With this convention, large values of k correspond to high resolutions, and therefore to large basis sets. At the resolution level k, the considered functions are expanded as a linear combination of basis functions based on both scaling functions and wavelets of different resolutions. More precisely, they are defined as

$$f_k(x) = \sum_{i=0}^{2^k-1} c_i^k \phi_i^k(x) + \sum_{i=0}^{2^k-1} d_i^k \psi_i^k(x),$$

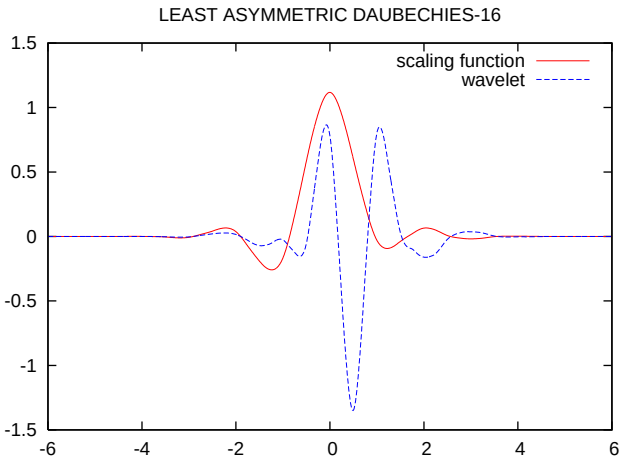

Figure 1.6 – Wavelet example: Daubechies mother scaling function and wavelet used in BigDFT [191].

where c_i^k and d_i^k are real coefficients, or equivalently as

$$f_k(x) = c_0^0\phi_0^0(x) + d_0^0\psi_0^0(x) + \sum_{i=0}^{1} d_i^1\psi_i^1(x) + \sum_{i=0}^{3} d_i^2\psi_i^2(x) + \ldots + \sum_{i=0}^{2^{k-1}-1} d_i^{k-1}\psi_i^{k-1}(x).$$

The two previous expressions span the same space because a scaling function at resolution level k is always a linear combination of a scaling function and a wavelet at the coarser level $k-1$.

For the orbitals, which are 3-dimensional functions, a tensor product of one-dimensional wavelets is used. Therefore, an orbital φ is represented in cartesian coordinates as

$$\varphi(x,y,z) = \sum_{i,j,k} \alpha_{i,j,k} f_{k_x}(x) f_{k_y}(y) f_{k_z}(z),$$

where $\alpha_{i,j,k}$ are real coefficients.

The wavelet method is very interesting for its possible adaptivity. Indeed, the basis set can be improved in two different systematic ways. It is possible to increase the number of grid points in a uniform way, or to increase the resolution level k in given grid points, which can be independently chosen from each other. Therefore, this method appears suited for adaptivity. For example, in the code BigDFT [191], a two-level basis set can be used: it is possible to put in each grid point either one basis function (the scaling function), or eight (the scaling function plus seven wavelets). However, it is not possible to refine the basis at a given resolution level $k > 1$.

Note that the Laplace operator can be diagonally dominant if the wavelet basis is well-chosen, which may lead to nearly sparse representation of the operators, and efficient numerical schemes exploiting this sparsity.

1.5.5 Finite elements

The finite element method is widely used in engineering (see [35] for an introduction). In this method, the simulation domain is meshed, and the basis functions are localized polynomial functions around the vertices of the mesh. They are not necessarily smooth. A classical example is the P_1 method, where the discrete space is composed of continuous functions which are polynomials of degree 1 in each cell of the mesh. The finite element method is called conforming when the finite element space is included in the continuous space in which the solution lies, non-conforming otherwise.

In quantum chemistry, this method has been developed only recently, and in particular in the context of high performance computing. A review on the subject can be found in [198]. As in the planewave method, the number of basis functions necessary to represent the solutions is very high (up to 10^5 per atom in the case of uniform meshes). But it is systematically refinable as it suffices to refine the mesh to get a more accurate solution. Moreover, the mesh can be refined in a non-uniform way, especially close to the nuclei, to capture the singularities. This is particularly used for all-electron calculations.

The locality of the basis gives a chance to deal with very large systems, by decoupling the problem in different parts in space, and solving the different parts in parallel on different computers. This method is also capable of dealing with a large variety of boundary conditions, such as Dirichlet boundary conditions and periodic boundary conditions.

In this domain, many advances have been performed by Tsuchida, see e.g. [236]. A DFT code based on discontinuous Galerkin finite elements has been developed in DGDFT [142]. In this code, adaptive local basis functions are built at each step of the iterative algorithm used to solve the problem, leading to an accurate representation of the Kohn–Sham orbitals with a relatively small number of degrees of freedom. Note that this code is also massively parallel. Additionally, another finite element method based on higher order finite elements has been proposed in [193]. Finally, an h-p refinement procedure for DFT problems can be found in [187].

1.5.6 Finite differences

The finite difference method is based on Taylor expansions of the solution for computing its derivatives. It is therefore quite different from all the previous discretizations based on Galerkin discretizations. More precisely, the solution is represented on a grid. Then the derivatives at stake in the considered equation are computed by finite differences on the grid. To find the ground state, the Hamiltonian matrix is then built and diagonalized. An interesting feature of this method is the sparsity of Hamiltonian matrix, which allows to use efficient sparse-matrix eigensolvers and parallel architectures.

As this method is not variational, there is no guarantee that the exact energy is below the computed energy. However, this method is systematically improvable by using finer and finer grids. There exists different codes in quantum chemistry using finite differences, e.g. Parsec [219] and Octopus [188]. Let us also mention [125] on finite differences for electronic structure calculations.

1.6 Algorithms for mean-field models

To compute numerically the ground state of mean-field models in a discrete basis, there exist mainly two types of methods. The first one consists of solving the Euler–Lagrange equations associated to the problem, which are nonlinear eigenvalue problems, and read respectively (1.3.4) and (1.3.8) for the Hartree–Fock and Kohn–Sham models. These methods are presented in Section 1.6.1. The second type of methods consists of minimizing the energy directly, i.e. dealing with the constrained minimization problems (1.3.2) and (1.3.7). They are presented in Section 1.6.2.

1.6.1 Self-Consistent Field algorithms

We consider a Galerkin discretization of the Hartree–Fock and Kohn–Sham problems (1.3.4) and (1.3.8). In both cases, the orbitals $(\varphi_j)_{1\leq j\leq N}$ solutions to the problem are written as linear

combinations of basis functions $(\chi_\mu)_{1\leq\mu\leq\mathcal{N}}$:

$$\varphi_j = \sum_{\mu=1}^{\mathcal{N}} C_{\mu j}\chi_\mu,$$

where $\mathcal{N}$ is the basis set size. The discrete equations can be written in this framework as

$$\begin{cases} F(D)\,C = SCE \\ C^T SC = I_N \\ D = CC^T, \end{cases} \tag{1.6.1}$$

where $C \in \mathbb{R}^{\mathcal{N}\times N}$ is the coefficient matrix, $S \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}$ is the overlap matrix of the basis functions and has coefficients $S_{\mu\nu} = \langle\chi_\mu|\chi_\nu\rangle$, and E is the diagonal matrix of eigenvalues ε_j. The second equation corresponds to the normality constraint. The matrix D defined in the third equation is called the density matrix. The matrix $F(D)$ has elements $F(D)_{\mu\nu} = \langle\chi_\mu|\mathcal{F}_\Phi\chi_\nu\rangle$, where Φ is constructed from the density matrix D, and the definition of $\mathcal{F}$ depends on the problem under consideration. Note that the expression of the matrix $F(D)$, called Fock matrix in the Hartree–Fock case is the only difference between the Hartree–Fock and Kohn–Sham problems. However, the analysis for the algorithms are performed in many cases for the Hartree–Fock case only.

Since the matrix $F(D)$ depends on the density matrix D itself constructed from the unknown coefficients C, this problem is a nonlinear eigenvalue problem. Such problems are usually solved by means of iterative algorithms, and in particular fixed-point algorithms, also called self-consistent field (SCF) algorithms in the physics and chemistry literature. SCF algorithms consist of solving a at each iteration a linear eigenvalue problem.

Roothan algorithm

The simplest SCF algorithm is called Roothan algorithm [217]. Starting from an initial guess of the coefficient matrix $C^{[0]}$ verifying the constraint $(C^{[0]})^T SC^{[0]} = I_N$ and corresponding density matrix $D^{[0]} = C^{[0]}(C^{[0]})^T$, this algorithm consists of solving at each iteration the following linear eigenvalue problem, where the nonlinearity in the matrix $F(D)$ is frozen with the density matrix computed at the previous iteration

$$\begin{cases} F(D^{[k-1]})\,C^{[k]} = SC^{[k]}E^{[k]} \\ (C^{[k]})^T SC^{[k]} = I_N \\ D^{[k]} = C^{[k]}(C^{[k]})^T. \end{cases} \tag{1.6.2}$$

The algorithm stops when two consecutive iterates of the density matrix are close enough, i.e. given a tolerance ε, if

$$\|D^{[k]} - D^{[k-1]}\| \leq \varepsilon.$$

The stopping criterion can also be chose on the coefficients, in which case it reads

$$\|C^{[k]} - C^{[k-1]}\| \leq \varepsilon.$$

Unfortunately, this algorithm does not converge when the basis set is large. Therefore, the precision of the method is very limited. To overcome this convergence problem, alternative SCF algorithms have been proposed, and we present three of them below. For these algorithms, the stopping criterion is identical to the Roothan algorithm, therefore we do not repeat it.

Level-shifting algorithm

In the level-shifting algorithm [220], a shift parameter $b \geq 0$ is chosen. One starts from an initial guess of the coefficient matrix $C^{[0]}$ verifying the constraint $(C^{[0]})^T S C^{[0]} = I_N$ with density matrix $D^{[0]}$. At each iteration of the algorithm, one solves the following linear eigenvalue problem:

$$\begin{cases} \left(F(D^{[k-1]}) - bD^{[k-1]}\right) C^{[k]} = SC^{[k]}E^{[k]} \\ (C^{[k]})^T S C^{[k]} = I_N \\ D^{[k]} = C^{[k]}(C^{[k]})^T. \end{cases} \tag{1.6.3}$$

Compared to the Roothan algorithm, the global convergence of the algorithm is guaranteed in the Hartree–Fock case when the shift parameter is large enough [55]. But it offers only a slow speed of convergence.

Damping algorithm

To increase the convergence speed of the Roothan algorithm, another algorithm, called damping algorithm [254], has been proposed. It consists, at each iteration, of taking a linear combination of previous iterates for the density matrix $D^{[k]}$. More precisely, starting from an initial guess $C^{[0]}$ verifying the constraint $(C^{[0]})^T S C^{[0]} = I_N$ with density matrix $D^{[0]}$, one solves at each step the following problem

$$\begin{cases} F(D^{[k-1]})\, C^{[k]} = SC^{[k]}E^{[k]} \\ (C^{[k]})^T S C^{[k]} = I_N \\ \tilde{D}^{[k]} = C^{[k]}(C^{[k]})^T \end{cases} \tag{1.6.4}$$

and

$$D^{[k]} = \alpha \tilde{D}^{[k]} + (1-\alpha) D^{[k-1]},$$

where the damping parameter α is a chosen in $[0, 1]$. Note that it is also possible to mix the matrix $F(D^{[k]})$ or the eigenvectors $C^{[k]}$ instead of the density matrix. An analysis of this algorithm can be found in [55].

Direct Inversion in the Iterative Subspace (DIIS) algorithm

Generalizing the damping algorithm, the DIIS algorithm also called Pulay mixing [208] accelerates the convergence of the SCF algorithm, even though it does not always converges to a global minimum of the energy. Instead of considering a linear combination with the previous iterate only, the DIIS algorithm mixes at each iteration m previous density matrices, minimizing the least square of the residuals to determine the coefficients in the linear combination. Starting from an inital guess $C^{[0]}$ verifying the constraint $(C^{[0]})^T S C^{[0]} = I_N$ with density matrix $D^{[0]}$, one solves at each step the following problem

$$\begin{cases} F(D^{[k-1]})\, C^{[k]} = SC^{[k]}E^{[k]} \\ (C^{[k]})^T S C^{[k]} = I_N \\ \tilde{D}^{[k]} = C^{[k]}(C^{[k]})^T \end{cases} \tag{1.6.5}$$

and

$$D^{[k]} = c_0 \tilde{D}^{[k]} + \sum_{i=1}^{m-1} c_i D^{[k-i]},$$

where the c_i are such that $\sum_{i=0}^{m-1} c_i = 1$ and minimize the norm

$$\left\| c_0[F(D^{[k]}), D^{[k]}] + \sum_{i=1}^{m-1} c_i[F(D^{[i]}), D^{[i]}] \right\|^2 ,$$

where $[\cdot, \cdot]$ denotes the commutator defined by $[A, B] = AB - BA$. Note that $[F(D), D] = 0$ corresponds to the equations (1.6.1). Once again, it is possible to mix the matrix $F(D^{[k]})$ or coefficient matrices $C^{[k]}$ instead of the density matrices.

This algorithm converges very fast when it converges, which is not always the case in practice. It is still very popular and used in almost all computational codes. An analysis of this method has been provided in [215], see also [54].

1.6.2 Direct minimization methods

Different methods deal with the energy minimization problem directly. Some of these methods are based on gradient methods, other on Newton-type methods. Let us mention some of the latter type. The Bacskay algorithm [16] introduced in 1981 consists in reformulating the constrained minimization problem in the form of an unconstrained minimization problem, so that a standard Newton algorithm can be applied on the unconstrained problem. Some further developments of this algorithm replaced the Newton method by a quasi-Newton method, as in [108, 68]. Another type of methods called trust-region consists of minimizing at each step a quadratic function within a ball, whose radius changes during the iterative process. Such a method is presented for the Hartree–Fock case in [111].

All these methods converge toward a local minimum, but they rarely converge to the global minimum of the problem. We do not detail these methods, as they will not be used in the other parts of the thesis. We refer to [45] for more details.

Chapter 2

Error analysis for linear and nonlinear eigenvalue problems

2.1 Introduction

2.1.1 Importance of error control

When physical phenomena are modeled with partial differential equations, analytical solutions are barely available. For example, in electronic structure calculation, only the ground state of one-electron systems can be computed exactly. Therefore, for real-case systems, many approximations are resorted to, including modeling, discretization, resolution algorithms, as presented in Chapter 1.

The main objective of error control is to quantify the impact of these different approximations on the final result, in order to evaluate the difference between the computed values of the quantities of interest, and their exact values. In quantum chemistry, these quantities of interest include ground-state energies of molecules, dissociation energies, electronic densities and band gaps. A bound on the error allows then to guaranty the results up to a given tolerance.

Error control can also be used to optimize the simulation parameters to achieve efficient calculations, by means of adaptive algorithms. For instance, when the chosen model is very coarse, it is useless to perform calculations using a very fine discretization, because the precision of the results is then limited by the model error, whatever the basis set size. Determining the error arising from each approximation allows to detect which approximation leads to the largest error. It is then possible to refine only the corresponding parameter to decrease the overall error. In order to obtain the best ratio between computational cost and efficiency, a typical strategy is to balance the errors coming from the different approximations.

To be profitable, an error bound should have different properties. First, it should be fully computable, so that the true error can really be estimated. Second, it should be fully guaranteed, i.e. the error on the quantity of interest in a certain norm should be bounded by this error bound. Otherwise, the bound can only be used as an error indicator. In some cases, the bound is valid under some assumptions, e.g. smallness assumptions, and this can restrain the applicability of the error bound. But if these assumptions are *a posteriori* verifiable, it is possible to guaranty that the regime in which the bound applies is reached. Third, the error bound should be sharp, that is close to the real error. Optimally, the ratio between the true error and the bound should be close to one, or at least converge to one when the error goes to zero. This is called efficiency. Fourth, the error bound should be cheap to compute. Indeed, an error bound is in practice useless if the cost of its computation is of same order or higher than solving the problem itself. Fifth, the bound should allow for error separation.

Indeed, adaptivity is only possible when the error bound can be separated into components depending on the different approximation parameters. In the context of finite elements, a bound is particularly useful when it can be separated in local components, so that mesh refinement can be performed non uniformly. Of course, it is not always possible to gather all these properties, and sometimes, one needs to compromise.

Different methods have been developed to control the error. The *a priori* analysis usually offers convergence guaranty and convergence rates of the approximate solutions to the exact one, but it cannot be used to derive proper error estimates. The *a posteriori* analysis is one of the main tools developed to derive error bounds. It includes residual-based *a posteriori* analysis, where the error is bounded by the residuals of the approximate solutions. Error estimates can also be based on cheap refinements of the approximate solutions, which is the idea of post-processing methods.

At the moment, quantum chemistry codes do not generally include error bounds, and especially no guaranteed bounds. These codes could therefore largely benefit from developments on error control. But more importantly, the computations in this domain can be extremely expensive, especially for large systems. Hence, error balance could be used to reduce the computational cost of quantum chemistry computations.

As presented in Chapter 1, we focus in this thesis on ground state calculations. In particular, we aim at considering models requiring to solve nonlinear eigenvalue problems, e.g. the Kohn–Sham model presented in Section 1.3.3. The error estimation for such problems encloses many aspects, such as the chosen discretization method, the chosen possibly iterative algorithm, the number of eigenvalues to compute, and the nonlinearity under consideration. We progressively deal with these aspects, considering first linear eigenvalue problems, and then nonlinear eigenvalue problems.

In this chapter, we provide an introduction to error control together with a presentation of the main contributions of this thesis. We first describe the different natures of error arising in numerical simulations and in particular in electronic structure calculation in Section 2.1.2. We then present the state of the art regarding the error analysis for linear and nonlinear eigenvalue problems and show how the contributions of this thesis fall in this framework in the two following subsections. We focus on *a priori* analysis in Section 2.1.3 and on *a posteriori* analysis in Section 2.1.4. In the rest of the chapter, we present in more detail the different contributions of the thesis. In Section 2.2, we focus on an *a priori* estimation of the discretization error cancellation between two different nuclear configurations of a system. In Section 2.3, we present our contribution for the *a posteriori* analysis of the Laplace eigenvalue problem, in giving first a presentation of the problem in Section 2.3.1 and presenting the results obtained in the conforming case in Section 2.3.2, and in the nonconforming case in Section 2.3.3. In Section 2.4, we present our contribution to the *a posteriori* error analysis as well as error balance for a nonlinear eigenvalue problem, namely the Gross–Pitaevskii equation. Finally, in Section 2.4, we concentrate on post-processing methods for the Kohn–Sham model. As our contribution in this direction is based on perturbation theory, we present this theory in Section 2.5.1. We then describe how we apply it to linear Schrödinger operators in Section 2.5.2, and to Kohn–Sham equations in Section 2.5.3. In Section 2.6, we present different existing post-processing methods for linear and nonlinear eigenvalue problems and compare them in a unified framework. In Section 2.7, we propose some perspectives of this work.

2.1.2 An error arising from different sources

In the context of molecular simulation, many approximations are employed to compute the value of physical observables. These approximations cause the observed errors between the computed values and their experimental (exact) values. However, a large part of these appro-

ximations is not limited to quantum chemistry, and runs more generally in the context of partial differential equations modeling physical phenomena. We describe them in the following, see also the introduction of Chapter 3.

Model error

The *model error* encloses the difference between experimental results, or the results of a reference model considered fine enough to be exact, and the solution of the chosen model. The chosen model can for example neglect higher order phenomena, or nonlinearities. It can also arise from a dimensionality reduction, as the Hartree–Fock model derived from the Schrödinger model. In electronic structure calculation, the Schrödinger model is very often seen as a reference model, when relativistic effects are small (see Section 1.2.3). Unfortunately, we do not take the model error into account in the error estimations of this thesis. Therefore, the considered models, which are defined by partial differential equations over continuous spaces, are always considered as reference models.

Discretization error

The *discretization error* corresponds to the error between the exact solution of the chosen model and a discrete solution of this particular model. It corresponds to a simplification from infinite-dimensional to finite-dimensional equations. Different discretization methods were presented in Section 1.5. In fact, some estimations of the discretization error have already been provided for eigenvalue problems, some of which will be presented in Sections 2.1.3 and 2.1.4. Most of them handle systematically improvable discretization methods, such as planewaves or finite elements. In this thesis, we mainly focus on the discretization error. In all Chapters 3 to 9, we consider at least the discretization error in the error estimations.

Algorithm error

To solve the discretized equations arising in mean-field models, such as the Hartree–Fock and Kohn–Sham models presented in Sections 1.3.1 and 1.3.3, iterative algorithms presented in Section 1.6 are used. In the iterative procedure, the number of iterations performed is always finite. The *algorithm error* therefore corresponds to the difference between the exact solution of the discrete equations and the computed solution after a given number of iterations. The algorithm error is in particular handled in Chapter 6 for the Gross–Pitaevskii equation. Note that, for some problems and especially nonconvex problems, the algorithm might not converge, so that the difference between two iterates in the algorithm could become very small, with both iterates far from the exact solution.

An algorithm error can also arise from linear algebra problems. Indeed, these problems are often solved with iterative solvers, hence never exactly. When the problems are solved up to machine precision, this error is neglected. But when the solver tolerance can be chosen, this linear algebra error can be taken into account in the error balance in order to not fully converge the linear resolution, while keeping the same overall precision. This is proposed for the Laplace eigenvalue problem in the context of conforming finite elements in Chapter 4.

Implementation error

The *implementation error* can be due but is not limited to bugs. Round-off errors and numerical integration errors can also be present in the computations. They are supposed to be small but can accumulate, especially for large systems. In this thesis, we do not take this error into account.

	A priori analysis	*A posteriori* analysis and adaptive computations
Linear eigenvalue problems including $-\Delta u = \lambda u$ Section 2.3	[14], [85] [28] also [69], [146]	FE: [14, 19, 178, 224, 225] [17, 18, 179] [174]
Gross–Pitaevskii equation $-\Delta u + Vu + u^3 = \lambda u$ Section 2.4	PW, FE: [43] FE: [73]	PW: [98] FE: [80]
Kohn–Sham/Hartree–Fock equations $-\Delta\Phi + V_\rho\Phi = \Lambda\Phi$ Section 2.5	PW: [44] FE: [71]	FE: [70, 172]

Table 2.1 – Table of some contributions for *a priori* and *a posteriori* analysis of eigenvalue problems

Computer error

The *computer error* is due random hardware failures (miswritten or misread bits). It is usually neglected but is expected to have a greater importance in the futur, especially with the development of exascale architectures [168]. In this thesis, we consider this error to be always negligible.

2.1.3 *A priori* analysis for eigenvalue problems

A priori analysis aims at showing the quality of the numerical solutions of a given problem. Let us denote by u the exact solution of a partial differential equation, and by u_h its variational approximation in a given basis. *A priori* estimates take the form

$$|||u - u_h||| \leq Ch^k,$$

where $k > 0$ and $C > 0$ are constants, $|||\cdot|||$ is a given norm, and h is a discretization parameter, typically the size of the mesh in finite elements, or the inverse of the cutoff in energy or momentum space in planewaves. Such an estimation guaranties that the error goes to zero as h goes to zero, at the order k. This hence justifies the quality of the numerical method. However, the constant C depends on the exact solution u, i.e. $C = C(u)$, and is in practice uncomputable or highly overestimated. The error $|||u - u_h|||$ is therefore not bounded by a *computable* constant.

The *a priori* analysis of partial differential equations has been studied for a long time. Already in 1964, Céa proposed *a priori* error estimates for elliptic partial differential equations [67], including the Laplace boundary problem. For linear eigenvalue problems, *a priori* estimates are provided in [85, 14] for finite elements, and a review article can be found in [28], see also [69, 146]. Note that the solutions of the Laplace eigenvalue problem are explicitly known in planewaves, hence there is no need for error estimation in this case.

In this thesis, we are particularly interested in nonlinear eigenvalue problems in the context of electronic structure calculation. In this domain, *a priori* results are more recent. An analysis for a class of nonlinear eigenvalue problems, which can be seen as toy models for density functional theory discretized with finite elements or planewaves has been provided in [43]. For the orbital-free Thomas–Fermi–von Weizäcker model, *a priori* estimates can be found in [255, 73]. Moreover, the Thomas–Fermi–Dirac–von Weizäcker has been studied in [161]. For the Kohn–Sham model, the *a priori* analysis has been performed few years ago in [44], and was

later extended to the finite element case in [71]. As a summary, the left column of Table 2.1 presents these *a priori* references grouped by equation type.

Since the *a priori* error bounds are not fully computable, they do not provide guaranteed error bounds. However, they can still be seen as error indicators, and used for adaptivity. For example, knowing the convergence rate relative to two different approximation parameters, one can adapt the values of the parameters so that the convergence speed is optimal. This is for example the spirit of h-p refinement [124], where the solution can be refined both by using finer elements, or polynomials with higher degrees. This method is developed in the context of electronic structure calculation in [187], see also [182].

In Chapter 3, our study is based on an *a priori* analysis of the energy error of the problem. However, the purpose is not to provide error estimates *per se*, but to show that the phenomenon of error cancellation observed between different molecular configurations is not due to the improvement of the convergence rate, but lies in the prefactor.

2.1.4 *A posteriori* analysis for eigenvalue problems

Residual-based *a posteriori* analysis

Unlike the *a priori* analysis, the goal of the *a posteriori* analysis is to provide a computable bound on the error between the numerical approximation of the solution and the unknown exact solution. This bound should therefore be computable with the only knowledge of the approximate solution, and the parameters used for the computation of the approximate solution. In the general case, it reads

$$|||u - u_h||| \leq F(h, u_h),$$

where $|||\cdot|||$ is a given norm, and $F(h, u_h)$ is a fully computable quantity.

Many *a posteriori* estimations are based on the computation of the residual, which represents the error of the solution with respect to the equation. In the case of an abstract nonlinear eigenvalue problem reading: find (u, λ) such that

$$A(u)u = \lambda u,$$

the residual $\mathrm{Res}(u_h, \lambda_h)$ relative to an approximate solution (u_h, λ_h) is defined by

$$\mathrm{Res}(u_h, \lambda_h) = A(u_h)u_h - \lambda_h u_h.$$

Note that the residual of the exact solution (u, λ) is equal to zero. The main idea of residual-based *a posteriori* error estimation is to bound the error on the solution in a given norm by the dual norm of the residual in an appropriate Banach space. Note that if this dual norm is not well chosen, the norm of the residual might not even go to zero when the error goes to zero. An important difficulty is then to compute or estimate the dual norm of the residual in an accurate and efficient way.

One of the key results in *a posteriori* analysis called equilibrated fluxed has been developed by Prager and Synge in 1947 [207]. Since then, *a posteriori* analysis has been an active research area. *A posteriori* estimation of elliptic problems can for example be found in the two books [240, 4].

As a reference book for eigenvalue problems, we refer to [14]. Note that Galerkin discretization allows to obtain upper bounds on the eigenvalues very easily. Indeed, the Rayleigh–Ritz method guaranties that the computed eigenvalues are above the exact ones. However, lower bounds on the eigenvalues are not easy to obtain, and have been investigated for a long time. Already in 1928, Temple provided lower and upper bounds for the frequency of the gravest mode of a vibrating system [232]. Then, different methods have been developed, among which

Kato [145], Aronszajn [12], Weinstein [247]. In the context of quantum chemistry problems, Bazley proposed a study for the helium atom [17] in 1960, and Bazley–Fox [18] and Löwdin [179] proposed lower and upper bounds for the eigenvalues of linear eigenvalue problems. A clear introduction to these historical methods can be found in [205]. Note that these methods often include only bounds on the eigenvalues and not on the eigenfunctions.

More recently, there has been many publications related to the *a posteriori* estimation for eigenvalue problems, including [19, 180, 140, 141, 64, 253, 176, 159, 224, 225, 178, 174]. We refer to the introduction of Chapter 4 for more detail on the bibliography in the conforming finite element case, and to the introduction of Chapter 5 for nonconforming finite elements.

For nonlinear eigenvalue problems, there exist only few *a posteriori* estimations. In [70], the authors provide a residual-based *a posteriori* estimation for the Kohn–Sham model discretized with finite elements. The estimator is not fully guaranteed, but the authors use it as an error indicator to refine the mesh, and therefore obtain an accurate solution with a reasonable number of degrees of freedom. An *a posteriori* analysis is performed in [147] for a linear eigenvalue problem, which is in turn numerically applied for the nonlinear Kohn–Sham problem in the context of non-polynomial basis functions in a discontinous Galerkin framework. The Hartree–Fock problem was studied in [184], where the authors provide an *a posteriori* estimation based on a post-processing of the solution. Finally, in Chapter 6, we provide a guaranteed bound of the eigenfunction error for the Gross–Pitaevskii equation.

Post-processing methods

Post-processing methods can sometimes be used for *a posteriori* estimations. The main idea is to solve the discrete problem in a coarse grid, i.e. with a small basis set, and then to do a not-so-expensive computation in a fine grid, in order to get a more precise result than in the coarse grid, at a lower price than computing the solution on the fine grid. In the framework of eigenvalue problems, this post-processing step can be for example the resolution of a boundary problem, or the resolution of local problems. If the post-processed solution is accurate enough, it can be used as a reference solution and the error is then estimated by the difference between the post-processed solution and the approximate solution. In many cases, this error bound is not guaranteed, but it can still be used in practice as an error indicator. Let us remark that the first goal of these methods is often not to provide error bounds, but rather to offer accurate solutions at a low computational cost.

Such a method called two-grid method has been proposed in [251] for a linear eigenvalue problem. The full problem is solved on the coarse grid, and a boundary value problem is solved on the fine grid. This method has been extended to a class of nonlinear eigenvalue problems in [42]. In the latter, one can choose between three different problems on the fine grid: two boundary value problems, and one linear eigenvalue problem. As previously mentioned, a post-processing method for the Hartree–Fock problem can be found in [184]. We also present a post-processing method based on perturbation theory in Part V. Kohn already proposed a post-processing method based on perturbation theory in [151], which improves numerically the eigenfunctions of a linear eigenvalue problem locally. In [96], we describe this method in detail and compare it to the perturbation theory of Part V. All these post-processing methods will be compared in a unified framework in Section 2.6.

Error balance

As shown in Section 2.1.2, the overall error comes from different sources. In order to reduce the computational cost of the simulations, a possible approach is to separate the error bound obtained for example from a residual-based *a posteriori* analysis into several components, each

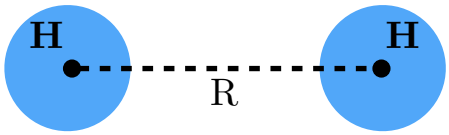

Figure 2.1 – Configurations of an H_2 molecule indexed by the parameter R.

of them depending only on one approximation parameter. In finite elements, this consists of writing the total error as a sum of local components, and to refine only the elements with large error components. A standard procedure for finite element refinement is called Dörfler marking [92]. Adaptive finite element procedures are very popular, and several have been proposed for electronic structure calculation [70, 173]. Another adaptive strategy for the Kohn–Sham problem can be found in [147].

It is also possible to incorporate different approximations and estimate each of them, in order to perform error balance. Some previous works take into account the errors due to the discretization and the algorithm used in the linear algebra resolution. This is for example proposed in [144] in the case of a diffusion problem. Another separation between the discretization and the iteration error is performed in [210] for a linear eigenvalue problem discretized with finite elements. Such separation between linear algebra and discretization error is presented for the Laplace eigenvalue problem in Chapter 4 in the case of conforming finite element methods.

Another possibility is to take into account the discretization error and the algorithm error arising in nonlinear problems, i.e. the linearization error in Newton or self-consistent schemes. This was performed in the case of a nonlinear diffusion-type problem in [101, 103]. In Chapter 6, we present a separation and balance of the error coming from the discretization and the iterative algorithm used to solve the nonlinear discrete Gross–Pitaevksii equation.

Let us remark that the model error is in general very hard to tackle. In this direction, no previous work is known to the author in the context of electronic structure calculations. In the context of the simulation for crystalline solids with defects, an adaptive simulation separating the model error from the discretization error for the atomistic/continuum coupling is provided in [244].

2.2 Error cancellation for the discretization error

In the previous presentation of error analysis, we have focused on the error estimation for a given problem. This corresponds in the quantum chemistry context to estimate the error for a given configuration of the nuclei. However, in many cases, for example in molecular dynamics, the difference between quantities of interest, which depend on molecular configurations, are very important. And if there is an error cancellation between these quantities, less accurate computations could be adequate to get overall accurate differences. In this section, we introduce the contribution presented in Chapter 3, where we study is a discretization error cancellation between different atomic configurations.

In the following, the configurations are indexed by a parameter R, which accounts for the geometry of the molecule. In the case of the H_2 molecule, R corresponds to the distance between the two hydrogen atoms (see Figure 2.1). For each configuration, it is possible to discretize the problem (e.g. Kohn–Sham problem) in a chosen finite basis characterized by a parameter N, typically the number of basis functions. What is expected is that if we consider two configurations R_1 and R_2 with corresponding ground-state exact energies E_{R_1} and E_{R_2},

and approximate energies $E_{R_1,N}$, $E_{R_2,N}$, there holds

$$|(E_{R_1,N} - E_{R_2,N}) - (E_{R_1} - E_{R_2})| \ll |E_{R_1,N} - E_{R_1}| + |E_{R_2,N} - E_{R_2}|,$$

that is the difference of the energy errors is way smaller than the energy errors themselves. In this contribution, we aim at characterizing this improvement.

First, we perform some simulations on simple molecular systems to assess this affirmation. We therefore compute the ground state of the H_2 molecule for different atomic positions using a planewave discretization method with the code Abinit [121] based on DFT. We also simulate a system composed of four hydrogen atoms and two oxygen atoms. We observe in both cases an improvement on the energy error, of a factor from 10 to 50 (presented in Figures 3.1 and 3.3). This improvement seems constant over the dimension of the planewave space, which suggests that the convergence rate of the energy difference is not better than the convergence rate of the energy itself, so that only the prefactor is smaller for the energy difference.

To justify this observation theoretically, we study a toy model, for which analytical solutions are available. Therefore, for two given parameters $z_1, z_2 > 0$, we consider the family of problems, indexed by $R \in (0,1)$, consisting in finding the ground state $(u_R, E_R) \in H^1_{\mathrm{per}} \times \mathbb{R}$ of

$$\begin{cases} \left(-\dfrac{d^2}{dx^2} - \displaystyle\sum_{m\in\mathbb{Z}} z_1\delta_m - \sum_{m\in\mathbb{Z}} z_2\delta_{m+R}\right) u_R = E_R u_R, \\ \displaystyle\int_0^1 u_R^2(x)dx = 1, \quad u_R \geq 0, \end{cases}$$

where δ_a denotes the Dirac mass at point $a \in \mathbb{R}$. For this specific problem, we can demonstrate an explicit *a priori* analysis for the energy when the problem is discretized with planewaves, showing that the error cancellation factor defined as

$$Q_N = \frac{|(E_{R_1,N} - E_{R_2,N}) - (E_{R_1} - E_{R_2})|}{(E_{R_1,N} - E_{R_1}) + (E_{R_2,N} - E_{R_2})}$$

converges to a fixed number $0 < Q_\infty < 1$ when N, the cut-off in momentum space in the planewave discretization, goes to infinity. This is presented in the following theorem, see also Theorem 3.3.2 for a detailed version.

Theorem (Asymptotic expressions of the energy error and of the error cancellation factor)**.** *For all $z_1, z_2 > 0$ and $R \in (0,1)$, we have for all $\epsilon > 0$,*

$$E_{R,N} - E_R = \frac{\alpha_R}{N} - \frac{\alpha_R}{2N^2} + \frac{\beta^{(1)}_{R,N}}{N} + \frac{\gamma_R}{N}\eta_{R,N} + o\left(\frac{1}{N^{3-\epsilon}}\right), \tag{2.2.1}$$

where $\alpha_R := \frac{z_1^2 u_R(0)^2 + z_2^2 u_R(R)^2}{2\pi^2}$, and $\gamma_R, \eta_{R,N}, \beta^{(1)}_{R,N}$ can be explicitely determined and estimated.

Moreover, $-\frac{\alpha_R}{2N^2} + \frac{\beta^{(1)}_{R,N}}{N} + \frac{\gamma_R}{N}\eta_{R,N}$ is of higher order than $\frac{\alpha_R}{N}$, and as a consequence, we have for all $z_1, z_2 > 0$ and all $R_1, R_2 \in (0,1)$,

$$\lim_{N\to+\infty} Q_N = \frac{|\alpha_{R_1} - \alpha_{R_2}|}{\alpha_{R_1} + \alpha_{R_2}} = \frac{\left|z_1^2\left(u_{R_1}(0)^2 - u_{R_2}(0)^2\right) + z_2^2\left(u_{R_1}(R_1)^2 - u_{R_2}(R_2)^2\right)\right|}{z_1^2(u_{R_1}(0)^2 + u_{R_2}(0)^2) + z_2^2(u_{R_1}(R_1)^2 + u_{R_2}(R_2)^2)}. \tag{2.2.2}$$

Note that the limit of Q_N depends only on z_1, z_2 and on $u_{R_1}(0)^2$, $u_{R_2}(0)^2$, $u_{R_1}(R_1)^2$, $u_{R_1}(R_2)^2$, i.e. on the values of the densities $\rho_{R_1} = u_{R_1}^2$ and $\rho_{R_1} = u_{R_2}^2$ at the singularities of the potential. This confirms that for this specific toy model, the energy error cancellation occurs not in the convergence rate, but only in the prefactor, which is smaller for the energy difference than for the energies themselves.

2.3 A posteriori error estimation for the Laplace eigenvalue problem

2.3.1 Presentation of the problem

In this section, we present the contributions of Chapters 4 and 5, which focus on the *a posteriori* analysis of a linear eigenvalue problem: the Laplace eigenvalue problem. We consider a domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ being a polygonal/polyhedral domain with a Lipschitz boundary. We denote by λ_i, u_i be the eigenvalues and associated eigenvectors of the Laplace operator $-\Delta$ on Ω with Dirichlet boundary conditions. The problem reads: find eigenvector and eigenvalue pairs (u_k, λ_k), with u_k satisfying a homogeneous Dirichlet boundary condition over $\partial\Omega$ and subject to the constraint $\|u_k\| = 1$, where $\|v\|^2 := \int_\Omega v^2 \,\mathrm{d}\mathbf{x}$ such that $-\Delta u_k = \lambda_k u_k$ in Ω. In a weak form, $(u_k, \lambda_k) \in V \times \mathbb{R}^+$ with $\|u_k\| = 1$ and

$$(\nabla u_k, \nabla v) = \lambda_k (u_k, v) \qquad \forall v \in V. \tag{2.3.1}$$

The problem is discretized using finite elements, and the aim of this contribution is to provide *a posteriori* bounds on both an arbitrary simple Laplace eigenvalue and the associated eigenvector for conforming and nonconforming methods. We focus on the discretization error, but in the conforming case, we also account for inexact solvers in the resolution of the problem.

The provided bounds are guaranteed, fully computable, and converge with optimal speed respectively to the given exact eigenvalue and eigenvector. Moreover, the only hypothesis needed is that the approximate eigenvalue is separated from the following smaller and larger ones, i.e.

$$\lambda_{i-1} < \lambda_{ih} < \lambda_{i+1}.$$

This can be checked in practice, by finding respectively upper and lower bounds of the exact eigenvalues λ_{i-1} and λ_{i+1}.

Note that in the estimations, there appears no unknown (solution-, regularity-, or polynomial-degree-dependent) constant, and no convexity/regularity assumption on the computational domain/exact eigenvector(s) is needed. The computation of the bounds only requires the resolution of local Neumann problems, hence is very cheap compared to the full computation of the solution. Therefore, these error estimates seem to fulfill the properties presented in Section 2.1.1.

In the conforming and the nonconforming cases, we use a similar approach. As is very common in *a posteriori* analysis, our estimation relies on the use of the residual defined in the conforming case as follows.

Definition 2.3.1 (Residual and its dual norm)**.** *Let V' stand for the dual of V. For any pair $(u_{ih}, \lambda_{ih}) \in V \times \mathbb{R}$, define the residual* $\mathrm{Res}(u_{ih}, \lambda_{ih}) \in V'$ *by*

$$\langle \mathrm{Res}(u_{ih}, \lambda_{ih}), v \rangle_{V',V} := \lambda_{ih}(u_{ih}, v) - (\nabla u_{ih}, \nabla v) \qquad \forall v \in V. \tag{2.3.2a}$$

Its dual norm is then

$$\|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1} := \sup_{v \in V,\, \|\nabla v\| = 1} \langle \mathrm{Res}(u_{ih}, \lambda_{ih}), v \rangle_{V',V}. \tag{2.3.2b}$$

The *Riesz representation* of the residual $\mathscr{r}_{(ih)} \in V$ is also very useful and is defined as

$$(\nabla \mathscr{r}_{(ih)}, \nabla v) = \langle \mathrm{Res}(u_{ih}, \lambda_{ih}), v \rangle_{V',V} \qquad \forall v \in V, \tag{2.3.3a}$$

$$\|\nabla \mathscr{r}_{(ih)}\| = \|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}. \tag{2.3.3b}$$

The *a posteriori* analysis is performed in two steps. First, the dual norm of the residual $\|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}$ is related to the error between the approximate solution u_{ih} and the exact solution u_i in L^2- and H^1-norms, i.e. to $\|u_i - u_{ih}\|$ and $\|\nabla(u_i - u_{ih})\|$. Once such an estimation is provided, the remaining difficulty is to estimate the dual norm of the residual $\|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}$. Indeed, this dual norm is in general not straightforward to compute and naively requires the resolution of problem (2.3.3a). Here, it is performed by means of equilibrated flux reconstruction, and strongly relies on [105]. Once these two elements are in order, one can assemble the estimations to get the full *a posteriori* estimate.

2.3.2 Results in the conforming case

For simplicity, we present now the results in the conforming case supposing that the eigenvalue solver is exact. This means that the approximate eigenpair (u_{ih}, λ_{ih}) is solution to the following equation on the approximate finite element space V_h written in weak form as

$$(\nabla u_{ih}, \nabla v_h) = \lambda_{ih}(u_{ih}, v_h) \qquad \forall v \in V_h.$$

We announce the main theorems leading to the derivation of the *a posteriori* bound.

Generic equivalences

To estimate the errors $\lambda_{ih} - \lambda_i$ and $\|\nabla(u_i - u_{ih})\|$, the errors $\|u_i - u_{ih}\|$ and $\|\nabla(u_i - u_{ih})\|$ are first evaluated in terms of the dual norm of the residual $\|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}$. First, the L^2-norm of the error $\|u_i - u_{ih}\|$ is related to the gradient norm of the Riesz representation of the residual $\|\nabla \mathfrak{r}_{(ih)}\|$ in the following lemma. Note that it can also be related to the norm of the Riesz representation $\|\mathfrak{r}_{(ih)}\|$. Let $i \geq 1$ and define (disregarding the left term for $i = 1$)

$$\widetilde{C}_{ih} := \min\left\{\lambda_{i-1}\left(1 - \frac{\lambda_{ih}}{\lambda_{i-1}}\right)^2, \lambda_{i+1}\left(1 - \frac{\lambda_{ih}}{\lambda_{i+1}}\right)^2\right\}. \tag{2.3.4}$$

Lemma 2.3.2 ($L^2(\Omega)$ bound via a quadratic residual inequality)**.** *Let $(u_{ih}, \lambda_{ih}) \in V \times \mathbb{R}^+$ with $\|u_{ih}\| = 1$ and $(u_i, u_{ih}) \geq 0$ be the i-th approximate eigenvector-eigenvalue pair, $i \geq 1$. Let λ_i be simple and let $\lambda_{i-1} < \lambda_{ih}$ when $i > 1$ and $\lambda_{ih} < \lambda_{i+1}$. Then*

$$\|u_i - u_{ih}\| \leq \alpha_{ih} := \sqrt{2}\widetilde{C}_{ih}^{-\frac{1}{2}}\|\nabla \mathfrak{r}_{(ih)}\|. \tag{2.3.5}$$

The assumption $(u_i, u_{ih}) \geq 0$ can actually be easily checked in practice (see Lemma 4.3.3). The second step is to relate the error on the eigenvalue $\lambda_{ih} - \lambda_i$ with the error on the eigenvector, and then to relate the error on the eigenvector $\|\nabla(u_i - u_{ih})\|$ to the dual norm of the residual $\|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}$. This is achieved in the two following theorems, which are presented and proved later on in Theorems 4.3.4 and 4.3.5.

Theorem 2.3.3 (Eigenvalue bounds)**.** *Let $u_{ih} \in V$ with $\|u_{ih}\| = 1$, $i \geq 1$, be arbitrary subject to $\|u_i - u_{ih}\| \leq \alpha_{ih}$ for some $\alpha_{ih} \in \mathbb{R}^+$ and $\|\nabla u_{ih}\|^2 = \lambda_{ih}$. Then*

$$\|\nabla(u_i - u_{ih})\|^2 - \lambda_i\alpha_{ih}^2 \leq \lambda_{ih} - \lambda_i \leq \|\nabla(u_i - u_{ih})\|^2. \tag{2.3.6}$$

Note that when the algebraic solver is not exact or in the non-conforming case, λ_{ih} is possibly different from $\|\nabla u_{ih}\|^2$, and the eigenvalue λ_{ih} has to be replaced by $\|\nabla u_{ih}\|^2$ in the estimation. Define

$$\overline{C}_{ih} := 1 \text{ if } i = 1, \quad \overline{C}_{ih} := \max\left\{\left(\frac{\lambda_{ih}}{\lambda_1} - 1\right)^2, 1\right\} \text{ if } i > 1. \tag{2.3.7}$$

Theorem 2.3.4 (Eigenvector bounds)**.** *Let* $(u_{ih}, \lambda_{ih}) \in V \times \mathbb{R}^+$ *with* $\|u_{ih}\| = 1$*,* $i \geq 1$*, be arbitrary subject to* $\|u_i - u_{ih}\| \leq \alpha_{ih}$ *for some* $\alpha_{ih} \in \mathbb{R}^+$*. Then*

$$\|\nabla(u_i - u_{ih})\|^2 \leq \|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}^2 + (\lambda_{ih} + \lambda_i)\alpha_{ih}^2, \tag{2.3.8a}$$

$$\|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}^2 \leq \frac{\|\nabla(u_i - u_{ih})\|^4}{\lambda_i} + \overline{C}_{ih}\|\nabla(u_i - u_{ih})\|^2. \tag{2.3.8b}$$

Up to this step, the equivalences are independent of the discretization method, and can be derived for a larger class of operators. An extension to the case of lower-bounded self-adjoint operators with compact resolvent is presented in the Appendix 5.10.1. Also, the only constants contained in the estimators depend on the exact eigenvalues. The error bounds can be therefore be estimated as soon as a (coarse) bound on the eigenvalues is available.

Dual norm of the residual estimation

In order to estimate the dual norm of the residual $\|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}$, we use a reconstruction method proposed in [105] for the Poisson problem, adapted here in the case of an eigenvalue problem. To motivate, note that from (2.3.1), it is straightforward that $-\nabla u_i \in \mathbf{H}(\mathrm{div}, \Omega)$, with the weak divergence equal to $\lambda_i u_i$. On the discrete level, however, we have, in general, $-\nabla u_{ih} \notin \mathbf{H}(\mathrm{div}, \Omega)$, and a fortiori $\nabla\cdot(-\nabla u_{ih}) \neq \lambda_{ih} u_{ih}$. We thus introduce an *equilibrated flux reconstruction*, a vector field $\boldsymbol{\sigma}_{ih}$ constructed from (u_{ih}, λ_{ih}), satisfying when the algebraic solver is exact

$$\boldsymbol{\sigma}_{ih} \in \mathbf{H}(\mathrm{div}, \Omega), \tag{2.3.9a}$$

$$\nabla\cdot\boldsymbol{\sigma}_{ih} = \lambda_{ih} u_{ih}. \tag{2.3.9b}$$

Naturally, only specific reconstructions $\boldsymbol{\sigma}_{ih}$ will give tight estimates, and we present in the following a *local construction* relying on the solution of homogeneous Neumann (Neumann–Dirichlet close to the boundary) problems by mixed finite elements. In the nonconforming case, as $u_{ih} \notin V$, one also has to introduce a reconstruction of the eigenvector, a scalar-valued function s_{ih} constructed from u_{ih} and satisfying $s_{ih} \in V$. Actually, both $\boldsymbol{\sigma}_{ih}$ and s_{ih} will be piecewise polynomials defined in standard finite element subspaces of $\mathbf{H}(\mathrm{div}, \Omega)$ and V, respectively.

The reconstruction method is based on the resolution of local problems on patches around vertices of the mesh. We denote by $\mathcal{T}_h$ the chosen mesh for solving the discrete problem. The set of vertices is denoted by $\mathcal{V}_h$, with interior vertices $\mathcal{V}_h^{\mathrm{int}}$, vertices located on the boundary $\mathcal{V}_h^{\mathrm{ext}}$, and a generic vertex $\mathbf{a}$. We call $\mathcal{T}_\mathbf{a}$ the patch of elements of $\mathcal{T}_h$ which share the vertex $\mathbf{a} \in \mathcal{V}_h$, $\omega_\mathbf{a}$ the corresponding subdomain, and $\mathbf{n}_{\omega_\mathbf{a}}$ its outward unit normal. We will often tacitly extend functions defined on $\omega_\mathbf{a}$ by zero outside of $\omega_\mathbf{a}$, whereas $V_h(\omega_\mathbf{a})$ stands for the restriction of the finite element space denoted by V_h to $\omega_\mathbf{a}$. Next, $\psi_\mathbf{a}$ for $\mathbf{a} \in \mathcal{V}_h$ stands for the piecewise affine "hat" function taking value 1 at the vertex $\mathbf{a}$ and zero at the other vertices. Remarkably, $(\psi_\mathbf{a})_{\mathbf{a}\in\mathcal{V}_h}$ form a partition of unity via $\sum_{\mathbf{a}\in\mathcal{V}_h} \psi_\mathbf{a} = 1|_\Omega$.

To define the reconstructed flux using problems on patches, we work with the Raviart–Thomas–Nédélec (RTN) mixed finite element spaces denoted by $\mathbf{V}_h \times Q_h \subset \mathbf{H}(\mathrm{div}, \Omega) \times L^2(\Omega)$. We define by $\mathbb{P}_s(K)$, $s \geq 0$ the space of polynomials of total degree at most s on $K \in \mathcal{T}_h$, and $\mathbb{P}_s(\mathcal{T}_h)$ the space of piecewise polynomials on $\mathcal{T}_h$, without any continuity requirement. We thus have, for degree $s \geq 0$, $\mathbf{V}_h := \{\mathbf{v}_h \in \mathbf{H}(\mathrm{div}, \Omega); \mathbf{v}_h|_K \in [\mathbb{P}_s(K)]^d + \mathbb{P}_s(K)\mathbf{x}\}$ and $Q_h := \mathbb{P}_s(\mathcal{T}_h)$. We also denote by Π_{Q_h} the $L^2(\Omega)$-orthogonal projection onto Q_h.

We now need to assume some properties on the approximate eigenpair (u_{ih}, λ_{ih}). More precisely, the pair (u_{ih}, λ_{ih}) must be orthogonal to all hat functions $\psi_\mathbf{a}$. Second, the approximate eigenvector u_{ih} must be a piecewise polynomial from $\mathbb{P}_p(\mathcal{T}_h)$, $p \geq 1$. Third, the construction of

the reconstructed flux $\boldsymbol{\sigma}_{ih}$ must be carried out in a sufficiently rich (order $p+1$) RTN space. These assumptions are precised below.

Assumption 2.3.5 (Galerkin orthogonality of the residual to $\psi_{\mathbf{a}}$)**.** *There holds*

$$\lambda_{ih}(u_{ih},\psi_{\mathbf{a}})_{\omega_{\mathbf{a}}} - (\nabla u_{ih},\nabla\psi_{\mathbf{a}})_{\omega_{\mathbf{a}}} = \langle \mathrm{Res}(u_{ih},\lambda_{ih}),\psi_{\mathbf{a}}\rangle_{V',V} = 0 \qquad \forall \mathbf{a}\in\mathcal{V}_h^{\mathrm{int}}.$$

Assumption 2.3.6 (Piecewise polynomials)**.** *There holds*

$$u_{ih}\in\mathbb{P}_p(\mathcal{T}_h),\ p\geq 1,\ \textit{and the spaces } \mathbf{V}_h\times Q_h \textit{ are of order } p+1.$$

If $u_{ih}\in V$ verifies these assumptions, we can construct concretely $\boldsymbol{\sigma}_{ih}$ by the following *local constrained minimizations*:

Definition 2.3.7 (Equilibrated flux reconstruction)**.** *Let $u_{ih}\in V$ satisfy Assumption 2.3.5. For $\mathbf{a}\in\mathcal{V}_h$, set*

$$\begin{aligned}
&\mathbf{V}_h^{\mathbf{a}} := \{\mathbf{v}_h\in\mathbf{V}_h(\omega_{\mathbf{a}});\ \mathbf{v}_h{\cdot}\mathbf{n}_{\omega_{\mathbf{a}}}=0 \text{ on } \partial\omega_{\mathbf{a}}\},\\
&Q_h^{\mathbf{a}} := \{q_h\in Q_h(\omega_{\mathbf{a}});\ (q_h,1)_{\omega_{\mathbf{a}}}=0\},
\end{aligned} \qquad \mathbf{a}\in\mathcal{V}_h^{\mathrm{int}},$$

$$\begin{aligned}
&\mathbf{V}_h^{\mathbf{a}} := \{\mathbf{v}_h\in\mathbf{V}_h(\omega_{\mathbf{a}});\ \mathbf{v}_h{\cdot}\mathbf{n}_{\omega_{\mathbf{a}}}=0 \text{ on } \partial\omega_{\mathbf{a}}\setminus\partial\Omega\},\\
&Q_h^{\mathbf{a}} := Q_h(\omega_{\mathbf{a}}),
\end{aligned} \qquad \mathbf{a}\in\mathcal{V}_h^{\mathrm{ext}}.$$

Then define $\boldsymbol{\sigma}_{ih} := \sum_{\mathbf{a}\in\mathcal{V}_h}\boldsymbol{\sigma}_{ih}^{\mathbf{a}}\in\mathbf{V}_h$, where $\boldsymbol{\sigma}_{ih}^{\mathbf{a}}\in\mathbf{V}_h^{\mathbf{a}}$ solve

$$\boldsymbol{\sigma}_{ih}^{\mathbf{a}} := \arg \min_{\substack{\mathbf{v}_h\in\mathbf{V}_h^{\mathbf{a}},\\ \nabla\cdot\mathbf{v}_h=\Pi_{Q_h}(\psi_{\mathbf{a}}\lambda_{ih}u_{ih}-\nabla u_{ih}\cdot\nabla\psi_{\mathbf{a}})}} \|\psi_{\mathbf{a}}\nabla u_{ih}+\mathbf{v}_h\|_{\omega_{\mathbf{a}}} \qquad \forall\mathbf{a}\in\mathcal{V}_h. \tag{2.3.10}$$

Note that the Euler–Lagrange equations for (2.3.10) give the standard *mixed finite element formulation*, cf. [105, Remark 3.7]: find $\boldsymbol{\sigma}_{ih}^{\mathbf{a}}\in\mathbf{V}_h^{\mathbf{a}}$ and $p_h^{\mathbf{a}}\in Q_h^{\mathbf{a}}$ such that

$$(\boldsymbol{\sigma}_{ih}^{\mathbf{a}},\mathbf{v}_h)_{\omega_{\mathbf{a}}} - (p_h^{\mathbf{a}},\nabla{\cdot}\mathbf{v}_h)_{\omega_{\mathbf{a}}} = -(\psi_{\mathbf{a}}\nabla u_{ih},\mathbf{v}_h)_{\omega_{\mathbf{a}}} \qquad \forall\mathbf{v}_h\in\mathbf{V}_h^{\mathbf{a}}, \tag{2.3.11a}$$

$$(\nabla{\cdot}\boldsymbol{\sigma}_{ih}^{\mathbf{a}},q_h)_{\omega_{\mathbf{a}}} = (\psi_{\mathbf{a}}\lambda_{ih}u_{ih}-\nabla u_{ih}\cdot\nabla\psi_{\mathbf{a}},q_h)_{\omega_{\mathbf{a}}} \qquad \forall q_h\in Q_h^{\mathbf{a}}. \tag{2.3.11b}$$

Under Assumption 2.3.5, this construction guaranties that $\nabla{\cdot}\boldsymbol{\sigma}_{ih}=\lambda_{ih}u_{ih}$, cf e.g., [105, Lemma 3.5]. This is actually crucial for deriving an upper bound of $\|\mathrm{Res}(u_{ih},\lambda_{ih})\|_{-1}$.

To give a lower bound of the dual norm of the residual $\|\mathrm{Res}(u_{ih},\lambda_{ih})\|_{-1}$, we actually introduce primal conforming local residual Riesz representations. To this aim, we solve *homogeneous local Neumann* (Neumann–Dirichlet close to the boundary) *problems* on the patches $\omega_{\mathbf{a}}$ via conforming primal counterparts of problems (2.3.11). Note that for all $\mathbf{a}\in\mathcal{V}_h$,

$$-(\psi_{\mathbf{a}}\nabla u_{ih},\nabla v)_{\omega_{\mathbf{a}}} + (\psi_{\mathbf{a}}\lambda_{ih}u_{ih}-\nabla u_{ih}\cdot\nabla\psi_{\mathbf{a}},v)_{\omega_{\mathbf{a}}} = \langle\mathrm{Res}(u_{ih},\lambda_{ih}),\psi_{\mathbf{a}}v\rangle_{V',V} \qquad \forall v\in V.$$

Let $\mathbf{a}\in\mathcal{V}_h$ and consider a patch $\omega_{\mathbf{a}}$ around the vertex $\mathbf{a}$. Define

$$H^1_*(\omega_{\mathbf{a}}) := \{v\in H^1(\omega_{\mathbf{a}});\ (v,1)_{\omega_{\mathbf{a}}}=0\}, \qquad \mathbf{a}\in\mathcal{V}_h^{\mathrm{int}}, \tag{2.3.12a}$$

$$H^1_*(\omega_{\mathbf{a}}) := \{v\in H^1(\omega_{\mathbf{a}});\ v=0 \text{ on } \partial\omega_{\mathbf{a}}\cap\partial\Omega\}, \quad \mathbf{a}\in\mathcal{V}_h^{\mathrm{ext}}, \tag{2.3.12b}$$

and let $X_h^{\mathbf{a}}$ stand for an arbitrary finite-dimensional subspace of $H^1_*(\omega_{\mathbf{a}})$. Typically, for $u_{ih}\in\mathbb{P}_p(\mathcal{T}_h)$, we choose $X_h^{\mathbf{a}} := \mathbb{P}_{p+1}(\mathcal{T}_{\mathbf{a}})\cap H^1_*(\omega_{\mathbf{a}})$.

Definition 2.3.8 (Conforming local Neumann problems)**.** *Define $r_{ih}^{\mathbf{a}}\in X_h^{\mathbf{a}}$ by*

$$(\nabla r_{ih}^{\mathbf{a}},\nabla v_h)_{\omega_{\mathbf{a}}} = \langle\mathrm{Res}(u_{ih},\lambda_{ih}),\psi_{\mathbf{a}}v_h\rangle_{V',V} \qquad \forall v_h\in X_h^{\mathbf{a}}$$

for each $\mathbf{a}\in\mathcal{V}_h$. Then set $r_{ih} := \sum_{\mathbf{a}\in\mathcal{V}_h}\psi_{\mathbf{a}}r_{ih}^{\mathbf{a}}$.

The functions $r_{ih}^{\mathbf{a}}$ are discrete Riesz projections of the local residual with hat-weighted test functions. Note that $r_{ih}^{\mathbf{a}} \notin V$ (when extended by zero outside of $\omega_{\mathbf{a}}$) but $\psi_{\mathbf{a}} r_{ih}^{\mathbf{a}} \in H_0^1(\omega_{\mathbf{a}})$, whence the sum r_{ih} indeed belongs to the space V.

Using Definitions 2.3.7 and 2.3.8, an estimation for the dual norm of the residual is provided in the following Theorem, presented for the conforming case with inexact algebraic solvers in Theorem 4.4.3.

Theorem 2.3.9 (Residual equivalences)**.** *Let $(u_{ih}, \lambda_{ih}) \in V \times \mathbb{R}$ verifying Assumptions 2.3.5 and 2.3.6 be arbitrary. Let $\boldsymbol{\sigma}_{ih}$ be constructed via Definition 2.3.7 and r_{ih} via Definition 2.3.8. Then*

$$\frac{\langle \mathrm{Res}(u_{ih}, \lambda_{ih}), r_{ih} \rangle_{V',V}}{\|\nabla r_{ih}\|} \leq \|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1} \leq \|\nabla u_{ih} + \boldsymbol{\sigma}_{ih}\|, \tag{2.3.13a}$$

$$\|\nabla u_{ih} + \boldsymbol{\sigma}_{ih}\| \leq (d+1) C_{\mathrm{st}} C_{\mathrm{cont,PF}} \|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}, \tag{2.3.13b}$$

where C_{st} and $C_{\mathrm{cont,PF}}$ are computable constants.

Final estimations

From (2.3.13) and the previous estimation (2.3.8), we can deduce estimates for the eigenvalue and the eigenvector errors. They are summarized in the following theorem, which is a simplified version of Theorems 4.5.1 and 4.5.7 in the conforming case with exact solvers.

Theorem 2.3.10 (Guaranteed lower bounds for the i-th eigenvalue and eigenvector)**.** *Let the i-th eigenvalue, $i \geq 1$, be simple and suppose the auxiliary bounds $\underline{\lambda}_1 \leq \lambda_1$, $\lambda_i \leq \overline{\lambda}_i$, $\underline{\lambda}_{i+1} \leq \lambda_{i+1}$, as well as $\lambda_{i-1} \leq \overline{\lambda}_{i-1}$ when $i > 1$, for $\underline{\lambda}_1, \overline{\lambda}_i, \underline{\lambda}_{i+1}, \overline{\lambda}_{i-1} > 0$. Let (u_{ih}, λ_{ih}) be any element of $\mathbb{P}_p(\mathcal{T}_h) \cap V \times \mathbb{R}^+$ verifying $\|u_{ih}\| = 1$, $\|\nabla u_{ih}\|^2 = \lambda_{ih}$, and the inequalities*

$$\overline{\lambda}_{i-1} < \lambda_{ih} \text{ when } i > 1, \qquad \lambda_{ih} < \underline{\lambda}_{i+1}. \tag{2.3.14}$$

Let next $\boldsymbol{\sigma}_{ih}$ and r_{ih} be respectively constructed following Definitions 2.3.7 and 2.3.8, and define

$$\eta_{i,\mathrm{res}} := \|\nabla u_{ih} + \boldsymbol{\sigma}_{ih}\|.$$

Set

$$\tilde{c}_{ih} := \max\left\{ \overline{\lambda}_{i-1}^{-\frac{1}{2}} \left(\frac{\lambda_{ih}}{\overline{\lambda}_{i-1}} - 1 \right)^{-1}, \underline{\lambda}_{i+1}^{-\frac{1}{2}} \left(1 - \frac{\lambda_{ih}}{\underline{\lambda}_{i+1}} \right)^{-1} \right\}, \tag{2.3.15a}$$

with the left terms in the max *disregarded for $i = 1$. If $(u_i, u_{ih}) \geq 0$ is known to hold, define $\overline{\alpha}_{ih} := \sqrt{2}\tilde{c}_{ih}\eta_{i,\mathrm{res}}$; if only $(u_{ih}, \chi_i) > 0$ holds, set $\overline{\alpha}_{ih} := \sqrt{2}(1 - \|u_{ih} - \Pi_i u_{ih}\|)^{-\frac{1}{2}} \tilde{c}_{ih}\eta_{i,\mathrm{res}}$, where $\Pi_i u_{ih}$ stands for the $L^2(\Omega)$-orthogonal projection of u_{ih} on the span of χ_i. Then for the eigenvalue*

$$\lambda_{ih} - \eta_i^2 \leq \lambda_i, \tag{2.3.16}$$

and the eigenvector error can be bounded via

$$\|\nabla(u_i - u_{ih})\| \leq \eta_i, \tag{2.3.17}$$

with

$$\eta_i^2 := \eta_{i,\mathrm{res}}^2 + (\lambda_{ih} + \overline{\lambda}_i)\overline{\alpha}_{ih}^2. \tag{2.3.18}$$

Note that these estimators can actually be refined under a smallness assumption on the error. On top of that, under a supplementary elliptic regularity assumption, the bounds can be optimally improved in the sense that the efficiency of the error bound goes to one when the mesh size h goes to zero. More generally, the efficiency of the estimator is shown in Theorem 4.5.7.

In these *a posteriori* estimators, the only uncomputable constants are the exact eigenvalues, which we aim at estimating. Luckily, we only need coarse bounds for the eigenvalues, and they can be estimated in different ways, for example using domain inclusions or other estimation methods for the eigenvalues, such as [176] in a coarse mesh. We refer to Remark 4.5.4 on this issue. Note that from a pratical point of view, it is possible to use the approximate eigenvalues as bounds for the lower and upper eigenvalues. The error bounds are not guaranteed any more, but they can still be used in practice, and numerically, the accuracy of the bounds is barely modified.

Use of inexact algebraic solvers

Up to now, we have presented results in the conforming case, supposing that the eigensolver was exact. In fact, in Chapter 4, the results are extended to take into account inexact algebraic solvers. The theory is similar, but the notations in the estimations are more involved. Also, the property $\|\nabla u_{ih}\|^2 = \lambda_{ih}$ does not hold anymore. Moreover, the reconstruction based on equilibrated fluxes takes into account the solver error, and mostly relies on [104].

The numerical results obtained in this case are presented in Section 4.7 for a set of test problems. They show that the assumptions under which the bounds are valid are satisfied already for very coarse meshes. Moreover, the bounds are shown to be very sharp, and converge at the right speed, both for the eigenvalues and for the eigenvectors.

2.3.3 Results in the nonconforming case

In Chapter 5, we extend this contribution to the case of nonconforming methods, including nonconforming, discontinuous Galerkin, and mixed finite elements. The formalism is heavier since in nonconforming methods, the property $\|\nabla u_{ih}\|^2 = \lambda_{ih}$ does not hold in general. Moreover, a discrete gradient depending on the method under consideration has to be defined to allow for irregular discrete functions.

The generic equivalences presented previously in the conforming case are very similar. In the reconstruction procedure, one needs however to employ an eigenvector reconstruction on top of the equilibrated flux reconstruction. The eigenvector reconstruction requires to solve local unconstrained minimization problems. Despite their apparent complexity, the final estimates presented in Theorems 5.6.1 and Theorems 5.6.3 are similar to the bounds presented in the conforming case in Theorems 4.5.1 and Theorems 4.5.2.

The numerical results provided in Section 5.8 for the nonconforming finite element method of order one and the symmetric discontinuous Galerkin finite element method of order one show the accuracy of these *a posteriori* bounds.

2.4 A posteriori error estimation for the Gross–Pitaevskii equation

In the previous section, we were interested in a linear eigenvalue problem. We now turn to the study of a nonlinear eigenvalue problem: the Gross–Pitaevskii equation. We summarize here the contribution presented in Chapter 6 for the *a posteriori* estimation of this equation.

The setting is chosen as simple as possible, within this nonlinear eigenvalue framework. The problem is 1-dimensional, whereas in density functional theory, the problems are posed in

$\mathbb{R}^3$. Moreover, the Gross–Pitaevskii energy functional is convex, which is not the case of the Kohn–Sham energy. We use a periodic setting, as is common for condensed matter systems.

The aim of this contribution is twofold. First, we provide a guaranteed *a posteriori* estimate for this particular nonlinear problem, based on the residual of the equation. Second, we separate the error components coming from two sources of error: the discretization and the algorithm used to solve the nonlinear problem. We are then able to determine which source is responsible for the largest error, and we balance the error components in order to save some computational resources.

To start with, the considered minimization problem reads

$$I = \inf \left\{ E(v),\ v \in X,\ \int_\Omega v^2 = 1 \right\}, \tag{2.4.1}$$

where the Gross–Pitaevskii energy functional is

$$E(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 + \frac{1}{2}\int_\Omega V v^2 + \frac{1}{4}\int_\Omega v^4,$$

and $X = H^1_\#(\Omega)$ is the Sobolev space defined in the more general settings for any $s \in \mathbb{R}$,

$$H^s_\#(\Omega) = \{ v|_\Omega,\ v \in H^s_{\mathrm{loc}}(\mathbb{R}) \mid v \text{ is 1-periodic} \},$$

provided with the norm denoted by $\|.\|_{H^s}$. We assume that $V \in L^p(\Omega)$ for some $p \geq 2$. The Euler–Lagrange equations of this constrained minimization problem write in a strong form: find $(u, \lambda) \in X \times \mathbb{R}$ such that

$$\left\{ \begin{array}{l} -\Delta u + Vu + u^3 = \lambda u \\ \|u\|_{L^2} = 1. \end{array} \right. \tag{2.4.2}$$

This nonlinear eigenvalue problem is discretized using planewaves (see Section 1.5.3). The discrete space denoted by X_N is defined by

$$X_N = \mathrm{Span}\left\{ e_k : x \mapsto e^{2ik\pi x}, |k| \leq N, k \in \mathbb{Z} \right\}.$$

The problem is solved using a fixed point algorithm, which is *not* a self-consistent field algorithm, but instead inspired from the inverse power method. An initial guess for the eigenpair (u_N^0, λ_N^0) is chosen. Then, at each iteration, the following linear boundary value problem is solved: find $u_N^{k*} \in X_N$ such that

$$\Pi_N \left(-\Delta u_N^{k*} + V u_N^{k*} + (u_N^{k-1})^2 u_N^{k*} \right) = \lambda_N^{k-1} u_N^{k-1}, \tag{2.4.3}$$

where Π_N is the orthogonal projection on X_N. The discrete solution u_N^{k*} is completely determined by the knowledge of $(\lambda_N^{k-1}, u_N^{k-1})$. Since u_N^{k*} is *a priori* a non-normalized vector, we normalize it and define u_N^k by

$$u_N^k = \frac{u_N^{k*}}{\|u_N^{k*}\|_{L^2}}. \tag{2.4.4}$$

Finally, we define the approximation of the eigenvalue λ_N^k as a Rayleigh quotient being

$$\lambda_N^k = \int_\Omega (\nabla u_N^k)^2 + \int_\Omega V (u_N^k)^2 + \int_\Omega (u_N^k)^4. \tag{2.4.5}$$

The approximate solution in a planewave basis with parameter N and after k iterations in the fixed-point algorithm is denoted by (u_N^k, λ_N^k). The residual R_N^k of this approximate solution is defined by

$$R_N^k = -\Delta u_N^k + V u_N^k + (u_N^k)^3 - \lambda_N^k u_N^k.$$

The *a posteriori* error estimation proposed in this contribution strongly relies on the *a priori* analysis that was performed in [43], detailed for an improved estimation of the involved constants. The *a posteriori* analysis is composed of two steps. In a first step, we use arguments based on a theory developed by Caloz–Rappaz in [40] to provide an upper bound of the error, which is guaranteed under *a posteriori* computable hypotheses. This result is presented in Lemma 6.3.1, and reads in a simplified version as

Lemma 2.4.1. *If the computable conditions detailed in Lemma 6.3.1 (which imply that N and k are large enough) are satisfied, there exists a unique solution $(\tilde{u}, \tilde{\lambda})$ to equation (2.4.2) in the ball $B((u_N^k, \lambda_N^k), 2\gamma\varepsilon)$, where γ is fully computable, and $\varepsilon = \|-\Delta u_N^k + V u_N^k + (u_N^k)^3 - \lambda_N^k u_N^k\|_{H^{-1}}$ is the dual norm of the residual and goes to zero as N and k go to infinity. Moreover,*

$$\|\tilde{u} - u_N^k\|_{H^1} + |\tilde{\lambda} - \lambda_N^k| \leq 2\gamma \| - \Delta u_N^k + V u_N^k + (u_N^k)^3 - \lambda_N^k u_N^k\|_{H^{-1}} \tag{2.4.6}$$

and there exists a computable condition depending on $\|\tilde{u} - u_N^k\|_{H^1}$, $|\tilde{\lambda} - \lambda_N^k|$, u_N^k, λ_N^k, and other computable quantities guaranteeing that $(\tilde{u}, \tilde{\lambda})$ is the ground state (u, λ) of (2.4.1).

This estimate guaranties that the error is bounded by a computable quantity times the dual norm of the residual. The computation of this estimate requires nevertheless to solve a linear eigenvalue problem on the discrete space. Note that compared to finite elements, the dual norm of the residual is easily computable here, since the Laplace operator is diagonal in planewaves.

As a main limitation, this bound overestimates the real error by a factor close to 2γ. Therefore, we designed another *a posteriori* error estimate, which is valid in the asymtotic regime, i.e. under a smallness assumption on the error. This assumption can effectively be guaranteed by Lemma 2.4.1.

Determining the higher order terms in the estimation of the H^1-norm of the error $\|u - u_N^k\|_{H^1}$ and neglecting them, one can show that asymtotically, the error $\|u - u_N^k\|_{H^1}$ is bounded by the dual norm of the residual plus some other terms that are computable. And in the nice case where the L^∞-norm of the negative part of $V + 3(u_N^k)^2 - \lambda_N^k - 1$ denoted by $\|(V + 3(u_N^k)^2 - \lambda_N^k - 1)_-\|_{L^\infty}$ is equal to zero, we have the following error estimate, valid for any $\alpha > 1$ and as close to 1 as we wish as the error goes to zero:

$$\|u - u_N^k\|_{H^1} \leq \alpha \|R_N^k\|_{H^{-1}}. \tag{2.4.7}$$

Using this error bound, it is possible to estimate the error coming from the two successive approximations: the discretization with parameter N and the iterative scheme with k iterations. To this aim, we separate this global error into two components, each of them depending mainly on one parameter associated with the above sources of error. In fact, one only needs to decompose the residual into two parts. The contribution based on the discretization corresponds to the residual relative to the numerical scheme and is defined as

$$R_{disc} = -\Delta u_N^k + V u_N^k + (u_N^{k-1})^2 u_N^k - \|u_N^{k*}\|_{L^2}^{-1} \lambda_N^{k-1} u_N^{k-1}. \tag{2.4.8}$$

The quantity $\|R_{disc}\|_{H^{-1}}$ measures the discretization error and depends on the finite dimension $(2N + 1)$ of the Fourier space X_N on which we solve the problem. The iteration residual is then defined such that $R_N^k = R_{disc} + R_{iter}$. Hence

$$R_{iter} = (u_N^k)^3 - (u_N^{k-1})^2 u_N^k - \lambda_N^k u_N^k + \|u_N^{k*}\|_{L^2}^{-1} \lambda_N^{k-1} u_N^{k-1}. \tag{2.4.9}$$

The quantity $\|R_{iter}\|_{H^{-1}}$ corresponds to the algorithm error and depends mainly on the finite number of iterations k. Hence, the previous error estimate can be split into two parts

$$\|u - u_N^k\|_{H^1} \leq \alpha \Big(\|R_{disc}\|_{H^{-1}} + \|R_{iter}\|_{H^{-1}} \Big). \tag{2.4.10}$$

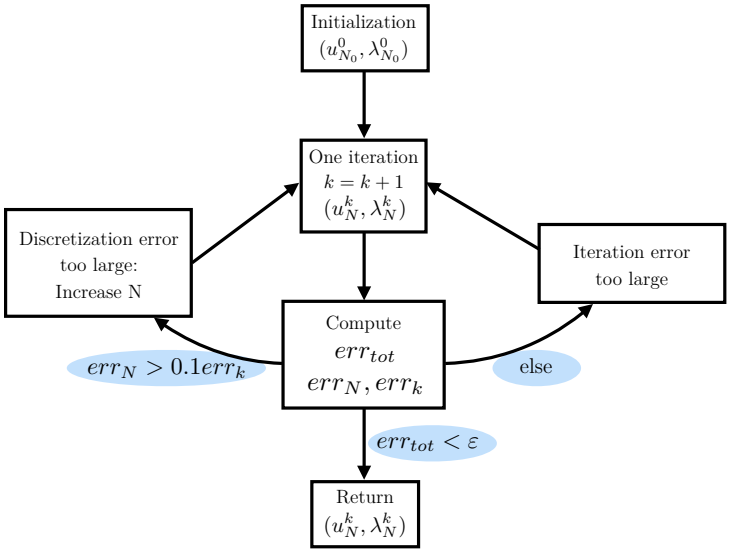

Figure 2.2 – Error balance algorithm given an error tolerance ε.

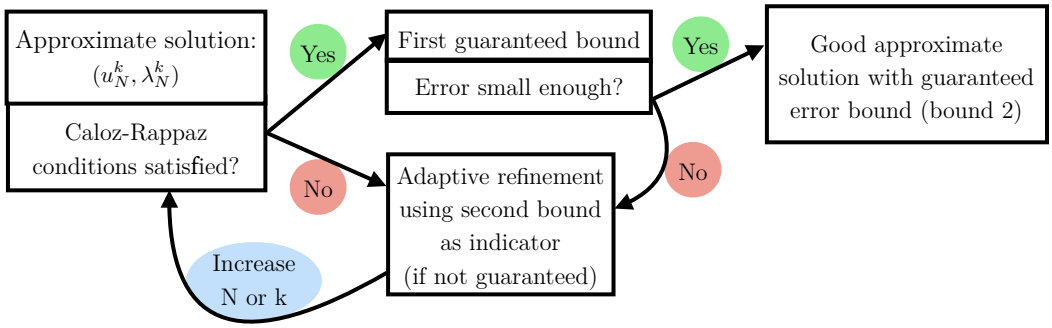

Figure 2.3 – Illustration of the guaranteed estimation and error balance strategy.

In this case, the error components are given by

$$err_N = \|R_{disc}\|_{H^{-1}},$$

and

$$err_k = \|R_{iter}\|_{H^{-1}}.$$

In the general case where $\|(V + 3(u^k_N)^2 - \lambda^k_N - 1)_-\|_{L^\infty} \neq 0$, a similar estimation can be performed. The terms involved are just slightly more complicated, and are presented in Theorem 6.3.1.

Some numerical results are presented to show that the error components err_N and err_k numerically depend mostly respectively on N and k. An adaptive algorithm described in Figure 2.2 has also been provided, refining at each step the parameter corresponding to the largest error component, proving the concept of error balance in this context.

Note that while the sharp error estimate is not guaranteed, it can still be used as an error indicator, and the adaptive refinement can be performed. When the second bound is guaranteed, we have then a guaranteed upper bound of the error, and an adaptive strategy to spare some computational resources. This adaptive strategy is illustrated in Figure 2.3.

2.5 Post-processing for the Kohn–Sham model

2.5.1 Perturbation theory

In Part V, we present a post-processing method for linear and nonlinear eigenvalue problems, including Kohn–Sham models. This method consists of performing a full computation for the problem under consideration in a coarse basis, and then to compute corrections on the eigenvalues and eigenfunctions in a fine basis, at a low computational cost. These corrections are turned into corrections on the ground-state density matrix and the ground-state energy, which are both theoretically and numerically improved.

This method is based on Rayleigh–Schrödinger perturbation theory [146]. Let us start by introducing this theory in the case of a linear eigenvalue problem. To adopt the notations of Part V, we consider a periodic setting. The unit cell is denoted by Ω and the periodic lattice by $\mathcal{R}$. Let $\mathcal{H}$ be a bounded below self-adjoint operator on $L^2_\#(\Omega)$ with domain $H^2_\#(\Omega)$ and compact resolvent. The operator $\mathcal{H}$ can be diagonalized in an orthonormal basis: there exists a non-decreasing sequence $(\lambda_i^0)_{i\geq 1}$ of real numbers going to infinity and an orthonormal basis $(\phi_i^0)_{i\geq 1}$ of $L^2_\#(\Omega)$ consisting of functions of $H^2_\#(\Omega)$ such that

$$\forall i \geq 1, \quad \mathcal{H}\,\phi_i^0 = \lambda_i^0 \phi_i^0. \tag{2.5.1}$$

Let us now assume that we are only able to determine the eigenpairs $(\phi_{i,N_\mathrm{c}}, \lambda_{i,N_\mathrm{c}})_{i\geq 1}$ of another, simpler operator denoted by $\mathcal{H}_{N_\mathrm{c}}$, which is also bounded below, self-adjoint on $L^2_\#(\Omega)$ with domain $H^2_\#(\Omega)$ and compact resolvent. The eigenpairs $(\phi_{i,N_\mathrm{c}}, \lambda_{i,N_\mathrm{c}})_{i\geq 1}$ verify

$$\forall i \geq 1, \quad \mathcal{H}_{N_\mathrm{c}}\phi_{i,N_\mathrm{c}} = \lambda_{i,N_\mathrm{c}}\phi_{i,N_\mathrm{c}}. \tag{2.5.2}$$

In our case, N_c is a planewave discretization parameter defined in Section 1.5.3. Using a Galerkin discretization of equation (2.5.1), $\mathcal{H}_{N_\mathrm{c}}$ would typically be equal to $\mathcal{H}_{N_\mathrm{c}} = \Pi_{N_\mathrm{c}}\mathcal{H}\,\Pi_{N_\mathrm{c}}$, with Π_{N_c} the orthogonal projector on the discrete space for the $L^2_\#(\Omega)$ scalar product.

In standard perturbation theory, one considers the operator $\mathcal{H}\,(\beta) = \mathcal{H}_{N_\mathrm{c}} + \beta(\mathcal{H} - \mathcal{H}_{N_\mathrm{c}})$, where β is a real parameter. Thus, for $\beta = 0$, $\mathcal{H}\,(0) = \mathcal{H}_{N_\mathrm{c}}$ which is the coarse operator, and for $\beta = 1$, $\mathcal{H}\,(1) = \mathcal{H}$, i.e. the exact operator. Then, the eigenvectors $\phi_i^0(\beta)$ and eigenvalues $\lambda_i^0(\beta)$ of the operator $\mathcal{H}\,(\beta)$ are expanded in terms of power series of β. We write formally

$$\forall i \geq 1, \quad \phi_i^0(\beta) = \sum_{k\geq 0} \beta^k \phi_{i,N_\mathrm{c}}^{(k)}, \qquad \lambda_i^0(\beta) = \sum_{k\geq 0} \beta^k \lambda_{i,N_\mathrm{c}}^{(k)},$$

where $\lambda_{i,N_\mathrm{c}}^{(0)} = \lambda_{i,N_\mathrm{c}}$ and $\phi_{i,N_\mathrm{c}}^{(0)} = \phi_{i,N_\mathrm{c}}$. The equation depending of β can be written as: for all $i = 1, \ldots, N$,

$$\left\{ \begin{aligned} (\mathcal{H}_{N_\mathrm{c}} + \beta(\mathcal{H} - \mathcal{H}_{N_\mathrm{c}}))\left(\sum_{k=0}^{\infty} \beta^k \phi_{i,N_\mathrm{c}}^{(k)}\right) &= \left(\sum_{k=0}^{\infty} \beta^k \lambda_{i,N_\mathrm{c}}^{(k)}\right)\left(\sum_{k=0}^{\infty} \beta^k \phi_{i,N_\mathrm{c}}^{(k)}\right) \\ \left\|\sum_{k=0}^{\infty} \beta^k \phi_{i,N_\mathrm{c}}^{(k)}\right\|_{L^2}^2 &= \sum_{k,l=0}^{\infty} \beta^{k+l} \int_\Omega \phi_{i,N_\mathrm{c}}^{(k)} \phi_{i,N_\mathrm{c}}^{(l)} = 1. \end{aligned} \right. \tag{2.5.3}$$

Then, separating the different orders in β and taking $\beta = 1$, we obtain the perturbed equations at different orders. Note that the series will converge only if the difference between the exact and approximate operators $\mathcal{H} - \mathcal{H}_{N_\mathrm{c}}$ is small. In general, the perturbation equations are only developed up to few orders (one or two, four maximum). In our post-processing method, we only compute the corrections at first-order for the eigenfunctions and second-order for the eigenvalues.

For a given $i \geq 1$, developing equation (2.5.3) at 0^{th} order, one obtains exactly the approximate equation (2.5.2). At 1^{st} order, the equation reads

$$\mathcal{H}_{N_c}\phi^{(1)}_{i,N_c} + (\mathcal{H} - \mathcal{H}_{N_c})\phi_{i,N_c} = \lambda_{i,N_c}\phi^{(1)}_{i,N_c} + \lambda^{(1)}_{i,N_c}\phi_{i,N_c}.$$

Projecting this equation on ϕ_{i,N_c} gives

$$\langle \phi_{i,N_c}, \mathcal{H}_{N_c}\phi^{(1)}_{i,N_c}\rangle + \langle \phi_{i,N_c}, (\mathcal{H} - \mathcal{H}_{N_c})\phi_{i,N_c}\rangle = \lambda_{i,N_c}\langle \phi_{i,N_c}, \phi^{(1)}_{i,N_c}\rangle + \lambda^{(1)}_{i,N_c}\langle \phi_{i,N_c}, \phi_{i,N_c}\rangle.$$

As $\mathcal{H}_{N_c}$ is self-adjoint, $\langle \phi_{i,N_c}, \mathcal{H}_{N_c}\phi^{(1)}_{i,N_c}\rangle = \langle \mathcal{H}_{N_c}\phi_{i,N_c}, \phi^{(1)}_{i,N_c}\rangle = \lambda_{i,N_c}\langle \phi_{i,N_c}, \phi^{(1)}_{i,N_c}\rangle$, from (2.5.2). Therefore,

$$\lambda^{(1)}_{i,N_c} = \langle \phi_{i,N_c}, (\mathcal{H} - \mathcal{H}_{N_c})\phi_{i,N_c}\rangle.$$

Note that this quantity is equal to zero if $\mathcal{H}_{N_c} = \Pi_{N_c}\mathcal{H}\,\Pi_{N_c}$. This gives in general for $\phi^{(1)}_{i,N_c}$

$$(\mathcal{H}_{N_c} - \lambda_{i,N_c})\phi^{(1)}_{i,N_c} = \left(\langle \phi_{i,N_c}, (\mathcal{H} - \mathcal{H}_{N_c})\phi_{i,N_c}\rangle - (\mathcal{H} - \mathcal{H}_{N_c})\right)\phi_{i,N_c}.$$

Let us now project this equation on ϕ_{j,N_c} for $j \geq 1$. We obtain

$$\langle \phi_{j,N_c}, (\mathcal{H}_{N_c} - \lambda_{i,N_c})\phi^{(1)}_{i,N_c}\rangle = \langle \phi_{j,N_c}, \left(\langle \phi_{i,N_c}, (\mathcal{H} - \mathcal{H}_{N_c})\phi_{i,N_c}\rangle - (\mathcal{H} - \mathcal{H}_{N_c})\right)\phi_{i,N_c}\rangle.$$

Using the self-adjointness of $\mathcal{H}_{N_c}$ leads to

$$(\lambda_{j,N_c} - \lambda_{i,N_c})\langle \phi_{j,N_c}, \phi^{(1)}_{i,N_c}\rangle = \langle \phi_{j,N_c}, \left(\langle \phi_{i,N_c}, (\mathcal{H} - \mathcal{H}_{N_c})\phi_{i,N_c}\rangle - (\mathcal{H} - \mathcal{H}_{N_c})\right)\phi_{i,N_c}\rangle\,.$$

Therefore, for $i \geq 1$ such that λ_{i,N_c} is a simple eigenvalue, the correction at first order is given by

$$\phi^{(1)}_{i,N_c} = \sum_{j\neq i} \frac{\langle \phi_{j,N_c}, \left(\langle \phi_{i,N_c}, (\mathcal{H} - \mathcal{H}_{N_c})\phi_{i,N_c}\rangle - (\mathcal{H} - \mathcal{H}_{N_c})\right)\phi_{i,N_c}\rangle}{\lambda_{j,N_c} - \lambda_{i,N_c}}\phi_{j,N_c}. \tag{2.5.4}$$

The corrected eigenfunctions at first order write

$$\begin{aligned}\widetilde{\phi_{i,N_c}} &= \phi_{i,N_c} + \phi^{(1)}_{i,N_c} \\ &= \phi_{i,N_c} + \sum_{j\neq i} \frac{\langle \phi_{j,N_c}, \left(\langle \phi_{i,N_c}, (\mathcal{H} - \mathcal{H}_{N_c})\phi_{i,N_c}\rangle - (\mathcal{H} - \mathcal{H}_{N_c})\right)\phi_{i,N_c}\rangle}{\lambda_{j,N_c} - \lambda_{i,N_c}}\phi_{j,N_c}.\end{aligned}$$

This is a classical development which can be found in many courses about perturbation theory (see e.g. [135, Chapter 14]).

2.5.2 Application to linear Schrödinger operators

In the post-processing method proposed in [49, 50], the operator under consideration is of the form $\mathcal{H} = -\frac{1}{2}\Delta + \mathcal{V}$, where $\mathcal{V}$ is a multiplicative potential, satisfying the following assumption:

Assumption 2.5.1. *$\mathcal{V}$ is a $\mathcal{R}-$periodic potential such that $\mathcal{V} \in L^\infty_\#(\Omega)$ and $\nabla\mathcal{V} \in L^3_\#(\Omega)$.*

We are interested in the lowest N eigenvalues and corresponding eigenfunctions of the operator $\mathcal{H}$, denoted respectively by $\Lambda^0 = \mathrm{diag}(\lambda^0_1, \cdots, \lambda^0_N)$ and $\Phi^0 = (\phi^0_1, \cdots, \phi^0_N)^T$. We use a planewave discretization with parameter N_c. Recall that the Laplace operator is diagonal in this basis. We denote by X_{N_c} the discrete space on which we solve the problem and Π_{N_c} the orthogonal projection for any $H^s_\#$ scalar product ($s \in \mathbb{R}$). The discrete eigenvalues and eigenfunctions denoted respectively by $\Lambda_{N_c} = \mathrm{diag}(\lambda_{1,N_c}, \cdots, \lambda_{N,N_c})$ and $\Phi_{N_c} = (\phi_{1,N_c}, \cdots, \phi_{N,N_c})^T$ are

computed from the operator $-\frac{1}{2}\Pi_{N_c}\Delta\Pi_{N_c}+\Pi_{N_c}\mathcal{V}\Pi_{N_c}$. Luckily, the operator $-\frac{1}{2}\Delta+\Pi_{N_c}\mathcal{V}\Pi_{N_c}$ can be decomposed as

$$\underbrace{\begin{pmatrix} \boxed{-\frac{1}{2}\Pi_{N_c}\Delta\Pi_{N_c}+\Pi_{N_c}\mathcal{V}\Pi_{N_c}} & 0 \\ 0 & \boxed{-\frac{1}{2}\Delta} \end{pmatrix}}_{X_{N_c}\quad\quad X_{N_c}^{\perp}} \begin{matrix} \} X_{N_c} \\ \} X_{N_c}^{\perp} \end{matrix}$$

On $X_{N_c}^{\perp}$, the operator $-\frac{1}{2}\Delta$ is diagonal with a smallest eigenvalue proportional to N_c^2. Therefore, as soon as the eigenvalues $(\lambda_{1,N_c},\cdots,\lambda_{N,N_c})$ are smaller than the lowest eigenvalue of $-\frac{1}{2}\Delta$ on $X_{N_c}^{\perp}$, they are also the lowest eigenvalues of the operator

$$\mathcal{H}_{N_c}=-\frac{1}{2}\Delta+\Pi_{N_c}\mathcal{V}\Pi_{N_c}.$$

Note that for $i\leq N$, $\phi_{i,N_c}\in X_{N_c}$ and that for $i>\dim(X_{N_c})$, the eigenvalues of $\mathcal{H}_{N_c}$ are explicitly known and correspond to the eigenvalues of the operator $-\frac{1}{2}\Delta$.

If this assumption is verified, which can be checked in practice, the corrections on the eigenfunctions defined in (2.5.4) corresponding to simple eigenvalues become in this case

$$\begin{aligned}
\phi_{i,N_c}^{(1)} =& \sum_{j\neq i}\frac{\langle\phi_{j,N_c},(\langle\phi_{i,N_c},(\mathcal{V}-\Pi_{N_c}\mathcal{V}\Pi_{N_c})\phi_{i,N_c}\rangle-(\mathcal{V}-\Pi_{N_c}\mathcal{V}\Pi_{N_c}))\,\phi_{i,N_c}\rangle}{\lambda_{j,N_c}-\lambda_{i,N_c}}\phi_{j,N_c}\\
=&-\sum_{j\neq i}\frac{\langle\phi_{j,N_c},(\mathcal{V}-\Pi_{N_c}\mathcal{V}\Pi_{N_c})\phi_{i,N_c}\rangle}{\lambda_{j,N_c}-\lambda_{i,N_c}}\phi_{j,N_c}\\
=&-\sum_{j>\dim(X_{N_c})}\frac{\langle\phi_{j,N_c},(\mathcal{V}-\Pi_{N_c}\mathcal{V}\Pi_{N_c})\phi_{i,N_c}\rangle}{\lambda_{j,N_c}-\lambda_{i,N_c}}\phi_{j,N_c}\\
=&-(-\frac{1}{2}\Delta-\lambda_{i,N_c})^{-1}(\mathcal{V}-\Pi_{N_c}\mathcal{V}\Pi_{N_c})\phi_{i,N_c}.
\end{aligned}$$

Using this last formula, the corrections can be defined even with multiple eigenvalues. In fact, in the proofs, we will only need the following assumption.

Assumption 2.5.2. *There is a gap between the N^{th} and the $(N+1)^{\text{st}}$ eigenvalues of $\mathcal{H}$, i.e.*

$$g:=\lambda_{N+1}^0-\lambda_N^0>0.$$

Since $-\frac{1}{2}\Delta\phi_{i,N_c}+\Pi_{N_c}\mathcal{V}\Pi_{N_c}\phi_{i,N_c}=\lambda_{i,N_c}\phi_{i,N_c}$, the residual Res_{i,N_c} of the eigenpair $(\phi_{i,N_c},\lambda_{i,N_c})$ can be written as

$$\begin{aligned}
\mathrm{Res}_{i,N_c} =& -\frac{1}{2}\Delta\phi_{i,N_c}+\mathcal{V}\phi_{i,N_c}-\lambda_{i,N_c}\phi_{i,N_c}\\
=&-\frac{1}{2}\Delta\phi_{i,N_c}+\Pi_{N_c}\mathcal{V}\Pi_{N_c}\phi_{i,N_c}-\lambda_{i,N_c}\phi_{i,N_c}+(\mathcal{V}-\Pi_{N_c}\mathcal{V}\Pi_{N_c})\phi_{i,N_c}\\
=&(\mathcal{V}-\Pi_{N_c}\mathcal{V}\Pi_{N_c})\phi_{i,N_c}.
\end{aligned}$$

Therefore the post-processed eigenfunctions can be defined as

$$\widetilde{\phi_{i,N_c}}=\phi_{i,N_c}-(-\frac{1}{2}\Delta-\lambda_{i,N_c})^{-1}\mathrm{Res}_{i,N_c}. \tag{2.5.5}$$

This is the formula obtained in [48]. A similar development at second order gives for the eigenvalues

$$\widetilde{\lambda_{i,N_\mathrm{c}}} = \lambda_{i,N_\mathrm{c}} + \lambda^{(2)}_{i,N_\mathrm{c}} = \lambda_{i,N_\mathrm{c}} + \langle \phi^{(1)}_{i,N_\mathrm{c}} | \mathrm{Res}_{i,N_\mathrm{c}} \rangle = \lambda_{i,N_\mathrm{c}} - \langle \mathrm{Res}_{i,N_\mathrm{c}} | \left(-\frac{1}{2}\Delta - \lambda_{i,N_\mathrm{c}} \right)^{-1} | \mathrm{Res}_{i,N_\mathrm{c}} \rangle. \tag{2.5.6}$$

To compute these post-processed eigenfunctions and eigenvalues, the residual $\mathrm{Res}_{i,N_\mathrm{c}}$ has to be computed on a finer grid than the grid used for the calculations, but then, one only needs to invert a diagonal operator on a finer grid, which is very cheap.

To characterize the improvement brought by the corrections, we do not work directly on the eigenfunctions and eigenvalues, especially since we consider possible multiple eigenvalues. Instead, we consider the ground-state density matrix of the system and the ground-state energy. The density matrix is the orthogonal projector on the space spanned by the N considered eigenfunctions. Using the Dirac bra-ket notation, the exact and approximate density matrices are respectively defined by

$$\gamma_0 := \sum_{i=1}^{N} |\phi^0_i\rangle\langle\phi^0_i|, \qquad \gamma_{0,N_\mathrm{c}} = \sum_{i=1}^{N} |\phi_{i,N_\mathrm{c}}\rangle\langle\phi_{i,N_\mathrm{c}}|.$$

In the linear case, the ground-state energy corresponds to the sum of the lowest N eigenvalues. The exact and approximate energies are respectively defined by

$$\mathcal{E}_0 := \mathrm{Tr}\left(\mathcal{H}\,\gamma_0\right) = \sum_{i=1}^{N} \lambda^0_i, \qquad \mathcal{E}_{0,N_\mathrm{c}} = \mathrm{Tr}\left(\mathcal{H}\,\gamma_{0,N_\mathrm{c}}\right) = \sum_{i=1}^{N} \lambda_{i,N_\mathrm{c}}.$$

The traces of operators Tr will be properly introduced in Chapter 8.

It is then possible to define corrections and the density-matrix and the energy as follows (see Definition 8.3.1).

Definition 2.5.3 (Perturbed density matrix, and energy)**.** *For all N_c such that the corrections on the eigenfunctions and the orbitals are well-defined. The perturbed density matrix is defined as*

$$\widetilde{\gamma_{N_\mathrm{c}}} = \gamma_{0,N_\mathrm{c}} + \gamma^{(1)}_{N_\mathrm{c}},$$

with

$$\gamma^{(1)}_{N_\mathrm{c}} = \sum_{i=1}^{N} |\phi^{(1)}_{i,N_\mathrm{c}}\rangle\langle\phi_{i,N_\mathrm{c}}| + \sum_{i=1}^{N} |\phi_{i,N_\mathrm{c}}\rangle\langle\phi^{(1)}_{i,N_\mathrm{c}}|, \tag{2.5.7}$$

and the perturbed energy as

$$\widetilde{\mathcal{E}_{0,N_\mathrm{c}}} = \sum_{i=1}^{N} \widetilde{\lambda_{i,N_\mathrm{c}}} = \mathrm{Tr}\left(\gamma_{0,N_\mathrm{c}}\mathcal{H}\,\widetilde{\gamma_{N_\mathrm{c}}}\right). \tag{2.5.8}$$

The main results of this contribution is the following theorem, which states the asymptotic improvement on the post-processed density matrix and energy.

Theorem 2.5.4. *Under Assumptions 2.5.1–2.5.2, there exist $C \in \mathbb{R}_+$ and $N^0_\mathrm{c} \in \mathbb{N}$, such that for $N_\mathrm{c} \geq N^0_\mathrm{c}$,*

$$\|(1-\Delta)^{1/2}(\widetilde{\gamma_{N_\mathrm{c}}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)} \leq C N_\mathrm{c}^{-2} \|(1-\Delta)^{1/2}(\gamma_{0,N_\mathrm{c}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}, \tag{2.5.9}$$

and

$$\left|\widetilde{\mathcal{E}_{0,N_\mathrm{c}}} - \mathcal{E}_0\right| \leq C N_\mathrm{c}^{-2} \left|\mathcal{E}_{0,N_\mathrm{c}} - \mathcal{E}_0\right|. \tag{2.5.10}$$

The improvement on the density matrix is provided in terms of Hilbert–Schmidt norm. In the case where all eigenvalues are simple, the quantity $\|(1-\Delta)^{1/2}(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}$ easily relates to the $H^1_\#$-norm of the error between the discrete eigenfunctions and exact eigenfunctions, and $\|(1-\Delta)^{1/2}(\widetilde{\gamma_{N_\mathrm{c}}}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}$ to the $H^1_\#$-norm of the error between the corrected eigenfunctions and exact eigenfunctions. This will be exposed in more detail in Chapter 8, together with the proof of this theorem. Therefore, we gain asymptotically a factor N_c^{-2} for the energy and for the density matrix, at the price of computing N residuals on a finer basis, which requires two FFTs per eigenvalue.

2.5.3 Application to the Kohn–Sham model

This post-processing method can be extended to Kohn–Sham models (see Section 1.3.3) as is presented in Chapters 7 and 9. We focus on the periodic versions of the Kohn–Sham LDA model with pseudopotentials, which we do not detail here, but can be found in Chapter 7. The main goal of this section is to show how the previous post-processing method can be applied in a nonlinear case. Let $N \in \mathbb{N}^*$ be the number of particles in the system. We consider an Hamiltonian $\mathcal{H}_{[\rho]}$ depending on the electronic density ρ defined by

$$\mathcal{H}_{[\rho]} = -\frac{1}{2}\Delta + V_\mathrm{ion} + V_\mathrm{coul}(\rho) + V_\mathrm{xc}(\rho), \tag{2.5.11}$$

where the electronic density is defined for a set of orbitals $\Psi = (\psi_1, \dots, \psi_N)$ by

$$\rho_{[\Psi]}(\mathbf{r}) = 2\sum_{i=1}^{N} |\psi_i(\mathbf{r})|^2.$$

In the Hamiltonian, V_ion is called the ionic potential and contains a local and a nonlocal part, V_coul denotes the Coulomb potential, and V_xc the exchange-correlation potential. The corresponding energy depends nonlinearly on the orbitals Ψ and is denoted by $\mathcal{E}^\mathrm{KS}_{0,\Omega}(\Psi)$. The precise definition of the energy can be found in Chapter 7.

We are looking for the lowest N eigenvalues and corresponding eigenfunctions respectively denoted by and $\Lambda^0 = \mathrm{diag}(\lambda_1^0, \cdots, \lambda_N^0)$ and $\Phi^0 = (\phi_1^0, \cdots, \phi_N^0)^T$, verifying

$$\forall i = 1, \dots, N, \quad \mathcal{H}_{[\rho^0]}\,\phi_i^0 = \lambda_i^0 \phi_i^0, \qquad \text{and} \qquad \forall i, j = 1, \dots, N, \quad \langle \phi_i^0 | \phi_j^0 \rangle = \delta_{ij}, \tag{2.5.12}$$

where

$$\rho^0 = \rho_{[\Phi^0]}.$$

We denote by $\mathcal{H}_0 = \mathcal{H}_{[\rho^0]}$ the exact Hamiltonian. As in the previous section, it is a bounded-below self-adjoint operator on $L^2_\#(\Omega)$ with domain $H^2_\#(\Omega)$ and compact resolvent. The ground-state density matrix of the system is defined by

$$\gamma_0 := \sum_{i=1}^{N} |\phi_i^0\rangle\langle\phi_i^0|, \tag{2.5.13}$$

and the ground-state energy by

$$\mathcal{E}_0 := \mathcal{E}^\mathrm{KS}_{0,\Omega}(\Phi^0).$$

This problem is discretized with planewaves. The discrete space is denoted by X_{N_c} and the orthogonal projector on X_{N_c} by Π_{N_c}. The Euler equations of the discrete minimization problem can be written as: find $(\phi_{i,N_\mathrm{c}}, \lambda_{i,N_\mathrm{c}})_{i=1,\dots,N}$ satisfying

$$\forall i = 1, \dots, N, \quad \mathcal{H}_{N_\mathrm{c},\mathrm{proj}}\,\phi_{i,N_\mathrm{c}} = \lambda_{i,N_\mathrm{c}}\,\phi_{i,N_\mathrm{c}}, \quad \text{and} \quad \forall i, j = 1, \dots, N, \quad \langle \phi_{i,N_\mathrm{c}} | \phi_{j,N_\mathrm{c}} \rangle = \delta_{ij},$$

$\lambda_{1,N_c} \leq \lambda_{2,N_c} \leq \ldots \leq \lambda_{N,N_c}$. Here we define $\mathcal{H}_{N_c,\mathrm{proj}} : X_{N_c} \to X_{N_c}$ by

$$\mathcal{H}_{N_c,\mathrm{proj}} = \Pi_{N_c}\mathcal{H}_{[\rho_{N_c}]}\Pi_{N_c} = -\frac{1}{2}\Pi_{N_c}\Delta\Pi_{N_c} + \Pi_{N_c}\Big[V_{\mathrm{ion}} + V_{\mathrm{coul}}(\rho_{N_c}) + V_{\mathrm{xc}}(\rho_{N_c})\Big]\Pi_{N_c}, \quad (2.5.14)$$

with $\rho_{N_c} = \rho_{[\Phi_{N_c}]}$, $\Phi_{N_c} = (\phi_{1,N_c}, \ldots, \phi_{N,N_c})^T$ and where $\mathcal{H}_{[\rho_{N_c}]}$ is defined by (2.5.11) for the approximate ground state density ρ_{N_c}. The discrete ground-state density matrix denoted by γ_{0,N_c} is defined as

$$\gamma_{0,N_c} = \sum_{i=1}^{N} |\phi_{i,N_c}\rangle\langle\phi_{i,N_c}|, \quad (2.5.15)$$

and the corresponding ground-state energy is defined as

$$\mathcal{E}_{0,N_c} = \mathcal{E}_0^{\mathrm{KS}}(\Phi_{N_c}).$$

As in the linear case, the eigenvalues of the Laplace operator are explicitly known on $X_{N_c}^{\perp}$. Therefore, as soon as λ_{N,N_c}, the N^{th} eigenvalue of the operator $\mathcal{H}_{N_c,\mathrm{proj}}$ defined in (2.5.14), is smaller than the lowest eigenvalue of the operator $-\frac{1}{2}\Delta$ on $X_{N_c}^{\perp}$, then Φ_{N_c} is also the ground-state of the following Kohn–Sham problem

$$\forall i = 1, \ldots, N, \quad \mathcal{H}_{N_c}\,\phi_{i,N_c} = \lambda_{i,N_c}\,\phi_{i,N_c}, \qquad \text{and} \qquad \forall i,j = 1, \ldots, N, \quad \langle\phi_{i,N_c}|\phi_{j,N_c}\rangle = \delta_{ij}, \quad (2.5.16)$$

$\lambda_{1,N_c} \leq \lambda_{2,N_c} \leq \ldots \leq \lambda_{N,N_c}$, where

$$\mathcal{H}_{N_c} = -\frac{1}{2}\Delta + \Pi_{N_c}\Big[V_{\mathrm{ion}} + V_{\mathrm{coul}}(\rho_{N_c}) + V_{\mathrm{xc}}(\rho_{N_c})\Big]\Pi_{N_c}.$$

Therefore the exact solution $(\phi_i^0, \lambda_i^0)_{j=1,\ldots,N}$ satisfies

$$(\mathcal{H}_{N_c} + \mathcal{V}_{N_c}^{\perp} + \mathcal{W}_{N_c})\,\phi_i^0 = \lambda_i^0\,\phi_i^0, \qquad \int_\Omega \phi_i^0\,\phi_j^0 = \delta_{ij},$$

where

$$\mathcal{V}_{N_c}^{\perp} = [V_{\mathrm{ion}} + V_{\mathrm{coul}}(\rho_{N_c}) + V_{\mathrm{xc}}(\rho_{N_c})] - \Pi_{N_c}\Big[V_{\mathrm{ion}} + V_{\mathrm{coul}}(\rho_{N_c}) + V_{\mathrm{xc}}(\rho_{N_c})\Big]\Pi_{N_c}, \quad (2.5.17)$$

and

$$\mathcal{W}_{N_c} = [V_{\mathrm{coul}}(\rho^0) + V_{\mathrm{xc}}(\rho^0)] - [V_{\mathrm{coul}}(\rho_{N_c}) + V_{\mathrm{xc}}(\rho_{N_c})], \quad (2.5.18)$$

whereas the approximate solution $(\phi_{i,N_c}, \lambda_{i,N_c})_{i=1,\ldots,N}$ satisfies (2.5.16), at least if N_c is large enough. The computation of the correction using Rayleigh–Schrödinger perturbation theory can be done exactly as in (2.5.4). The corrections should therefore be defined as

$$\phi_{i,N_c}^{(1)} = \sum_{j\neq i} \frac{\langle\phi_{j,N_c}, \big(\langle\phi_{i,N_c}, \mathcal{W}_{N_c}\phi_{i,N_c}\rangle - (\mathcal{V}_{N_c}^{\perp} + \mathcal{W}_{N_c})\big)\,\phi_{i,N_c}}{\lambda_{j,N_c} - \lambda_{i,N_c}}\phi_{j,N_c}.$$

Unfortunately, $\mathcal{W}_{N_c}$ contains the exact density ρ^0 and is not computable. But the contributions containing $\mathcal{W}_{N_c}$ can be shown *a priori* small. Hence, we neglect these terms and finally define the corrections on the eigenfunctions as

$$\phi_{i,N_c}^{(1)} = -\sum_{j\neq i} \frac{\langle\phi_{j,N_c}, \mathcal{V}_{N_c}^{\perp}\phi_{i,N_c}\rangle}{\lambda_{j,N_c} - \lambda_{i,N_c}}\phi_{j,N_c}.$$

Then, exactly as in the linear case, this can be simplified as

$$\phi^{(1)}_{i,N_c} = -(-\frac{1}{2}\Delta - \lambda_{i,N_c})^{-1}\mathcal{V}^{\perp}_{N_c}\phi_{i,N_c} = -(-\frac{1}{2}\Delta - \lambda_{i,N_c})^{-1}\mathrm{Res}_{i,N_c},$$

where the residual Res_{i,N_c} is defined by

$$\mathrm{Res}_{i,N_c} = -\frac{1}{2}\Delta\phi_{i,N_c} + \Big[V_{\mathrm{ion}} + V_{\mathrm{coul}}(\rho_{N_c}) + V_{\mathrm{xc}}(\rho_{N_c})\Big]\phi_{i,N_c} - \lambda_{i,N_c}\phi_{i,N_c}.$$

From this definition of the corrections, one can define post-processed orbitals, ground-state density matrix, and energy as in Definition 2.5.5 below. Since the post-processed orbitals are not *a priori* orthonormalized, we define two sets of post-processed orbitals, the second one being the orthonormalized version of the first one. This leads to the definition of two different post-processed density matrices and energies.

Definition 2.5.5 (Perturbed eigenvectors, density matrix and energy)**.** *For all $i = 1, \dots, N$, we define the perturbed eigenvectors as*

$$\widetilde{\phi_{i,N_c}} = \phi_{i,N_c} + \phi^{(1)}_{i,N_c}.$$

We also define orthonormal perturbed eigenvectors as an orthonormalization of $(\widetilde{\phi_{i,N_c}})_{i=1,\dots,N}$. More precisely, for all $i = 1, \dots, N$, define

$$\widetilde{\widetilde{\Phi_{N_c}}} = {S_{N_c}}^{-1/2}\widetilde{\Phi_{N_c}},$$

where S_{N_c}, the $N \times N$ overlap matrix of $\widetilde{\Phi_{N_c}} = (\widetilde{\phi_{1,N_c}}, \dots, \widetilde{\phi_{N,N_c}})$, is defined as

$$\forall i, j = 1, \dots, N, \quad (S_{N_c})_{i,j} = \langle\widetilde{\phi_{i,N_c}}|\widetilde{\phi_{j,N_c}}\rangle. \tag{2.5.19}$$

We define the perturbed density matrix as

$$\boxed{\widetilde{\gamma_{N_c}} = \sum_{i=1}^{N} |\widetilde{\phi_{i,N_c}}\rangle\langle\widetilde{\phi_{i,N_c}}| = \gamma_{0,N_c} + \gamma^{(1)}_{N_c} + \sum_{i=1}^{N} |\phi^{(1)}_{i,N_c}\rangle\langle\phi^{(1)}_{i,N_c}|,} \tag{2.5.20}$$

where

$$\gamma^{(1)}_{N_c} = \sum_{i=1}^{N} |\phi^{(1,1)}_{i,N_c}\rangle\langle\phi_{i,N_c}| + \sum_{i=1}^{N} |\phi_{i,N_c}\rangle\langle\phi^{(1)}_{i,N_c}|. \tag{2.5.21}$$

We also define a orthonormalized perturbed density matrix as

$$\boxed{\widetilde{\widetilde{\gamma_{N_c}}} = \sum_{i=1}^{N} |\widetilde{\widetilde{\phi_{i,N_c}}}\rangle\langle\widetilde{\widetilde{\phi_{i,N_c}}}|.} \tag{2.5.22}$$

We define respectively the perturbed energy and the orthonormalized perturbed energy as

$$\widetilde{\mathcal{E}_{0,N_c}} = \mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\widetilde{\Phi_{N_c}}), \qquad \widetilde{\widetilde{\mathcal{E}_{0,N_c}}} = \mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\widetilde{\widetilde{\Phi_{N_c}}}). \tag{2.5.23}$$

As in the linear case, we show that asymptotically, the post-processed density-matrices and energies are more precise than the approximate quantities on the coarse grid. This is presented in the following theorem, for which the proofs are presented together with Theorem 9.5.1. We gain a factor N_c^{-2} for all considered quantities except for the energy computed from the post-processed normalized orbitals, for which we gain a factor N_c^{-4}. The results partly rely on the *a priori* estimation performed in [44], which has been translated in the density matrix formalism to be used in this context.

Theorem 2.5.6 (Improved convergence for the density matrix and the energy)**.** *There exist* $C \in \mathbb{R}^+$ *and* $N_{\mathrm{c}}^0 \in \mathbb{N}$ *such that for* $N_{\mathrm{c}} \geq N_{\mathrm{c}}^0$,

$$\|(1-\Delta)^{1/2}(\gamma_0 - \widetilde{\gamma_{N_{\mathrm{c}}}})\|_{\mathfrak{S}_2(L^2_\#)} \leq C N_{\mathrm{c}}^{-2} \|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_{\mathrm{c}}})\|_{\mathfrak{S}_2(L^2_\#)}, \tag{2.5.24}$$

$$\|(1-\Delta)^{1/2}(\gamma_0 - \widetilde{\widetilde{\gamma_{N_{\mathrm{c}}}}})\|_{\mathfrak{S}_2(L^2_\#)} \leq C N_{\mathrm{c}}^{-2} \|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_{\mathrm{c}}})\|_{\mathfrak{S}_2(L^2_\#)}, \tag{2.5.25}$$

$$|\widetilde{\mathcal{E}_{0,N_{\mathrm{c}}}} - \mathcal{E}_0| \leq C N_{\mathrm{c}}^{-2} |\mathcal{E}_{0,N_{\mathrm{c}}} - \mathcal{E}_0|, \tag{2.5.26}$$

and

$$|\widetilde{\widetilde{\mathcal{E}_{0,N_{\mathrm{c}}}}} - \mathcal{E}_0| \leq C N_{\mathrm{c}}^{-4} |\mathcal{E}_{0,N_{\mathrm{c}}} - \mathcal{E}_0|. \tag{2.5.27}$$

Once again, the improvement on the density matrix is provided in terms of Hilbert–Schmidt norm, which can be related to H^1-norms of the orbital error in the case of simple eigenvalues.

Together with the theoretical results, numerical results were performed with KSSOLV [252], a Matlab toolbox for solving the Kohn–Sham equations. They are presented in Section 7.5. We observe an improvement of about a factor 10 to 50 for the energy of simple molecules, such as the alanine molecule, which contains 18 valence electrons. The cost of the post-processing is very limited, and is about 3% over the different calculations. Note that in practice, it it not possible to compute the corrections on the whole space, hence they are computed on a fine grid. Details are provided in Chapter 7.

In this contribution, the post-processing method is presented as a way to improve the accuracy of the ground-state energy and density matrix at a low computational cost. However, since the post-processed quantities are more precise than the coarse ones, they could be used as error indicators for the coarse solution. In this framework, the energy error could be estimated by the difference between the post-processed energy and the coarse energy, and respectively for the density matrix error. Thus, this method does not offer a proper guaranteed *a posteriori* estimation for the Kohn–Sham equations, but can still be very useful for the estimation of the energy and density-matrix errors.

2.6 A posteriori estimation and post-processing methods in a unified framework

In the previous sections, we have presented different methods for the *a posteriori* error estimation and post-processing of linear and nonlinear eigenvalue problems. In the following, we compare these different methods in a unified framework. This section is a preliminary version of [99]. First, we relate the error estimation for the Gross–Pitaevskii equation presented in Chapter 6 to a Taylor expansion of the residual. We then present the *a posteriori* estimation for the Hartree–Fock problem derived in [184]. Finally, we mention the two-grid methods proposed in [251] for the linear case, and [42] for the nonlinear case, as well as post-processing methods based on perturbation theory [49, 50].

2.6.1 Generic linear and nonlinear eigenvalue problems

In order to compare the different methods more easily, we adopt the same notations for the eigenvalues problems. Let us first consider an abstract *linear* eigenvalue problem. Let X be a separable Hilbert space endowed with a scalar product denoted by $\langle \cdot, \cdot \rangle$, and corresponding norm $\| \cdot \|$. Let A be a bounded-below self-adjoint operator on X with domain $D(A)$ and compact resolvent. There exists a non-decreasing sequence of real numbers $(\lambda_k)_{k \geq 1}$ such that $\lambda_k \to \infty$ and an orthonormal basis $(u_k)_{k \geq 1}$ of $\mathcal{H}$ consisting of vectors of $D(A)$ such that

$$\forall k \geq 1, \quad A\, u_k = \lambda_k u_k.$$

For simplicity, we restrict ourselves here to the problem of finding the lowest eigenvalue of the problem, even though the post-processing methods presented in [49, 50] and the *a posteriori* estimation [184] are not restricted to the lowest eigenvalue. The problem can be stated as: Find $(u, \lambda) \in X \times \mathbb{R}$, such that $\|u\| = 1$ and

$$Au = \lambda u,$$

or written in a weak form

$$\forall v \in X, \quad \langle Au, v\rangle = \lambda \langle u, v\rangle.$$

We consider here a Galerkin discretization of this equation using planewaves. The approximate space X_{N_c} is characterized by a cutoff N_c in momentum space. The discrete solution $(u_{N_c}, \lambda_{N_c}) \in X_{N_c} \times \mathbb{R}$ satisfies $\|u_{N_c}\| = 1$ and

$$\forall v_{N_c} \in X_{N_c}, \quad \langle Au_{N_c}, v_{N_c}\rangle = \lambda_{N_c} \langle u_{N_c}, v_{N_c}\rangle. \tag{2.6.1}$$

This problem is called linear eigenvalue problem because the operator A itself does not depend on the eigenvector. However, strictly speaking, this problem is nonlinear in the solution (u, λ).

Let us now consider an abstract *nonlinear* eigenvalue problem, i.e. a problem where the operator A depends on the eigenvector. We characterize this dependency by a function f. In our applications, this operator f depends on the electronic density ρ in the Hartree–Fock and Kohn–Sham problems, which is equal to u^2 in the case of only one eigenfunction. Therefore, we denote by $f(u^2)$ the nonlinear part of the operator. Thus, we consider the following problem: Find $(u, \lambda) \in X \times \mathbb{R}$, such that $\|u\| = 1$ and

$$Au + f(u^2)u = \lambda u, \tag{2.6.2}$$

where A is a linear operator independent of u having the properties described in the linear case. A Galerkin discretization of this problem writes: Find $(u_{N_c}, \lambda_{N_c}) \in X_{N_c} \times \mathbb{R}$ with $\|u_{N_c}\| = 1$ such that

$$\forall v_{N_c} \in X_{N_c}, \quad \langle \left(A + f(u_{N_c}^2)\right) u_{N_c}, v_{N_c}\rangle = \lambda_{N_c} \langle u_{N_c}, v_{N_c}\rangle. \tag{2.6.3}$$

As before, we choose to present the methods with a planewave discretization.

2.6.2 Some convergence results

Let us now recall some *a priori* results on the convergence of the eigenfunctions and eigenvalues of the problems presented above. First, under suitable and realistic hypotheses on A, there holds

$$\|u - u_{N_c}\|_{H^1} \lesssim \min_{v_{N_c} \in X_{N_c}} \|u - v_{N_c}\|_{H^1}.$$

Also, the eigenvalues converge quadratically compared to the eigenvectors, i.e.

$$|\lambda_{N_c} - \lambda| \lesssim \|A^{1/2}(u - u_{N_c})\|^2,$$

and in the case where $A = -\Delta + V$ (see e.g. [43]),

$$|\lambda_{N_c} - \lambda| \lesssim \|u - u_{N_c}\|_{H^1}^2.$$

Moreover, the L^2-norm of the error $u - u_{N_c}$ as well as negative Sobolev norms converge faster than the H^1-norm of the error. Finally, the dual norm of the residual behaves like the H^1-norm of the error, i.e.

$$\|Au_{N_c} - \lambda_{N_c} u_{N_c}\|_{H^{-1}} \lesssim \|u - u_{N_c}\|_{H^1}.$$

This is not trivial as the problem is an eigenvalue problem, but it was shown for the Gross–Pitaevskii equation in [98], and for the Laplace eigenvalue problem in [52].

2.6.3 A posteriori estimation for the Gross–Pitaevskii equation

In the *a posteriori* estimation for the Gross–Pitaevskii equation presented in Chapter 6, which is nonlinear, the first error bound relies on Caloz–Rappaz theory [40], with a Taylor development of the residual. The equation corresponds to (2.6.2) with $A = -\Delta + V$ and f being the identity.

Let us denote the residual by F, i.e.

$$\forall v \in X,\ \forall \mu \in \mathbb{R}, \quad F(v,\mu) = Av + f(v^2)v - \mu v.$$

For the exact eigenpair $(u,\lambda) \in X \times \mathbb{R}$, there holds

$$F(u,\lambda) = Au + f(u^2)u - \lambda u = 0.$$

Writing a first-order development of the residual around the coarse eigenpair $(u_{N_c}, \lambda_{N_c}) \in X_{N_c} \times \mathbb{R}$, where X_{N_c} is a planewave space, gives

$$F(u,\lambda) = F(u_{N_c},\lambda_{N_c}) + DF_{(u_{N_c},\lambda_{N_c})}(u - u_{N_c}, \lambda - \lambda_{N_c}) + h.o.t., \tag{2.6.4}$$

where *h.o.t.* means higher order terms, and DF is the differential of the residual F, and is defined by

$$\forall w \in X, \tau \in \mathbb{R}, \quad DF_{(u_{N_c},\lambda_{N_c})}(w,\tau) = \begin{pmatrix} Aw + 2f'(u_{N_c}^2)u_{N_c}^2 w + f(u_{N_c}^2)w - \tau u_{N_c} - \lambda_{N_c} w \\ 2\int u_{N_c} w \end{pmatrix}$$

Loosely speaking, if the approximate solution (u_{N_c}, λ_{N_c}) is close to the exact solution (u,λ), there holds

$$0 \simeq F(u_{N_c},\lambda_{N_c}) + DF_{(u_{N_c},\lambda_{N_c})}(u - u_{N_c}, \lambda - \lambda_{N_c}).$$

Therefore, if DF is invertible,

$$\begin{pmatrix} u - u_{N_c} \\ \lambda - \lambda_{N_c} \end{pmatrix} \simeq DF^{-1}_{(u_{N_c},\lambda_{N_c})} F(u_{N_c},\lambda_{N_c}).$$

In order to get a guaranteed bound of the error in Chapter 6, Caloz–Rappaz theorem provides conditions ensuring that there exists an exact solution in the vicinity of the approximate solution, and that the error between the exact and approximate solution is bounded by

$$\|u - u_{N_c}\|_{H^1} + |\lambda - \lambda_{N_c}| \leq 2\|DF^{-1}_{(u_{N_c},\lambda_{N_c})}\|_{H^{-1},H^1} \|F(u_{N_c},\lambda_{N_c})\|_{H^{-1}}.$$

The factor 2 in this estimation allows to absorb the higher order terms of the Taylor development. In the *a posteriori* estimation of Chapter 6, an important part of the contribution consists of showing that DF is invertible at (u_{N_c}, λ_{N_c}) and to bound the norm of its inverse. Indeed, the main part of DF^{-1} is $(\Delta)^{-1}$, that is an isometry between H^{-1} and H^1, and the remaining part in DF^{-1} is of lower order in terms of differential operator. In the refined bound presented in Section 6.3.2, the term $DF^{-1}_{(u_{N_c},\lambda_{N_c})} F(u_{N_c},\lambda_{N_c})$ is directly estimated, and these lower order terms are shown to be small thanks to the first bound. Asymptotically, we obtain

$$\|u - u_{N_c}\|_{H^1} \leq \alpha \|F(u_{N_c},\lambda_{N_c})\|_{H^{-1}},$$

where α can be taken as close to 1 as we wish when N_c goes to infinity.

2.6.4 A posteriori estimation for the Hartree–Fock problem

The *a posteriori* estimation for the Hartree–Fock problem [184] is mainly performed on the energy, but also offers a convergence rate for the orbitals. We refer to Section 1.3.1 for the description of this nonlinear eigenvalue problem. The *a posteriori* estimation is based on a post-processing of the coarse solution (u_{N_c}, λ_{N_c}) on a fine space X_{N_f} at a cheap computational cost. Hence, this method can be seen as a two-grid method, as we will see in the following subsection. Here N_f denotes the cutoff in momentum space for the fine planewave space, and $N_f > N_c$.

The post-processed orbitals are defined from a second-order Taylor development on the energy, which corresponds to a first-order development on the residual. Indeed, translated in this formalism, neglecting the nonlinearity in the operator, the fine-grid problem writes: Find $u_{N_f} \in X_{N_f}$ such that

$$(A - \lambda_{N_c})_{|X_{N_c}^{\perp}} u_{N_f} = -(A - \lambda_{N_c}) u_{N_c} \quad \text{in} \quad X_{N_f}. \tag{2.6.5}$$

Since the convergence of the eigenvalues is quadratic compared to the convergence of the H^1-norm of the orbital error, we can neglect the dependency of the residual in the eigenvalue, and rather consider the residual as a function of u defined by:

$$F(u) = (A - \lambda_{N_c})u.$$

The differential of F at u writes

$$\forall v \in X, \quad DF_u(v) = (A - \lambda_{N_c})v.$$

Therefore, the equation (2.6.5) is equivalent to a Newton step with respect to F which can be written as: Find $u_{N_f} \in X_{N_f}$ such that

$$\forall v_{N_f} \in X_{N_f}, \quad \langle F(u_{N_c}), v_{N_f} \rangle + \langle DF_{u_{N_c}}(u_{N_f} - u_{N_c}), v_{N_f} \rangle = 0. \tag{2.6.6}$$

If we consider now the nonlinear problem (2.6.3), $F(u)$ is defined as

$$F(u) = (A + f(u^2) - \lambda_{N_c})u.$$

In this case, the differential DF of F at u writes

$$\forall v \in X, \quad DF_u(v) = (A + 2f'(u^2)u^2 + f(u^2) - \lambda_{N_c})v,$$

and the corresponding Newton step in the fine grid reads in a strong form: Find $u_{N_f} \in X_{N_f}$ such that

$$(A + 2f'(u_{N_c}^2)u_{N_c}^2 + f(u_{N_c}^2) - \lambda_{N_c})_{|X_{N_c}^{\perp}} u_{N_f} = -(A + f(u_{N_c}^2) - \lambda_{N_c})u_{N_c} \quad \text{in} \quad X_{N_f}. \tag{2.6.7}$$

Seen as a Newton step, this proposed post-processing naturally doubles the convergence rate of the energy and the H^1-norm of the orbitals. Then, the difference between the post-processed energy and the approximate energy allows to estimate the error between the exact energy and the approximate one. Compared to the previous *a posteriori* estimation, the bounds here are not guaranteed. Nevertheless, they converge at the right speed to the true error.

Numerically, this method requires to solve a boundary value problem. Note that the computation of DF might not be straightforward, especially since the Hartree–Fock problem is nonlinear. Moreover, in order to get a doubled convergence rate on the energy, the post-processed eigenfunctions need to be normalized, which requires the inversion of a (relatively small) matrix.

2.6.5 Two-grid methods

Let us now present the two-grid methods proposed in [251, 42]. In two-grid methods, the goal is to compute a better approximation of a coarse solution (u_{N_c}, λ_{N_c}) on a fine space X_{N_f} at a cheap computational cost. In fact, these methods can be seen as an approximate Newton step for the residual equation, hence can be compared to the post-processing of the Hartree–Fock equations presented above.

In the two-grid method for a linear eigenvalue problem presented in [251], the eigenvalue problem (2.6.1) is solved on the coarse space X_{N_c}, and then, the following boundary problem is solved on a fine space X_{N_f}: Find $(u_{N_f}, \lambda_{N_f}) \in X_{N_f} \times \mathbb{R}$ such that

$$\forall v_{N_f} \in X_{N_f}, \quad \langle A u_{N_f}, v_{N_f} \rangle = \lambda_{N_c} \langle u_{N_c}, v_{N_f} \rangle. \tag{2.6.8}$$

The fine eigenvalue is then defined as a Rayleigh quotient of the fine solution, i.e.

$$\lambda_{N_f} = \frac{\langle u_{N_f}, A u_{N_f} \rangle}{\langle u_{N_f}, u_{N_f} \rangle}.$$

In [251], the operator A is the Laplace operator, and the problem is discretized with finite elements. Note that in equation (2.6.8), the residual $(A - \lambda_{N_c}) u_{N_c}$ does not appear explicitly. However, substracting $\langle A u_{N_c}, v_{N_f} \rangle$ on both side of equation (2.6.8), we obtain

$$\forall v_{N_f} \in X_{N_f}, \quad \langle A(u_{N_f} - u_{N_c}), v_{N_f} \rangle = -\langle (A - \lambda_{N_c}) u_{N_c}, v_{N_f} \rangle, \tag{2.6.9}$$

where the residual $(A - \lambda_{N_c})$ is present on the right hand-side. The quantity $u_{N_f} - u_{N_c}$ now appears as a correction on the eigenfunction u_{N_c}.

We can also decompose the residual of the fine-grid solution in different terms in order to include the fine-grid equation. We obtain

$$A u_{N_f} - \lambda_{N_f} u_{N_f} = A u_{N_f} - \lambda_{N_c} u_{N_c} + \lambda_{N_c}(u_{N_c} - u_{N_f}) + (\lambda_{N_c} - \lambda_{N_f}) u_{N_f}, \tag{2.6.10}$$

which is very close to the first-order Taylor development of the residual (2.6.4). The dual norm of the fine residual $\|A u_{N_f} - \lambda_{N_f} u_{N_f}\|_{H^{-1}}$ is composed of the residual of the equation solved on the fine grid $\|A u_{N_f} - \lambda_{N_c} u_{N_c}\|_{H^{-1}}$, which is very small, the dual norm of the error $\|u_{N_c} - u_{N_f}\|_{H^{-1}}$ and the error on the eigenvalue $|\lambda_{N_c} - \lambda_{N_f}|$, which are both of higher order compared to the dual norm of the residual $\|A u_{N_c} - \lambda_{N_c} u_{N_c}\|_{H^{-1}}$. This explains the gain in accuracy for the fine-grid solution. Indeed, if the discretization parameters are well-chosen,

$$\|u - u_{N_f}\|_{H^1} \lesssim \|u - u_{N_c}\|_{H^1}^2, \quad \text{and} \quad |\lambda - \lambda_{N_f}| \lesssim |\lambda - \lambda_{N_c}|^2. \tag{2.6.11}$$

The cost of this method is the resolution of a boundary problem on the fine grid, i.e. (2.6.8).

In fact, the equation solved in the fine grid in this two-grid method is similar to the Newton step presented for the Hartree–Fock problem (2.6.5), except that the term $-\lambda_{N_c}(u_{N_f} - u_{N_c})$ is neglected. But this term is of higher order compared to $A(u_{N_f} - u_{N_c})$, which explains why the improved convergence results are similar.

In the two-grid method for a nonlinear eigenvalue problem presented in [42], the problem solved on the coarse grid is: Find $(u_{N_c}, \lambda_{N_c}) \in X_{N_c} \times \mathbb{R}$ such that

$$(A + f(u_{N_c}^2)) u_{N_c} = \lambda_{N_c} u_{N_c}, \quad \|u_{N_c}\|_{L^2} = 1, \quad \text{on} \quad X_{N_c}.$$

The operator A is $-\Delta + V$ and the discretization can be finite element or a planewave methods. Then three different problems on the fine grid X_{N_f} are proposed:

1- Solve a linear eigenvalue problem on the fine grid:

$$(A + f(u_{N_c}^2))u_{N_f} = \lambda_{N_f} u_{N_f}.$$

2- Solve a linear boundary problem on the fine grid:

$$(A + f(u_{N_c}^2))u_{N_f} = \lambda_{N_c} u_{N_c}.$$

3- Solve a linear boundary problem on the fine grid:

$$Au_{N_f} = -f(u_{N_c}^2)u_{N_c} + \lambda_{N_c} u_{N_c}.$$

An error analysis is provided for the first method. If the ratio between N_c and N_f is well-chosen, an improvement is expected on the eigenfunction and the eigenvalue, but not as much as in the Hartree–Fock case. Let us remark that compared to (2.6.7), these equations neglect the derivative of the functional f. This might explain why the convergence rates of the eigenfunctions and the eigenvalues are not doubled.

2.6.6 Post-processing methods based on perturbation theory

In the perturbative method [49], the idea is a bit different. Let us first consider the linear case. On the coarse grid X_{N_c}, we solve the problem (2.6.1), but we write it as a problem on the whole space with a modified operator A_{N_c}: the problem (2.6.1) corresponds to the problem

$$A_{N_c} u_{N_c} = \lambda_{N_c} u_{N_c}, \quad \text{on} \quad X,$$

or written in the weak form as

$$\forall v \in X, \quad \langle A_{N_c} u_{N_c}, v\rangle = \lambda_{N_c} \langle u_{N_c}, v\rangle.$$

In the particular case presented in [49], $A = -\Delta + V$, with V a multiplicative potential, X_{N_c} is a planewave discretization space, and $A_{N_c} = -\Delta + \Pi_{N_c} V \Pi_{N_c}$, with Π_{N_c} the projector on the space X_{N_c} for any H^s scalar product ($s \in \mathbb{R}$). Then, a perturbative expansion is performed, taking the previous equation as the unperturbed equation and the equation $Au = \lambda u$ as the perturbed one.

Denoting in this case respectively u_{N_f} and λ_{N_f} the perturbed eigenfunction and eigenvalue, we obtain at first order

$$(A - A_{N_c})u_{N_c} + A_{N_c}(u_{N_f} - u_{N_c}) = \lambda_{N_c}(u_{N_f} - u_{N_c}) + (\lambda_{N_f} - \lambda_{N_c})u_{N_c},$$

i.e.

$$(A_{N_c} - \lambda_{N_c})(u_{N_f} - u_{N_c}) - (\lambda_{N_f} - \lambda_{N_c})u_{N_c} = -(A - \lambda_{N_c})u_{N_c}.$$

It can be shown that the correction for the eigenvalue at first order is equal to zero, so that

$$(A_{N_c} - \lambda_{N_c})(u_{N_f} - u_{N_c}) = -(A - \lambda_{N_c})u_{N_c}. \tag{2.6.12}$$

The discretization being done in planewave, the operator $(A_{N_c} - \lambda_{N_c})$ is diagonal on the orthogonal complement to X_{N_c}, and since $(A - \lambda_{N_c})u_{N_c} \in X_{N_c}^{\perp}$, this equation can be easily solved. The corrected eigenvalue λ_{N_f} is computed with a second-order development.

If we write a development similar to (2.6.10), we obtain

$$Au_{N_f} - \lambda_{N_f} u_{N_f} = (A - \lambda_{N_c})u_{N_c} + (A_{N_c} - \lambda_{N_c})(u_{N_f} - u_{N_c}) - (\lambda_{N_f} - \lambda_{N_c})u_{N_f} + (A - A_{N_c})(u_{N_f} - u_{N_c}).$$

The first two terms in the right-hand side correspond to the perturbation equation (2.6.12). The third term is of higher order as it contains the eigenvalue difference $\lambda_{N_f} - \lambda_{N_c}$, and the last term is small only since $(A - A_{N_c})$ does not contain the Laplace operator any more and is equal to $V - \Pi_{N_c} V \Pi_{N_c}$, so that $u_{N_f} - u_{N_c}$ does not need to be measured in H^1-norm. In this case, we gain a factor N_c^{-2} both on the H^1-norm of the eigenvectors and on the eigenvalue.

The post-processing method presented in [50] in the nonlinear case is similar to the linear case. The uncomputable terms are shown to be asymptotically negligible. We refer to Chapter 9 for more details on this case. The gain on the eigenvector is similar to the linear case.

This perturbation method corresponds to a Taylor development of the residual, exactly as in (2.6.6) except that here, the residual is seen as a function of A, u, and λ:

$$F(A, u, \lambda) = (A - \lambda)u.$$

Thus, the perturbation method and the two grids method are similar, in the sense that the post-processed eigenvectors arise from a Taylor development where higher order terms are neglected. However, they are not equivalent. Indeed, the perturbation method does not require to solve a boundary value problem, since the operator to invert on the fine grid is diagonal. Moreover, the improvement of this perturbation method is limited to ${N_c}^{-2}$ by the difference in the operator $V - \Pi_{N_c} V \Pi_{N_c}$, whereas, in the two-grid case, the convergence rate of the eigenfunctions can be doubled, at least in the linear case.

The following table gives a summary of the different methods that were presented before, with their main characteristics.

Method	**Equation**	**Number of eigenvalues**	**Discretization method**	**Generic type of estimates**	**Cost**
Linear two-grid method [251]	$-\Delta u = \lambda u$	K lowest	Finite elements (FE)	$\|u - u_{N_f}\|_{H^1} \lesssim \|u - u_{N_c}\|_{H^1}^2$ $\|\lambda - \lambda_{N_f}\| \lesssim \|\lambda - \lambda_{N_c}\|^2$	Boundary Value Problem (BVP)
Nonlinear two-grid method [42]	$(-\Delta + V + f(u^2))u = \lambda u$	1 lowest	FE or planewaves (PW)	(in FE) if $\|u - u_{N_c}\|_{H^1} \lesssim {N_c}^{-\sigma}$, then $\|u - u_{N_f}\|_{H^1} \lesssim {N_c}^{-\sigma-2} + {N_f}^{-\sigma}$ $\|E(u) - E(u_{N_f})\| \lesssim {N_c}^{-2\sigma-4} + {N_f}^{-2\sigma}$	BVP or linear eigenvalue problem
Linear perturbation method [49]	$(-\Delta + V)u = \lambda u$	K lowest	PW	if N_{c} cutoff in momentum space, $\|u - u_{N_f}\|_{H^1} \lesssim N_{\mathrm{c}}^{-2}\|u - u_{N_c}\|_{H^1}$ $\|\lambda - \lambda_{N_f}\| \lesssim N_{\mathrm{c}}^{-2}\|\lambda - \lambda_{N_c}\|$	Residual computation
Nonlinear perturbation method [50]	$(-\Delta + V + V_{\rho_\Phi})\phi_i = \lambda_i\phi_i$ with $\Phi = (\phi_1, \ldots, \phi_K)$.	K lowest	PW	if N_{c} cutoff in momentum space, $\|u - u_{N_f}\|_{H^1} \lesssim N_{\mathrm{c}}^{-2}\|u - u_{N_c}\|_{H^1}$ Energy: $\|E(u) - E(u_{N_f})\| \lesssim N_{\mathrm{c}}^{-2}\|E(u) - E(u_{N_c})\|$	Residual computation
A posteriori estimation for Hartree–Fock [184]	$(-\Delta + V + (\rho_\Phi \star \frac{1}{\|x\|})\phi_i = \lambda_i\phi_i$, with $\Phi = (\phi_1, \ldots, \phi_K)$.	K lowest	Any	$\|u - u_{N_f}\|_{H^1} \lesssim \|u - u_{N_c}\|_{H^1}^2$ $\|E(u) - E(u_{N_f})\| \lesssim \|E(u) - E(u_{N_c})\|^2$	Resolution of a BVP

Table 2.2 – Comparative table of different post-processing methods.

2.7 Conclusion and perspectives

In this chapter, we have presented the different contributions of this thesis. We now present possible extensions of this work.

In the contribution presented in Section 2.2 and detailed in Part II, the phenomenon of discretization error cancellation is rigorously explained for an operator of the form $-\Delta + V$, where the potential V is a sum of Dirac potentials, hence not smooth. The potentials considered in molecular simulations being often more regular, e.g. with pseudopotentials, one can wonder whether the results apply in this case. In fact, the reason for choosing Dirac potentials in the mathematical model was that explicit solutions were available, and not directly a regularity issue. Hence, we expect similar results for smoother potentials, i.e. an improved prefactor on the energy error difference but no improvement in the convergence rate. This is also what we observed on the simulations for the H_2 molecule. To go beyond the discretization error cancellation, it would be interesting to study other types of error cancellation, such as the error in the supercell size. Such estimations would allow to guaranty the accuracy of differences of quantities of interest, which are in many cases the true quantities of interest, and possibly decrease the numerical cost of the simulations at the same time.

In Part III, the *a posteriori* analysis for the Laplace operator leading to fully computable, guaranteed and efficient error bounds for the eigenvalues and eigenvectors is valid for any simple eigenvalue. In practice, eigenvalues are often multiple or close to degenerate, in which case our analysis does not hold anymore. However, it is possible to provide similar results for clusters of eigenvalues. This is work in progress. The proofs have to be extended within the framework of density matrices, as in Chapters 8 and 9. The resulting assumption is the existence of a gap below and above the cluster of eigenvalues. This *a posteriori* analysis could also be generalized by considering a more involved operator, for example an operator of the form $-\Delta + V$, with a (possibly nonlinear) potential V, in order to get closer to a fully, computable, robust *a posteriori* analysis for the finite element discretization of the Kohn–Sham model. To this aim, the generic equivalence part has already been generalized to a class of bounded-below self-adjoint operators with compact resolvent. However, the reconstruction part based on equilibrated fluxed and dealing with the nonlinearity seem quite challenging.

Concerning the Gross–Pitaevskii equation presented in Part IV, a possible extension of our contribution on the *a posteriori* analysis for this equation would be to add a rotating magnetic field to the system, in order to model rotating Bose–Einstein condensates. Two main difficulties arise in this case. First, the energy of the problem is not convex anymore. Second, the magnetic field introduces a coupling between the real and imaginary parts of the solution, which cannot be considered real anymore. On top of that, the problem is then 3-dimensional, whereas our *a posteriori* analysis of the Gross–Pitaevskii equation has been simplified in a one-dimensional case.

For the Kohn–Sham model, we have not yet proposed a guaranteed *a posteriori* error bound. Instead, we have proposed a post-processing method which improves the accuracy of the energy and the orbitals at a low computational cost, and can be used to derive error indicators, as presented in Section 2.5.3. In this direction, one possibility would be to modify the post-processing method to guaranty a post-processed energy always above (or below) the exact energy. The post-processed energy could then be used as an upper (or lower) bound of the energy. This is actually work in progress, following the works of Feshbach–Schur and Löwdin. However, the non-convexity of the model has also to be dealt with. In a first step, asymptotic bounds, following the ideas of the second bound for the Gross–Pitaevskii equation presented in Chapter 6 could be derived.

A such *a posteriori* bound for the Kohn–Sham model could then be used to perform error separation and error balance. For the Gross–Pitaevskii equation, the computational gain ob-

tained through the error balance presented in Chapter 6 is somehow limited, especially in the one-dimensional case, as the cost of the whole computation is anyway not expensive. However, solving the Kohn–Sham equations for large systems is expensive, therefore we expect a larger computational cost reduction using error balance in this case.

At this point, one should note that this separation of error and error balance procedure may complicate the computation of the forces. Indeed, when the discretization and iteration errors are balanced, the algorithm presented in Figure 2.2 stops while the self-consistent field algorithm has not fully converged. It is exactly how we intend to spare computational resources. Unfortunately, in molecular dynamics, the full convergence of the SCF cycle is used to efficiently compute the forces thanks to Hellmann–Feynman theorem. This problem also appears in the post-processing method. For the moment, Hellmann–Feynman theorem does not hold for the post-processed quantities, which slows down the computation of the forces. Therefore, it would be interesting to either derive a strategy in which Hellmann–Feynman theorem can still be applied, or develop a cheap computation of the forces in the case where the orbitals are not fully converged. Ideas in this direction can be found in [13].

Up to now, the numerical results have only been performed with KSSOLV [252], a Matlab toolbox running for small molecules, mainly for a testing purpose. In the future, it would be interesting to implement guaranteed error bounds for the Kohn–Sham models in largely used quantum chemistry codes, such as Abinit [121] in the case of planewaves. This would make guaranteed computations with error bars available for the practitioners.

This thesis focuses on two different discretization methods: the finite elements and the planewaves. In Section 1.5, we presented other discretization methods, which could be studied as well. In particular, the molecular orbitals are massively used by chemists for their good compromise between accuracy and number of basis functions. But this accuracy is limited, as no one knows which function to add to decrease the error below a certain point. Therefore, it would be great to combine the intuition of the chemists, which leads to a very good accuracy with only few basis functions with the possible automatic optimization given by an adaptive refinement procedure. Another possibility would be to focus on the wavelet method, which easily allows for adaptivity. Even though this adaptivity is in practice somehow limited, as there are only two precisions levels in one of the main wavelet codes called BigDFT [191], it would be already interesting to propose adaptive procedures for this specific case.

Finally, and probably further in the future, one could look at model errors, for example try to obtain error bounds for the Configuration Interaction or Coupled Cluster methods with respect to the solution of the Schrödinger equation. There already exist adaptive algorithms for these methods [181, 194, 222], but for the moment, the adaptive selection of the orbitals is based on *a priori* arguments, e.g. the smallness of the coefficients, whereas it could be based on guaranteed error bounds. However, a possibly major difficulty would be the computation or estimation of the residual. Anyway, such procedure would lead to a better selection of the orbitals, together with an error bound on the final result.

Part II

Discretization error cancellation for a linear eigenvalue problem

Chapter 3

Discretization error cancellation in electronic structure calculation: toward a quantitative study

We expose in this chapter the results of [47]. This work was done in collaboration with Eric Cancès.

Abstract

It is often claimed that error cancellation plays an essential role in quantum chemistry and first-principle simulation for condensed matter physics and materials science. Indeed, while the energy of a large, or even medium-size, molecular system cannot be estimated numerically within chemical accuracy (typically 1 kcal/mol or 1 mHa), it is considered that the energy difference between two configurations of the same system can be computed in practice within the desired accuracy.

The purpose of this paper is to initiate the quantitative study of discretization error cancellation. Discretization error is the error component due to the fact that the model used in the calculation (e.g. Kohn-Sham LDA) must be discretized in a finite basis set to be solved by a computer. We first report comprehensive numerical simulations performed with Abinit [120, 121] on two simple chemical systems, the hydrogen molecule on the one hand, and a system consisting of two oxygen atoms and four hydrogen atoms on the other hand. We observe that errors on energy differences are indeed significantly smaller than errors on energies, but that these two quantities asymptotically converge at the same rate when the energy cut-off goes to infinity. We then analyze a simple one-dimensional periodic Schrödinger equation with Dirac potentials, for which analytic solutions are available. This allows us to explain the discretization error cancellation phenomenon on this test case with quantitative mathematical arguments.

3.1 Introduction

Error control is a central issue in molecular simulation. The error between the computed value of a given physical observable (e.g. the dissociation energy of a molecule) and the exact one, has several origins. First, there is always a discrepancy between the physical reality and the reference model, here the N-body Schrödinger equation, possibly supplemented with Breit terms to account for relativistic effects. However, at least for the atoms of the first three rows of the periodic table, this reference model is in excellent agreement with experimental data, and can be considered as exact in most situations of interest. The overall error is therefore the sum of the following components:

1. the *model error*, that is the difference between the value of the observable for the reference model, which is too complicated to solve in most cases, and the value obtained with the chosen approximate model (e.g. the Kohn-Sham LDA model), assuming that the latter can be solved exactly;

2. the *discretization error*, that is the difference between the value of the observable for the approximate model and the value obtained with the chosen discretization of the approximate model. Indeed, the approximate model is typically an infinite dimensional minimization problem, or a system of partial differential equations, which must be discretized to be solvable by a computer, using e.g. a Gaussian atomic basis set, or a planewave basis;

3. the *algorithmic error*, which is the difference between the value of the observable obtained with the exact solution of the discretized approximate model, and the value computed with the chosen algorithm. The discretized approximate models are indeed never solved exactly; they are solved numerically by iterative algorithms (e.g. SCF algorithms, Newton methods), which, in the best case scenario, only converge in the limit of an infinite number of iterations. In practice, stopping criteria are used to exit the iteration loop when the error at iteration k, measured in terms of differences between two consecutive iterates or, better, by some norm of some residual, is below a prescribed threshold. If the stopping criterion is very tight, the algorithmic error can become very small, ... or not! For instance, if the discretized approximate model is a non convex optimization problem, there is no guarantee that the numerical algorithm will converge to a global minimum. It may converge to a local, non-global minimum, leading to a non-zero algorithmic error even in the limit of an infinitely tight stopping criterion;

4. the *implementation error*, which may, obviously, be due to bugs, but does not vanish in the absence of bugs, because of round-off errors: in molecular simulation packages, most operations are implemented in double precision, and the resulting round-off errors can accumulate, especially for very large systems;

5. the *computer error*, due to random hardware failures (miswritten or misread bits). This component of the error is usually negligible in today's standard computations, but is expected to become critical in future exascale architectures [168].

Quantifying these different sources of errors is an interesting purpose for two reasons. First, guaranteed estimates on these five components of the error would allow one to supplement the computed value of the observable returned by the numerical simulation with guaranteed error bars (certification of the result). Second, they would allow one to choose the parameters of the simulation (approximate model, discretization parameters, algorithm and stopping criteria, data structures, etc.) in an optimal way in order to minimize the computational effort required to reach the target accuracy.

The construction of guaranteed error estimators for electronic structure calculation is a very challenging task. Some progress has however been made in the last few years, regarding notably the discretization and algorithmic errors for Kohn-Sham LDA calculations. *A priori* discretization error estimates have been constructed in [44] for planewave basis sets, and then in [71] for more general variational discretization methods. *A posteriori* error estimators of the discretization error have been proposed in [50, 70, 147]. A combined study of both the discretization and algorithmic errors was published in [49] (see also [98]). We also refer to [184, 74, 75, 158, 173, 128, 157, 202, 203, 216] and references therein for other works on error analysis for electronic structure calculation.

In all the previous works on this topic we are aware of, the purpose was to estimate, *for a given nuclear configuration R of the system*, the difference between the ground state energy E_R (or another observable) obtained with the continuous approximate model under consideration (e.g. Kohn-Sham LDA) and its discretized counterpart denoted by $E_{R,N}$, where N is the discretization parameter. The latter is typically the number of basis functions in the basis set for local combination of atomic orbitals (LCAO) methods [135], the inverse fineness of the grid or the mesh for finite difference (FD) and finite element (FE) methods [125, 219, 198, 193], the cut-off parameter in energy or momentum space for planewave (PW) discretization methods [120, 114, 153], or the inverse grid spacing and the coarse and fine region multipliers for wavelet (WL) methods [191]. In variational approximation methods (LCAO, FE, PW, and WL), the discretization error $E_{R,N} - E_R$ is always nonnegative by construction. In systematically improvable methods (FD, FE, PW, and WL), this quantity goes to zero when N goes to infinity with a well-understood rate of convergence depending on the smoothness of the pseudopotential (see [44] for the PW case). However, in most applications, the discretization parameters are not tight enough for the discretization error to be lower than the target accuracy, which is typically of the order of 1 kcal/mol or 1 mHa (recall that 1 mHa $\simeq$ 0.6275 kcal/mol $\simeq$ 27.2 meV, which corresponds to an equivalent temperature of about 316 K). It is often advocated that this is not an issue since the real quantity of interest is not the value of the energy E_R for a particular nuclear configuration R, but the energy difference $E_{R_1} - E_{R_2}$ between two different configurations R_1 and R_2. It is indeed expected that

$$|(E_{R_1,N} - E_{R_2,N}) - (E_{R_1} - E_{R_2})| \ll |E_{R_1,N} - E_{R_1}| + |E_{R_2,N} - E_{R_2}|,$$

that is, the numerical error on the energy difference between the two configurations is much smaller than the sum of the discretization errors on the energies of each configuration. This expected phenomenon goes by the name of (discretization) error cancellation in the Physics and Chemistry literatures.

Obviously, for variational discretization methods, $E_{R_j,N} - E_{R_j} \geq 0$ so that both discretization errors have the same sign, leading to

$$\begin{aligned} |(E_{R_1,N} - E_{R_2,N}) - (E_{R_1} - E_{R_2})| &= |(E_{R_1,N} - E_{R_1}) - (E_{R_2,N} - E_{R_2})| \\ &\leq \max\left(E_{R_1,N} - E_{R_1}, E_{R_2,N} - E_{R_2}\right), \end{aligned}$$

but this does not explain the magnitude of the error cancellation phenomenon. The commonly admitted *qualitative* argument usually raised to explain this phenomenon is that the errors $E_{R_1,N} - E_{R_1}$ and $E_{R_2,N} - E_{R_2}$ are of the same nature and almost annihilate one another.

The purpose of this article is to provide a *quantitive* analysis of discretization error cancellation for PW discretization methods. First, we report in Section 3.2 two systematic numerical studies on, respectively, the hydrogen molecule and a simple system consisting of six atoms. For these systems, we are able to perform very accurate calculations with high PW cut-offs and

tight convergence criteria, which provide excellent approximations of the ground state energy E_R. We then compute, for two different configurations R_1 and R_2, the error cancellation factor

$$0 \leq Q_N := \frac{|(E_{R_1,N} - E_{R_2,N}) - (E_{R_1} - E_{R_2})|}{|E_{R_1,N} - E_{R_1}| + |E_{R_2,N} - E_{R_2}|} \leq 1.$$

We observe that this ratio is indeed small (typically between 10^{-3} and 10^{-1} depending on the system and on the configurations R_1 and R_2), and that it does not vary much with N. In Section 3.3, we introduce a toy model consisting of seeking the ground state of a one-dimensional linear periodic Schrödinger equation with Dirac potentials:

$$\left(-\frac{d^2}{dx^2} - \sum_{m\in\mathbb{Z}} z_1\delta_m - \sum_{m\in\mathbb{Z}} z_2\delta_{m+R}\right) u_R = E_R u_R, \qquad \int_0^1 u_R^2(x)dx = 1,$$

for which we can prove that the error cancellation factor Q_N converges to a fixed number $0 < Q_\infty < 1$ when N goes to infinity. Interestingly, it is possible to obtain a simple explicit expression of Q_∞, which only depends on z_1, z_2 and on $u_{R_1}(0)^2$, $u_{R_2}(0)^2$, $u_{R_1}(R_1)^2$, $u_{R_1}(R_2)^2$, i.e. on the values of the densities $\rho_{R_1} = u_{R_1}^2$ and $\rho_{R_1} = u_{R_2}^2$ at the singularities of the potential.

An alternative way to estimate the error on the energy difference between two configurations R_1 and R_2 is to integrate the error on the atomic forces on a smooth path linking R_1 and R_2. We conclude Section 3.2 by showing that the latter approach is not efficient in general.

3.2 Discretization error cancellation in planewave calculations

We present here some numerical simulations on two systems: the H_2 molecule and a system consisting of two oxygen atoms and four hydrogen atoms. The simulations are done in a cubic supercell of size 10×10×10 bohrs with the Abinit simulation package [120, 121]. The chosen approximate model is the periodic Kohn-Sham LDA model [152] with the parametrization and the pseudopotential proposed in [117]. Note that, in this work, we consider the approximation consisting of replacing the original problem set on the whole space $\mathbb{R}^3$ with a problem set on a cubic supercell with periodic boundary conditions as a *model error*. Alternatively, this error could be regarded as a discretization error: the supercell problem can indeed be seen as a non-consistent, non-conforming approximation of the original problem set on the whole space (see [53], in which this point of view was adopted to study the case of a local defect embedded in a perfect crystal).

For each configuration R, we compute a reference ground state energy E_R taking a high energy cutoff $E_{\text{cut}} = 400$ Ha. We then compute approximate energies for $N = E_{\text{cut}}$ varying from 5 to 105 Ha by steps of 5 Ha. The so-obtained energies are denoted by $E_{R,N}$.

For two given configurations R_1 and R_2 of the same system, we compute S_N, the sum of the discretization errors on the energies of the two configurations (note that $E_{R,N} - E_R \geq 0$ since PW is a variational approximation method), and D_N, the discretization error on the energy difference:

$$S_N = (E_{R_1,N} - E_{R_1}) + (E_{R_2,N} - E_{R_2}) \quad \text{and} \quad D_N = |(E_{R_1,N} - E_{R_2,N}) - (E_{R_1} - E_{R_2})|,$$

as well as the error cancellation factor

$$Q_N = \frac{D_N}{S_N} = \frac{|(E_{R_1,N} - E_{R_2,N}) - (E_{R_1} - E_{R_2})|}{(E_{R_1,N} - E_{R_1}) + (E_{R_2,N} - E_{R_2})}.$$

The two chemical systems considered in this section are very simple. We can therefore safely assume that for each configuration, our numerical simulations provide good approximations

of the Kohn–Sham ground state. Besides, very tight convergence criteria are used, so that algorithmic errors are negligible. Implementation and computer errors are not expected to be significant in this context.

3.2.1 Ground state potential energy surface of the H_2 molecule

In all our calculations, the H_2 molecule lies on the x axis and is centered at the origin. The parameter R is here the interatomic distance in bohrs.

We numerically observe that D_N is smaller than S_N by a factor of 10 to 100, and that the error cancellation factor Q_N is smaller when the two interatomic distances are close to each other ($R_1 \simeq R_2$). Morevoer, Q_N is almost constant with respect to the cut-off energy N.

In Figure 3.1, we present detailed results for two different pairs of configurations. On the top, the configurations are rather close since the interatomic distances are $R_1 = 1.464$ and $R_2 = 1.524$ bohr. For this approximate model, the equilibrium distance is about $R_{\text{eq}} \simeq 1.464$ bohrs (the experimental value is $R_{\text{eq}}^{\text{exp}} \simeq 1.401$ bohrs). The energy difference is better approximated by a factor of about 50 compared to the energies ($Q_N \simeq 0.02$). Moreover the log-log plots of S_N and D_N are almost parallel, which suggests that there is no improvement in the order of convergence when considering energy differences instead of energies; only the prefactor is improved. This is confirmed by the plots of the error cancellation factor Q_N, showing that this ratio does not vary much with N. On the bottom, the configurations are further apart. The interatomic distances are $R_1 = 1.344$ and $R_2 = 1.704$ bohrs. We observe a similar behavior except that the error cancellation phenomenon is less pronounced ($Q_N \simeq 0.1$).

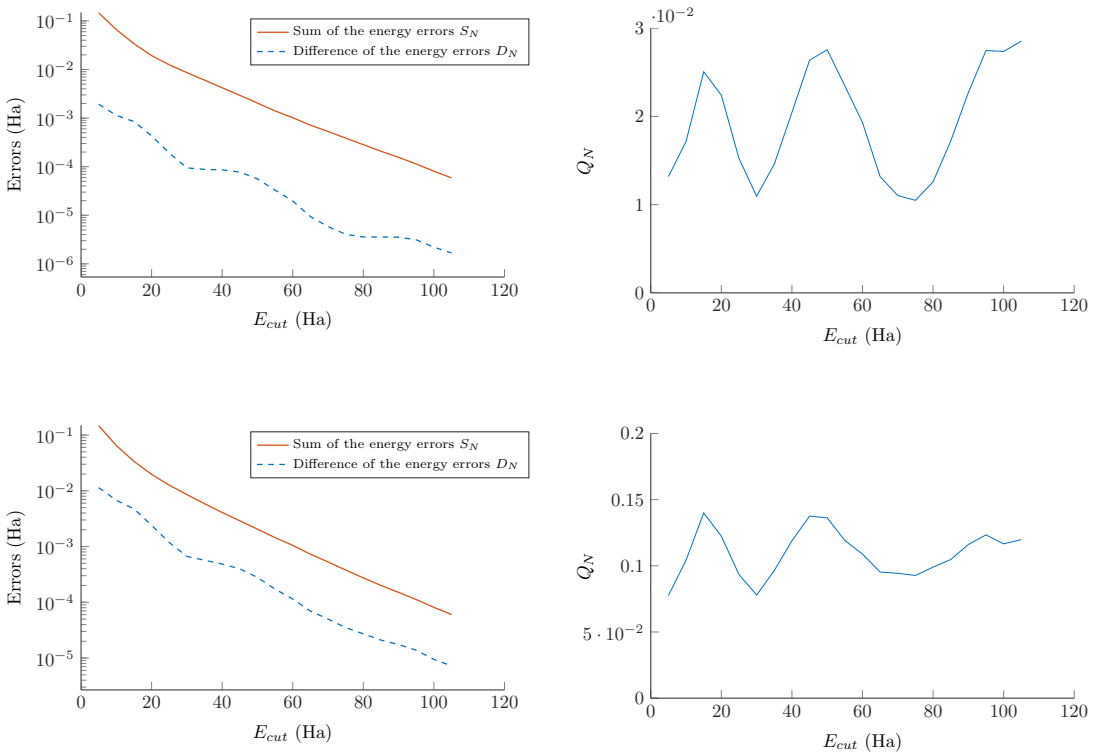

Figure 3.1 – Convergence plots of the quantities S_N and D_N (left) and of the error cancellation factor $Q_N = D_N/S_N$ (right) for two different pairs of interatomic distances for the H_2 molecule. Top: $R_1 = 1.464$ and $R_2 = 1.524$ bohrs. Bottom: $R_1 = 1.344$ and $R_2 = 1.704$ bohrs.

We then compare in Table 3.1 the values of S_N and D_N for different pairs of configurations and for two values of $N = E_{\text{cut}}$: a rather coarse energy cut-off $N = 30$ Ha, and a quite fine one

$N = 100$ Ha. One configuration is kept fixed ($R_1 = 1.284$ bohrs), while the second one varies from $R_2 = 1.344$ bohrs (close configurations) to $R_2 = 1.764$ bohrs (distant configurations). We also report, for each pair of configurations, the minimum, maximum, and mean values of Q_N over the different tested energy cutoffs $5 \leq N \leq 105$ Ha. We also observe that Q_N increases with $R_2 - R_1$ on the range $R_2 = [1.344, 1.764]$.

R_1	R_2	$S_{N=30}$	$D_{N=30}$	$S_{N=100}$	$D_{M=100}$	$\min(Q_N)$	$\max(Q_N)$	$\text{mean}(Q_N)$
1.284	1.344	9.410	0.1985	0.09157	0.00112	0.0103	0.0340	0.0212
1.284	1.404	9.268	0.3408	0.08990	0.00279	0.0216	0.0633	0.0413
1.284	1.464	9.160	0.4491	0.08772	0.00497	0.0375	0.0895	0.0610
1.284	1.524	9.065	0.5436	0.08552	0.00717	0.0544	0.1107	0.0802
1.284	1.584	8.969	0.6394	0.08380	0.00889	0.0713	0.1285	0.0985
1.284	1.644	8.863	0.7456	0.08274	0.00995	0.0841	0.1455	0.1151
1.284	1.704	8.744	0.8646	0.08213	0.01056	0.0983	0.1642	0.1302
1.284	1.764	8.615	0.9937	0.08154	0.01115	0.1072	0.1802	0.1440

Table 3.1 – Comparison of S_N, D_N and Q_N for different atomic configurations of the H_2 molecule. Distances are in bohrs, energies in mHa.

3.2.2 Energy of a simple chemical reaction

In this section, we consider the energy difference between two very different configurations of a system consisting of two oxygen atoms and four hydrogen atoms. The first configuration, denoted by R_1, corresponds to the chemical system 2 H_2O (two water molecules) and the second one, denoted by R_2, to the chemical system 2 H_2 + O_2, all these molecules being in their equilibrium geometry (see Figure 3.2). The energy difference between the two configurations thus provides a rough estimate of the energy of the chemical reaction

$$2\ H_2 + O_2 \longrightarrow 2\ H_2O.$$

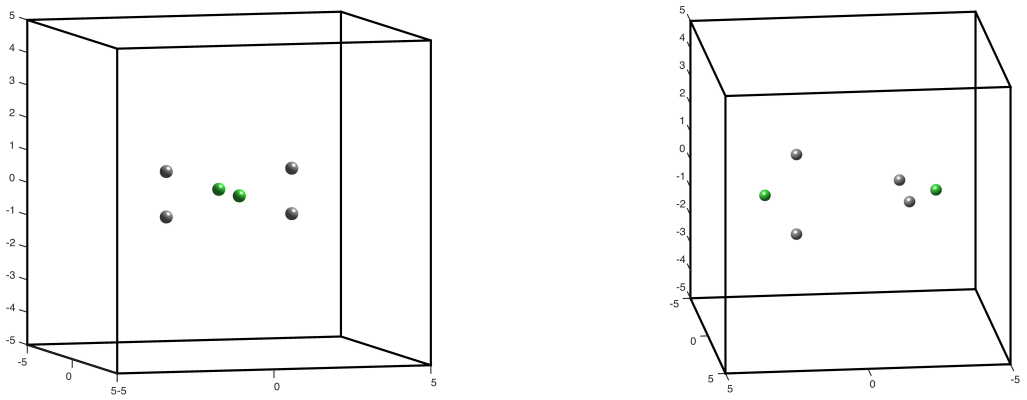

Figure 3.2 – Graphical representation of the two atomic configurations whose energies are compared. Oxygen atoms are in green, hydrogen atoms in black.

We can observe on Figure 3.3 and Table 3.2 a similar behavior as for H_2, but with a better error cancellation factor ($Q_N \simeq 0.005$).

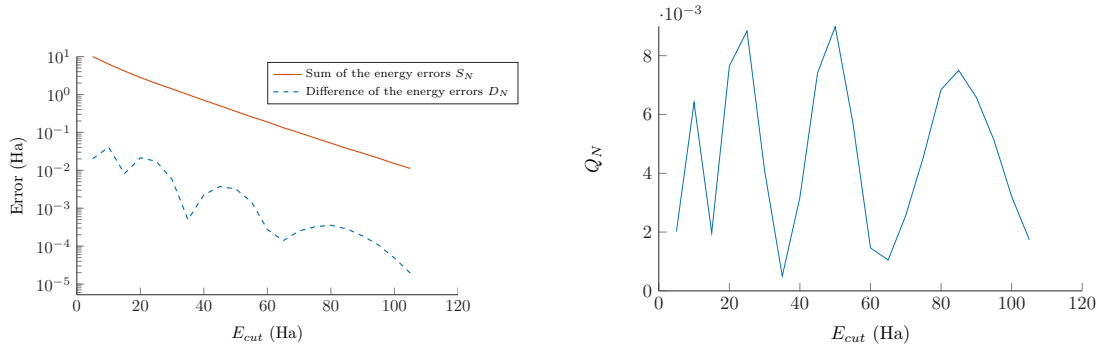

Figure 3.3 – Convergence plots of the quantities S_N and D_N (left) and of the error cancellation factor $Q_N = D_N/S_N$ (right) for the two different configurations displayed on Figure 3.2.

$S_{N=30}$	$D_{N=30}$	$S_{N=100}$	$D_{N=100}$	$\min(Q_N)$	$\max(Q_N)$	$\text{mean}(Q_N)$
1403	5.726	15.12	0.0485	0.0005036	0.008986	0.004640

Table 3.2 – Comparison of S_N, D_N (in mHa) and Q_N for the two different configurations displayed on Figure 3.2.

3.3 Mathematical analysis of a toy model

We now present a simple one-dimensional periodic linear Schrödinger model for which the discretization error cancellation phenomenon observed in the previous section can be explained with full mathematical rigor.

We denote by

$$L^2_{\text{per}} := \left\{ u \in L^2_{\text{loc}}(\mathbb{R}) \;\middle|\; u \text{ is } 1-\text{periodic} \right\}$$

the vector space of the 1-periodic locally square integrable real-valued functions on $\mathbb{R}$, and by

$$H^1_{\text{per}} := \left\{ u \in L^2_{\text{per}} \;\middle|\; u' \in L^2_{\text{per}} \right\}$$

the associated order-1 Sobolev space. For two given parameters $z_1, z_2 > 0$, we consider the family of problems, indexed by $R \in (0,1)$, consisting in finding the ground state $(u_R, E_R) \in H^1_{\text{per}} \times \mathbb{R}$ of

$$\left\{ \begin{array}{l} \left(-\dfrac{d^2}{dx^2} - \displaystyle\sum_{m\in\mathbb{Z}} z_1 \delta_m - \sum_{m\in\mathbb{Z}} z_2 \delta_{m+R} \right) u_R = E_R u_R, \\ \displaystyle\int_0^1 u_R^2(x)dx = 1, \quad u_R \geq 0, \end{array} \right. \tag{3.3.1}$$

where δ_a denotes the Dirac mass at point $a \in \mathbb{R}$. A variational formulation of the problem is: find the ground state $(u_R, E_R) \in H^1_{\text{per}} \times \mathbb{R}$ of

$$\left\{ \begin{array}{l} \forall v \in H^1_{\text{per}}, \displaystyle\int_0^1 u_R'(x)v'(x)dx - z_1 u_R(0)v(0) - z_2 u_R(R)v(R) = E_R \int_0^1 u_R(x)v(x)dx, \\ \displaystyle\int_0^1 u_R^2(x)dx = 1, \quad u_R \geq 0. \end{array} \right. \tag{3.3.2}$$

Remark 3.3.1. *The ground state eigenvalue E_R is negative. Indeed, using the variational*

characterization of the ground state energy, we get

$$E_R = \min_{v \in H^1_{\mathrm{per}} \setminus \{0\}} \frac{\displaystyle\int_0^1 v'(x)^2 dx - z_1 v(0)^2 - z_2 v(R)^2}{\displaystyle\int_0^1 v^2(x) dx} < 0,$$

since the Rayleigh quotient is equal to $-z_1 - z_2 < 0$ *for the constant test function* $v = 1$.

Denoting by $k_R = \sqrt{-E_R}$, we have

$$\begin{cases} u_R(x) = Ae^{k_R x} + Be^{-k_R x}, & \forall x \in [0, R], \\ u_R(x) = Ce^{k_R x} + De^{-k_R x}, & \forall x \in [R-1, 0), \end{cases} \tag{3.3.3}$$

where A, B, C, and D are real-valued constants. Since the function u_R is 1-periodic and continuous on $\mathbb{R}$ and its derivative satisfies the jump conditions $u_R'(m+0) - u_R'(m-0) = -z_1 u_R(m)$ and $u_R'(m+R+0) - u_R'(m+R-0) = -z_2 u_R(m+R)$ for all $m \in \mathbb{Z}$, the coefficients A, B, C, D solve the linear system

$$\underbrace{\begin{pmatrix} 1 & 1 & -1 & -1 \\ e^{k_R R} & e^{-k_R R} & -e^{k_R(R-1)} & -e^{-k_R(R-1)} \\ k_R + z_1 & -k_R + z_1 & -k_R & k_R \\ (k_R - z_2)e^{k_R R} & -(k_R + z_2)e^{-k_R R} & -k_R e^{k_R(R-1)} & k_R e^{-k_R(R-1)} \end{pmatrix}}_{M(k_R)} \begin{pmatrix} A \\ B \\ C \\ D \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

The wave vector k_R is the lowest positive root of the function $k \mapsto \det(M(k))$. The coefficients (A, B, C, D) are then uniquely determined by the normalization condition $\|u_R\|_{L^2_{\mathrm{per}}} = 1$ and the positivity of u_R. Exact solutions for two different values of the triplet of parameters (z_1, z_2, R) are plotted in Figure 3.4.

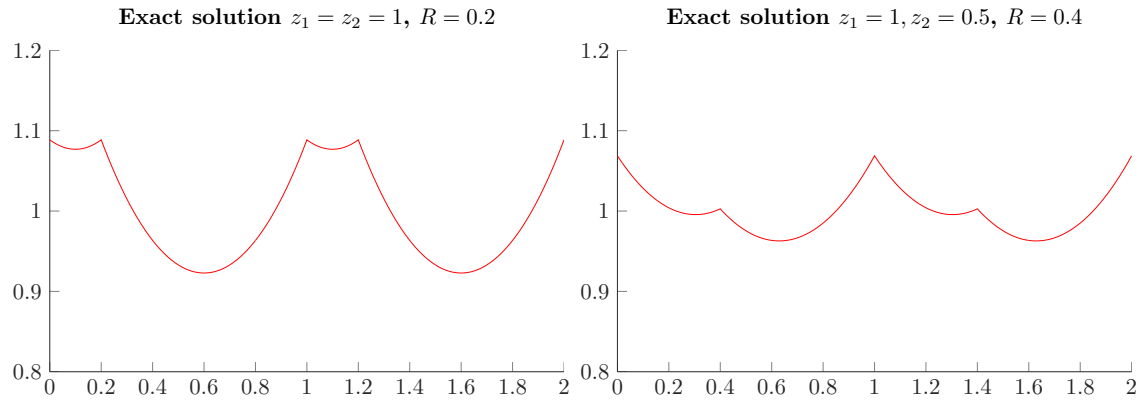

Figure 3.4 – Plot of the exact solutions of (3.3.1) for two sets of parameters.

An approximate solution of the problem is obtained using the PW discretization method. Denoting by

$$X_N := \mathrm{Span}\left\{ v_N(x) = \sum_{k \in \mathbb{Z},\ |k| \le N} \widehat{v}_k e^{2\pi i k x} \;\middle|\; \widehat{v}_k \in \mathbb{C},\ \widehat{v}_{-k} = \overline{\widehat{v}_k} \right\} \subset H^1_{\mathrm{per}},$$

the variational approximation of problem (3.3.2) in X_N consists in computing the ground state $(u_{R,N}, E_{R,N}) \in X_N \times \mathbb{R}$ of

$$\left\{ \begin{array}{l} \forall v_N \in X_N, \quad \displaystyle\int_0^1 u'_{R,N} v'_N - z_1 u_{R,N}(0) v_N(0) - z_2 u_{R,N}(R) v_N(R) \; = E_{R,N} \int_0^1 u_{R,N} v_N, \\ \displaystyle\int_0^1 u_{R,N}^2 = 1, \quad \int_0^1 u_{R,N} \geq 0. \end{array} \right. \tag{3.3.4}$$

The conditions $\widehat{v}_{-k} = \overline{\widehat{v}_k}$ in the definition of X_N is equivalent to imposing that the elements of X_N are real-valued functions. For convenience, the discretization parameter N here corresponds to the cut-off in momentum space. As above, we consider the error cancellation factor

$$Q_N = \frac{|(E_{R_1,N} - E_{R_2,N}) - (E_{R_1} - E_{R_2})|}{(E_{R_1,N} - E_{R_1}) + (E_{R_2,N} - E_{R_2})} \tag{3.3.5}$$

associated with the pair of configurations (R_1, R_2).

Note that imposing the condition $\int_0^1 u_{R,N} \geq 0$, we ensure that the discrete eigenfunction $u_{R,N}$ will approximate the positive eigenfunction u_R to the continuous problem (3.3.1) and not $-u_R$.

Theorem 3.3.2 (Asymptotic expressions of the energy error and of the error cancellation factor)**.** *For all $z_1, z_2 > 0$ and $R \in (0,1)$, we have for all $\epsilon > 0$,*

$$E_{R,N} - E_R = \frac{\alpha_R}{N} - \frac{\alpha_R}{2N^2} + \frac{\beta^{(1)}_{R,N}}{N} + \frac{\gamma_R}{N} \eta_{R,N} + o\left(\frac{1}{N^{3-\epsilon}}\right), \tag{3.3.6}$$

where

$$\alpha_R := \frac{z_1^2 u_R(0)^2 + z_2^2 u_R(R)^2}{2\pi^2}, \quad \gamma_R := \frac{z_1 z_2 u_R(0) u_R(R)}{\pi^2}, \quad \eta_{R,N} := N \sum_{k=N+1}^{+\infty} \frac{\cos(2\pi k R)}{k^2},$$

$$\beta^{(1)}_{R,N} := \frac{z_1^2 u_R(0)(u_{R,N}(0) - u_R(0)) + z_2^2 u_R(R)(u_{R,N}(R) - u_R(R))}{2\pi^2}.$$

In addition

$$|\eta_{R,N}| \leq \min\left(1, \frac{2 + \frac{\pi^3}{8}}{|\sin(\pi R)| N}\right),$$

and for all $\epsilon > 0$, there exists $C_\epsilon \in \mathbb{R}_+$ such that

$$|\beta^{(1)}_{R,N}| \leq \frac{C_\epsilon}{N^{1-\epsilon}}.$$

As a consequence, we have for all $z_1, z_2 > 0$ and all $R_1, R_2 \in (0,1)$,

$$\lim_{N \to +\infty} Q_N = \frac{|\alpha_{R_1} - \alpha_{R_2}|}{\alpha_{R_1} + \alpha_{R_2}} = \frac{\left|z_1^2 \left(u_{R_1}(0)^2 - u_{R_2}(0)^2\right) + z_2^2 \left(u_{R_1}(R_1)^2 - u_{R_2}(R_2)^2\right)\right|}{z_1^2 (u_{R_1}(0)^2 + u_{R_2}(0)^2) + z_2^2 (u_{R_1}(R_1)^2 + u_{R_2}(R_2)^2)}. \tag{3.3.7}$$

The proof of the above theorem is given in Appendix. We deduce from (3.3.6) that the discretization error $E_{R,N} - E_R$ on the energy of the configuration R is the sum of

1. a leading term $\alpha_R N^{-1}$ of order 1 (in N^{-1});
2. three terms $-1/2\alpha_R N^{-2}$, $\beta^{(1)}_{R,N} N^{-1}$, and $\gamma_R N^{-1} \eta_{R,N}$ which are roughly of order 2;

3. higher order terms which are roughly of order 3 and above.

The leading term $\alpha_R N^{-1}$ has a very simple expression and the prefactor α_R does not vary much with respect to R (see Figure 3.5). This explains the phenomenon of discretization error cancellation. Regarding the second order corrections on $E_{R,N} - E_R$, we have observed numerically (see Figure 3.6) that

- the terms $-\frac{1}{2}\alpha_R N^{-2}$ and $\gamma_R N^{-1}\eta_{R,N}$ are of about the same order of magnitude in absolute values, that the former is always negative (since $\alpha_R > 0$), but that the latter can be either positive or negative, so that the sum of these two contributions can be either significant or negligible;

- the term $\beta^{(1)}_{R,N} N^{-1}$ is smaller in absolute value than the other two terms, and seems to be always negative. Our numerical calculations indeed show that $u_{R,N}(0) < u_R(0)$ and $u_{R,N}(R) < u_R(R)$, which is not very surprising since the function u_R has cusps at points $x = 0$ and $x = R$ (see Figure 3.4). These inequalities have not been rigorously established though.

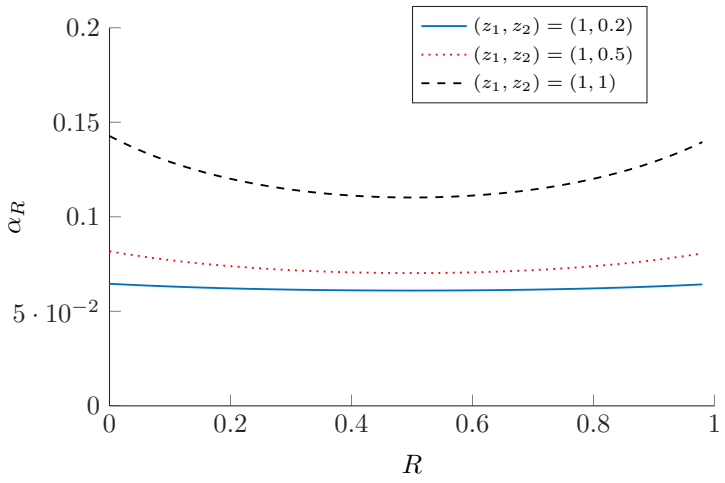

Figure 3.5 – Plots of the function $R \mapsto \alpha_R$ for three sets of parameters (z_1, z_2).

Finally, we observe on Figure 3.7 that Q_N converges to the asymptotic value Q_∞ when N goes to infinity very smoothly for large values of R, and with oscillations when R becomes close to zero. Moreover, $Q_N - Q_\infty$ is of order N^{-2}.

Remark 3.3.3. *The 1D model studied in this section involves Dirac potentials, for which the exact solutions* (3.3.3)*, as well as the lowest-order terms of the discretization error* (3.3.6)*, can be computed explicitly. It would have been possible to use more regular potentials with explicit solutions, such as piecewise constant potentials for instance. However, the calculations would have been more tedious than for the Dirac case, and we anticipate that, qualitatively, the results would have been similar. Loosely speaking, the faster convergence of the energy difference originates from the fact that the leading term of the error depends on the nuclear configuration, but not that much. This explains why the convergence rate is not improved, while the prefactor is improved. For smoother potentials, as well as for pseudopotentials, it is expected that most of the error on the energy remains concentrated in the vicinities of the core regions, where, for different nuclear configurations, the electronic orbitals change, but not much.*

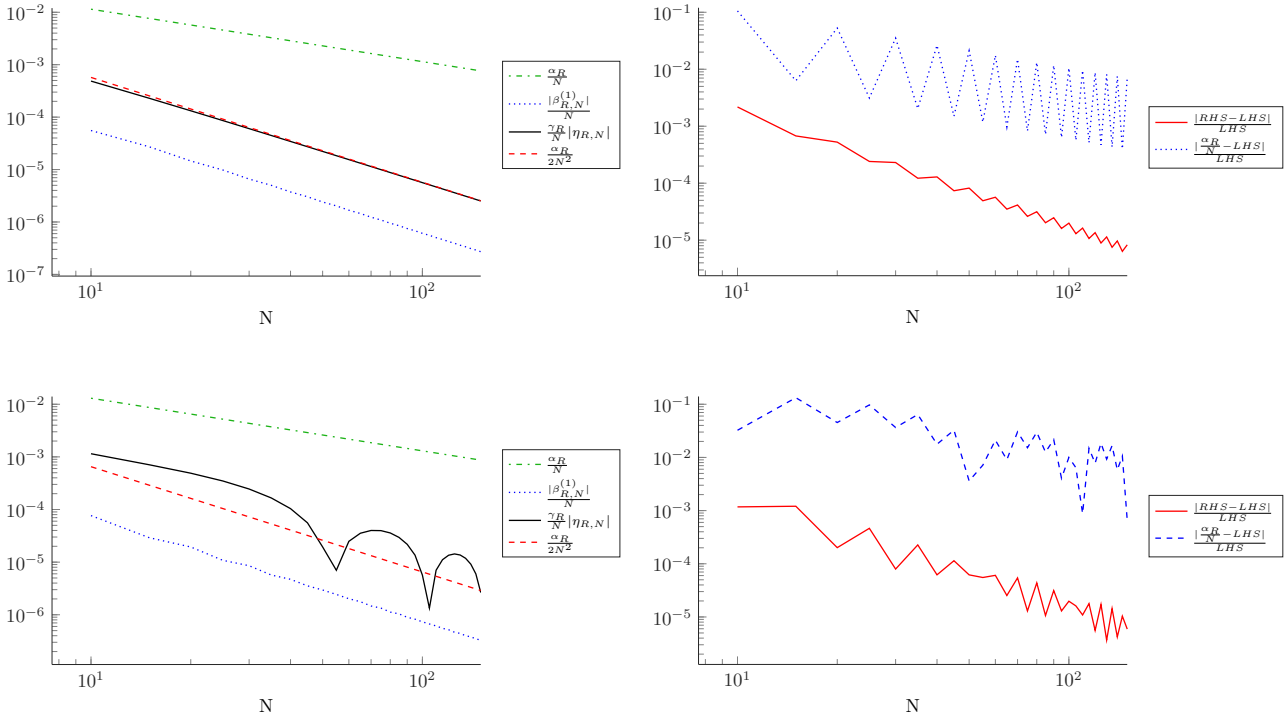

Figure 3.6 – Convergence plots of the four quantities $\frac{\alpha_R}{N}$, $\frac{\alpha_R}{2N^2}$, $\frac{|\beta^{(1)}_{R,N}|}{N}$, and $\frac{\gamma_R}{N}|\eta_{R,N}|$ (left) and plots of $\frac{|(\frac{\alpha_R}{N}-\frac{\alpha_R}{2N^2}+\frac{\beta^{(1)}_{R,N}}{N}+\frac{\gamma_R}{N}\eta_{R,N})-(E_{R,N}-E_R)|}{E_{R,N}-E_R}$ and $\frac{|\frac{\alpha_R}{N}-(E_{R,N}-E_R)|}{E_{R,N}-E_R}$ (right). Top: $z_1 = z_2 = 1, R = 0.3$. Bottom: $z_1 = z_2 = 1, R = 0.09$.

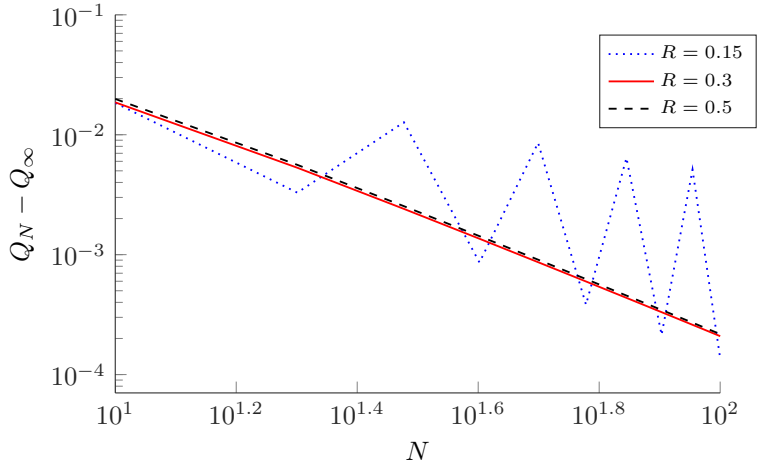

Figure 3.7 – Plot of $Q_N - Q_\infty$ for three values of R.

Remark 3.3.4. *Note that a variant of the projected augmented wave (PAW) method [25] was recently studied for the 1D model considered here [23]: it is shown that the error on the energy has two contributions, the first one scaling as* $r_c^{4N_0} N^{-1}$, *and the second one as* $r_c^{-p} N^{-(p+1)}$, *where* r_c *is the core radius,* N_0 *the number of pseudo-orbitals,* p *the degree of the (polynomial) pseudo-orbitals in the core region, and* N *the number of planewaves. However, it is not clear how to use the estimates in [23] to obtain estimates on energy differences. We intend to*

investigate this point in the future.

To conclude, let us comment on the alternative approach to estimate the error on the energy difference between two configurations consisting in integrating the error on the atomic forces along a path in the nuclear configuration space liking the two configurations. In this simple 1D setting, we have, for $R_1 < R_2$,

$$|(E_{R_1,N} - E_{R_2,N}) - (E_{R_1} - E_{R_2})| = \left| \int_{R_1}^{R_2} (F_{R,N} - F_R) \, dR \right|,$$

where

$$F_{N,R} := -\frac{dE_{R,N}}{dR} \text{ and } F_R := -\frac{dE_R}{dR}.$$

The use of a variational method guaranties that the energy error $E_{R,N} - E_R$ is nonnegative for all N and all R. On the other hand, the error on the force $F_{R,N} - F_R$ does not have a constant sign (it integrates to zero on the interval $[0, 1]$), so that, in general,

$$|(E_{R_1,N} - E_{R_2,N}) - (E_{R_1} - E_{R_2})| = \left| \int_{R_1}^{R_2} (F_{R,N} - F_R) \, dR \right| \leq \int_{R_1}^{R_2} |F_{R,N} - F_R| \, dR.$$

The left hand-side of the above inequality can *a priori* be much smaller than the right hand-side. In this case, using bounds on the error on the forces would lead to a dramatic overestimation of the error on the energy difference. This is confirmed by our numerical simulations. The functions

$$(R_1, R_2) \mapsto \left| \int_{R_1}^{R_2} (F_{R,N} - F_R) \, dR \right| \quad \text{and} \quad (R_1, R_2) \mapsto \int_{R_1}^{R_2} |F_{R,N} - F_R| \, dR, \tag{3.3.8}$$

plotted in Figure 3.8, are very different and the latter one is not a good approximation of the former one. Another interesting observation is the following. Numerical simulations show that

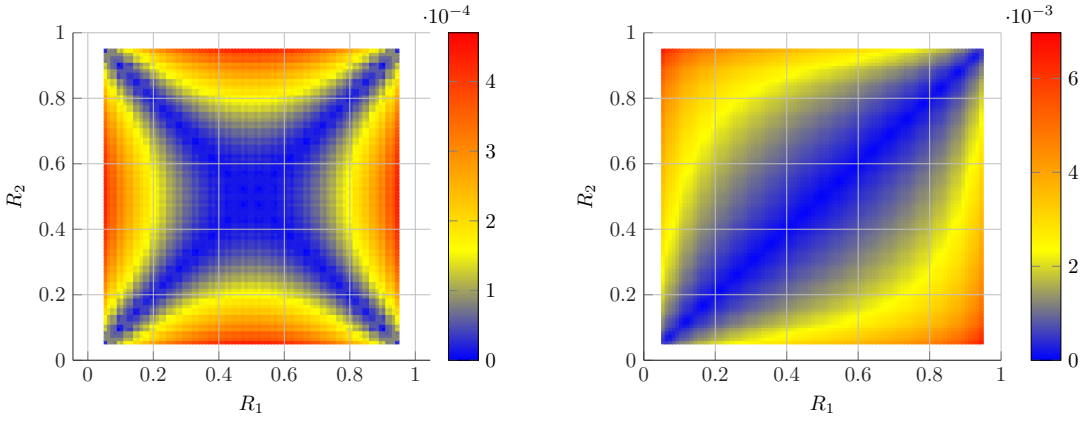

Figure 3.8 – Colorplots of the functions defined in (3.3.8). The forces were computed with centered finite difference with step size 10^{-6} and the integrals with Simpson's rule with step length 10^{-2}, chosen equal to the resolution of the figure.

the forces converge at the same rate as the energy, i.e. in $1/N$ (see Figure 3.9), and that, for each value of N in the range $[10, 100]$, the derivatives of the functions

$$R \mapsto E_{R,N} - E_R \quad \text{and} \quad R \mapsto \chi_{R,N} := \frac{\alpha_R}{N} - \frac{\alpha_R}{2N^2} + \frac{\beta^{(1)}_{R,N}}{N} + \frac{\gamma_R}{N} \eta_{R,N}$$

agree up to very small correction terms. Nevertheless, the derivative of the fourth term in $\chi_{R,N}$ (i.e. of $\gamma_R \eta_{R,N} N^{-1}$) can be much larger than the derivative of the first term (i.e. of $\alpha_R N^{-1}$). The leading term of the error on the force is therefore not in general (minus) the derivative of the leading term of the energy error. In Figure 3.10, the above functions are plotted for $N = 10$ (top) and $N = 100$ (bottom).

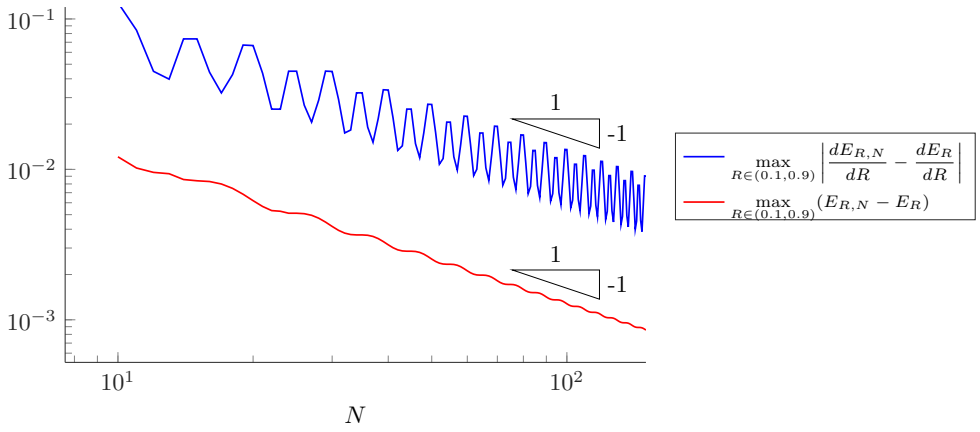

Figure 3.9 – Convergence of the errors on the energy (in red) and on the forces (in blue).

3.4 Appendix: proof of Theorem 3.3.2

In the sequel, z_1 and z_2 are fixed positive real numbers. We endow the functional spaces $L^2_{\rm per}$ and $H^1_{\rm per}$ with their usual scalar products

$$\langle u|v\rangle_{L^2_{\rm per}} := \int_0^1 u(x)v(x)\,dx \quad \text{and} \quad \langle u|v\rangle_{H^1_{\rm per}} := \langle u|v\rangle_{L^2_{\rm per}} + \langle u'|v'\rangle_{L^2_{\rm per}}.$$

More generally, we endow the Sobolev space

$$H^s_{\rm per} := \left\{ v(x) = \sum_{k\in\mathbb{Z}} \widehat{v}_k e^{2i\pi kx} \;\middle|\; \widehat{v}_k \in \mathbb{C},\ \widehat{v}_{-k} = \overline{\widehat{v}_k},\ \sum_{k\in\mathbb{Z}} (1+(2\pi k)^2)^s |\widehat{v}_k|^2 < \infty \right\},$$

$s \in \mathbb{R}$, with the scalar product defined by

$$\langle u|v\rangle_{H^s_{\rm per}} := \sum_{k\in\mathbb{Z}} (1+(2\pi k)^2)^s\, \overline{\widehat{u}_k}\, \widehat{v}_k.$$

Note that the above two definitions of $\langle u|v\rangle_{H^1_{\rm per}}$ coincide and that $H^0_{\rm per} = L^2_{\rm per}$. We also denote by Π_N the orthogonal projection on X_N for the $L^2_{\rm per}$ (and also $H^s_{\rm per}$) scalar product and by $\Pi_N^\perp = 1 - \Pi_N$.

We first recall some useful results on the convergence of $(u_{R,N}, E_{R,N})$ to (u_R, E_R).

Lemma 3.4.1. *Let $R \in (0,1)$. Let (u_R, E_R) be the ground state of the continuous problem (3.3.2), and $(u_{N,R}, E_{R,N})$ be a ground state of the discretized problem (3.3.4). Then, for all $\epsilon > 0$ and all $0 \le s < 3/2$, there exists $C_{s,\epsilon} \in \mathbb{R}_+$ such that*

$$\|u_{R,N} - u_R\|_{H^s_{\rm per}} \le \frac{C_{s,\epsilon}}{N^{3/2-s-\epsilon}}. \tag{3.4.1}$$

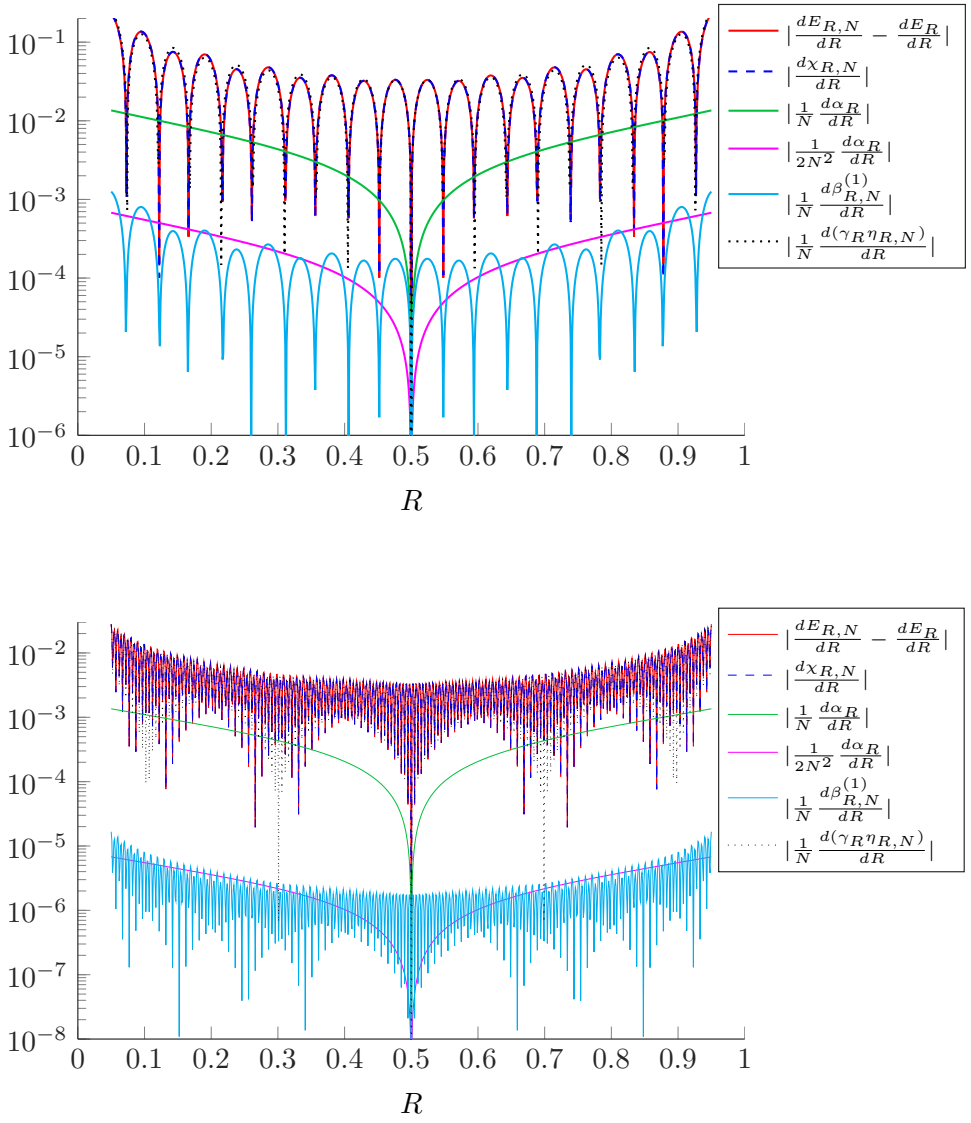

Figure 3.10 – Plots of the functions $R \mapsto \frac{dE_{R,N}}{dR} - \frac{dE_{R,N}}{dR}$ and $R \mapsto \frac{d\chi_{R,N}}{dR}$, and of the derivative of each of the four components of $\chi_{R,N}$, for $N = 10$ (top) and $N = 100$ (bottom). The derivatives were computed numerically by centered finite differences with step size 10^{-6}.

In addition, there exist $0 < c \leq C < \infty$ *such that*

$$c\|u_{R,N} - u_R\|^2_{H^1_{\mathrm{per}}} \leq E_{R,N} - E_R \leq C\|u_{R,N} - u_R\|^2_{H^1_{\mathrm{per}}}, \tag{3.4.2}$$

and for all $\epsilon > 0$*, there exists* $C_\epsilon \in \mathbb{R}_+$ *such that*

$$|u_{R,N}(0) - u_R(0)| + |u_{R,N}(R) - u_R(R)| \leq \frac{C_\epsilon}{N^{1-\epsilon}}. \tag{3.4.3}$$

Proof. We denote by C^0_{per} the space of continuous 1-periodic functions from $\mathbb{R}$ to $\mathbb{R}$ endowed with the norm defined by

$$\forall u \in C^0_{\mathrm{per}}, \quad \|u\|_{C^0_{\mathrm{per}}} := \max_{x\in\mathbb{R}} |u(x)|.$$

Recall that $H^s_{\rm per}$ is continuously embedded in $C^0_{\rm per}$ for all $s > 1/2$. In particular, $H^1_{\rm per} \hookrightarrow C^0_{\rm per}$ and there exists $K \in \mathbb{R}_+$ such that

$$\forall u \in H^1_{\rm per}, \quad \|u\|_{C^0_{\rm per}} \leq K\|u\|_{H^{3/4}_{\rm per}} \leq K\|u\|^{3/4}_{H^1_{\rm per}}\|u\|^{1/4}_{L^2_{\rm per}}. \tag{3.4.4}$$

In particular, the bilinear form

$$\forall (u,v) \in H^1_{\rm per} \times H^1_{\rm per}, \quad a_R(u,v) = \int_0^1 u'v' - z_1 u(0)v(0) - z_2 u(R)v(R)$$

is well-defined, symmetric, and continuous on $H^1_{\rm per} \times H^1_{\rm per}$, and we have

$$\begin{aligned}\forall u \in H^1_{\rm per}, \quad a_R(u,u) &\geq \|u\|^2_{H^1_{\rm per}} - (z_1+z_2)K^2\|u\|^{3/2}_{H^1_{\rm per}}\|u\|^{1/2}_{L^2_{\rm per}} - \|u\|^2_{L^2_{\rm per}} \\ &\geq \frac{1}{2}\|u\|^2_{H^1_{\rm per}} - \left(1 + \frac{27}{32}(z_1+z_2)^4K^8\right)\|u\|^2_{L^2_{\rm per}},\end{aligned}$$

using Young's inequality. The quadratic form $H^1_{\rm per} \ni u \mapsto a_R(u,u) \in \mathbb{R}$ therefore is bounded below and closed. We denote by H_R the unique self-adjoint operator on $L^2_{\rm per}$ associated to $a_R(\cdot,\cdot)$ (see e.g. [212, Theorem VIII.15]). Formally,

$$H_R = -\frac{d^2}{dx^2} - z_1\sum_{m\in\mathbb{Z}}\delta_m - z_2\sum_{m\in\mathbb{Z}}\delta_{m+R}.$$

The domain of H_R being a subspace of $H^1_{\rm per}$, which is itself compactly embedded in $L^2_{\rm per}$, the spectrum of H_R is purely discrete: it consists of an increasing sequence of eigenvalues of finite multiplicities going to $+\infty$. It is easily seen that its ground state eigenvalue E_R is simple. Let us denote by $\mu_R > 0$ the gap between the lowest two eigenvalues of H_R. A classical calculation shows that

$$\begin{aligned}E_{R,N} - E_R &= a_R(u_{R,N} - u_R, u_{R,N} - u_R) - E_R\|u_{R,N} - u_R\|^2_{L^2_{\rm per}} \\ &= \langle u_{R,N}|H_R|u_{R,N}\rangle - E_R.\end{aligned}$$

First, since $E_R < 0$, we have

$$E_{R,N} - E_R \leq a_R(u_{R,N} - u_R, u_{R,N} - u_R) \leq M_R\|u_{R,N} - u_R\|^2_{H^1_{\rm per}},$$

where M_R is the continuity constant of a_R, which proves the second inequality in (3.4.2). Second, since $\|u_R\|_{L^2_{\rm per}} = \|u_{R,N}\|_{L^2_{\rm per}} = 1$, we have on the one hand

$$\begin{aligned}E_{R,N} - E_R &= \langle u_{R,N}|H_R|u_{R,N}\rangle - E_R \\ &\geq \left(E_R|\langle u_{R,N}|u_R\rangle_{L^2_{\rm per}}|^2 + (E_R+\mu_R)\left(1 - |\langle u_{R,N}|u_R\rangle_{L^2_{\rm per}}|^2\right)\right) - E_R \\ &= \mu_R\left(1 - |\langle u_{R,N}|u_R\rangle_{L^2_{\rm per}}|^2\right) \geq \mu_R\left(1 - \langle u_{R,N}|u_R\rangle_{L^2_{\rm per}}\right) = \frac{\mu_R}{2}\|u_{R,N} - u_R\|^2_{L^2_{\rm per}},\end{aligned}$$

and, on the other hand,

$$E_{R,N} - E_R \geq \frac{1}{2}\|u_{R,N} - u_R\|^2_{H^1_{\rm per}} - \left(1 + \frac{27}{32}(z_1+z_2)^4K^8 + E_R\right)\|u_{R,N} - u_R\|^2_{L^2_{\rm per}}.$$

Combining the above two inequalities yields the first inequality in (3.4.2). Hence, (3.4.2) is proved.

We deduce from the min-max principle that for each $v_N \in X_N$ such that $\|v_N\|_{L^2_{\rm per}} = 1$, we have

$$E_{R,N} - E_R \leq a_R(v_N, v_N) - E_R = a_R(v_N - u_R, v_N - u_R) - E_R\|v_N - u_R\|^2_{L^2_{\rm per}} \\ \leq (M_R - E_R)\,\|v_N - u_R\|^2_{H^1_{\rm per}}.$$

Since $z_1 \sum_{m\in\mathbb{Z}} \delta_m + z_2 \sum_{m\in\mathbb{Z}} \delta_{m+R} \in H^{-1/2-\epsilon}_{\rm per}$ for all $\epsilon > 0$, we have that $u_R \in H^{3/2-\epsilon}_{\rm per}$. Applying the above estimate to $v_N = \|\Pi_N u_R\|^{-1}_{L^2_{\rm per}} \Pi_N u_R$, we get $E_{R,N} - E_R \leq \frac{C_\epsilon}{N^{1-\epsilon}}$. Combining with (3.4.2), we obtain (3.4.1) for $s = 1$. Together with (3.4.4), this implies in addition that $(u_{R,N})_{N\in\mathbb{N}}$ converges to u_R in $C^0_{\rm per}$. Since

$$-u''_{R,N} = z_1 u_{R,N}(0)\Pi_N\left(\sum_{k\in\mathbb{Z}} \delta_m\right) + z_2 u_{R,N}(R)\Pi_N\left(\sum_{k\in\mathbb{Z}} \delta_{m+R}\right) + E_{R,N}u_{R,N},$$

and the right hand-side converges to $-u''_R$ in $H^{-1/2-\epsilon}_{\rm per}$ for all $\epsilon > 0$, the sequence $(u_{R,N})_{N\in\mathbb{N}}$ converges to u_R in $H^{3/2-\epsilon}_{\rm per}$ for all $\epsilon > 0$. By interpolation, we then obtain (3.4.1) for all $1 \leq s < 3/2$. We finally obtain (3.4.1) for $s = 0$ by a classical Aubin-Nitsche argument, and we conclude by interpolation that the result also holds true for all $0 \leq s < 1$.

To prove (3.4.3), we infer from the Sobolev embedding $H^{1/2+\epsilon}_{\rm per} \hookrightarrow C^0_{\rm per}$, that

$$|u_{R,N}(0) - u_R(0)| + |u_{R,N}(R) - u_R(R)| \leq 2\|u_{R,N} - u_R\|_{C^0_{\rm per}} \leq 2C'_\epsilon \|u_{R,N} - u_R\|_{H^{1/2+\epsilon}_{\rm per}},$$

and we conclude using (3.4.1) with $s = 1/2 + \epsilon$. □

The following lemma provides an expression of the leading term of the energy difference $E_{R,N} - E_R$.

Lemma 3.4.2. *Let $z_1, z_2 > 0$. Let $R \in (0, 1)$. Let (u_R, E_R) be the ground state of the continuous problem* (3.3.2)*, and $(u_{R,N}, E_{R,N})$ be a ground state of the discretized problem* (3.3.4)*. Then, for all $\epsilon > 0$,*

$$E_{R,N} - E_R = z_1 u_{R,N}(0)(\Pi_N^\perp u_R)(0) + z_2 u_{R,N}(R)(\Pi_N^\perp u_R)(R) + o\left(\frac{1}{N^{3-\epsilon}}\right), \tag{3.4.5}$$

when N goes to $+\infty$.

Proof. The variational formulation (3.3.2) with $v = u_{R,N}$ gives

$$E_R \int_0^1 u_{R,N}u_R = \int_0^1 u'_{R,N}u'_R - z_1 u_{R,N}(0)u_R(0) - z_2 u_{R,N}(R)u_R(R).$$

The variational formulation (3.3.4) with $v_N = \Pi_N u_R$ gives

$$E_{R,N} \int_0^1 u_{R,N}(\Pi_N u_R) = \int_0^1 u'_{R,N}(\Pi_N u_R)' - z_1 u_{R,N}(0)(\Pi_N u_R)(0) - z_2 u_{R,N}(R)(\Pi_N u_R)(R).$$

Subtracting these two equalities, and noting first that $\displaystyle\int_0^1 u_{R,N}(\Pi_N u_R) = \int_0^1 u_{R,N}u_R$, and second that $\displaystyle\int_0^1 u'_{R,N}(\Pi_N u_R)' = \int_0^1 u'_{R,N}u'_R$, since $u_{R,N} \in X_N$ and the orthogonal projection Π_N and the derivation commute, we get

$$(E_{R,N} - E_R) \int_0^1 u_{R,N}u_R = z_1 u_{R,N}(0)(\Pi_N^\perp u_R)(0) + z_2 u_{R,N}(R)(\Pi_N^\perp u_R)(R).$$

Moreover, since $\displaystyle\int_0^1 u_R^2 = \int_0^1 u_{R,N}^2 = 1$, we have

$$\int_0^1 u_{R,N}u_R = 1 - \frac{1}{2}\int u_R^2 - \frac{1}{2}\int_0^1 {u_{R,N}}^2 + \int_0^1 u_{R,N}u_R = 1 - \frac{1}{2}\|u_{R,N} - u_R\|^2_{L^2_{\mathrm{per}}}.$$

Hence,

$$(E_{R,N} - E_R)\left(1 - \frac{1}{2}\|u_{R,N} - u_R\|^2_{L^2_{\mathrm{per}}}\right) = z_1 u_{R,N}(0)(\Pi_N^{\perp}u_R)(0) + z_2 u_{R,N}(R)(\Pi_N^{\perp}u_R)(R).$$

Using estimates (3.4.1) for $s = 0$ and (3.4.2), we obtain that for all $\epsilon > 0$,

$$1 - \frac{1}{2}\|u_{R,N} - u_R\|^2_{L^2_{\mathrm{per}}} = 1 + o\left(\frac{1}{N^{3-\epsilon}}\right), \quad \text{when } N \to +\infty.$$

This concludes the proof of Lemma 3.4.2. □

The following lemma provides an explicit expression of the quantities $(\Pi_N^{\perp}u_R)(0)$ and $(\Pi_N^{\perp}u_R)(R)$ appearing in (3.4.5).

Lemma 3.4.3. *Let $z_1, z_2 > 0$. For all $R \in (0,1)$, all $N \in \mathbb{N}$, and all $x \in \mathbb{R}$,*

$$(\Pi_N^{\perp}u_R)(x) = \sum_{k=N+1}^{+\infty} \frac{2}{k_R^2 + 4\pi^2k^2}\left(z_1 u_R(0)\cos(2\pi kx) + z_2 u_R(R)\cos(2\pi k(x-R))\right). \quad (3.4.6)$$

Proof. In order to estimate $(\Pi_N^{\perp}u_R)(x)$, we first need to compute the Fourier coefficients of u_R

$$\forall k \in \mathbb{Z}, \quad \widehat{u_R}(k) := \int_0^1 u_R(x)e^{-2i\pi kx}\,dx. \quad (3.4.7)$$

Using the periodicity of u_R, we can rewrite the first equation in (3.3.1) as

$$-u_R'' - z_1 u_R(0)\left(\sum_{m\in\mathbb{Z}}\delta_m\right) - z_2 u_R(R)\left(\sum_{m\in\mathbb{Z}}\delta_{m+R}\right) = E_R u_R.$$

Taking the Fourier transform, and using the relation $E_R = -k_R^2$, we obtain

$$4\pi^2k^2\widehat{u_R}(k) - z_1u_R(0) - z_2u_R(R)e^{-2i\pi kR} = -k_R^2\widehat{u_R}(k).$$

Hence, for all $k \in \mathbb{Z}$,

$$\widehat{u_R}(k) = \frac{1}{k_R^2 + 4\pi^2k^2}\left(z_1u_R(0) + z_2u_R(R)e^{-2i\pi kR}\right). \quad (3.4.8)$$

Consequently,

$$\begin{aligned}(\Pi_N^{\perp}u_R)(x) &= \sum_{k\in\mathbb{Z},\ |k|>N} \widehat{u_R}(k)e^{2i\pi kx} \\ &= \sum_{k\in\mathbb{Z},\ |k|>N} \frac{1}{k_R^2 + 4\pi^2k^2}\left(z_1u_R(0) + z_2u_R(R)e^{-2i\pi kR}\right)e^{2i\pi kx} \\ &= \sum_{k=N+1}^{+\infty} \frac{2}{k_R^2 + 4\pi^2k^2}\left(z_1u_R(0)\cos(2\pi kx) + z_2u_R(R)\cos(2\pi k(x-R))\right),\end{aligned}$$

which completes the proof of Lemma 3.4.3. □

The last technical lemma we need provides an estimates of the series in (3.4.6) for $x = 0$ and $x = R$.

Lemma 3.4.4. *Let $\mathbb{R} \ni R \mapsto k_R \in \mathbb{R}$ be a positive bounded function and $M = \sup_{R\in\mathbb{R}} k_R^2$. We denote by*

$$f_N(R) := \sum_{k=N+1}^{+\infty} \frac{1}{k_R^2 + 4\pi^2 k^2} \quad \textit{and} \quad g_N(R) := \sum_{k=N+1}^{+\infty} \frac{\cos(2\pi kR)}{k_R^2 + 4\pi^2 k^2}.$$

For all $R \in \mathbb{R} \setminus \mathbb{Z}$ we have

$$f_N(R) = \frac{1}{4\pi^2 N} a_N + \phi_N(R), \quad \textit{with} \quad a_N = N \sum_{k=N+1}^{+\infty} \frac{1}{k^2}, \quad |\phi_N(R)| \le \frac{M}{48\pi^4 N^3}, \tag{3.4.9}$$

and

$$g_N(R) = \frac{1}{4\pi^2 N} \eta_{N,R} + \psi_N(R), \quad \textit{with} \quad \eta_{N,R} = N \sum_{k=N+1}^{+\infty} \frac{\cos(2\pi kR)}{k^2}, \quad |\psi_N(R)| \le \frac{M}{48\pi^4 N^3}. \tag{3.4.10}$$

Besides,

$$a_N = 1 + \frac{1}{2N} + O\left(\frac{1}{N^2}\right) \qquad \textit{and} \qquad |\eta_{N,R}| \le \min\left(1, \frac{2 + \frac{\pi^3}{8}}{|\sin(\pi R)| N}\right). \tag{3.4.11}$$

Proof. The function f_N can be decomposed as

$$f_N(R) = \frac{1}{4\pi^2 N} a_N + \phi_N(R),$$

where

$$\phi_N(R) = f_N(R) - \frac{1}{4\pi^2 N} a_N = -\frac{k_R^2}{4\pi^2} \sum_{k=N+1}^{+\infty} \frac{1}{k^2 (k_R^2 + 4\pi^2 k^2)}.$$

We have on the one hand

$$\begin{aligned} a_N &= 1 + N \sum_{k=N+1}^{+\infty} \left(\frac{1}{k^2} - \int_{k-1}^{k} \frac{dt}{t^2} \right) \\ &= 1 + N \sum_{k=N+1}^{+\infty} \frac{1}{k^2} \int_0^1 \left(1 - \left(1 - \frac{s}{k}\right)^{-2} \right) ds = 1 + \frac{1}{2N} + O\left(\frac{1}{N^2}\right), \end{aligned}$$

and on the other hand, by a sum-integral comparison,

$$|\phi_N(R)| \le \frac{M}{4\pi^2} \sum_{k=N+1}^{+\infty} \frac{1}{4\pi^2 k^4} \le \frac{M}{48\pi^4 N^3}.$$

Thus, (3.4.9) and the first statement of (3.4.11) are proved. For $N \in \mathbb{N}$ and $R \in \mathbb{R}$, we set

$$h_N(R) := \sum_{k=N+1}^{+\infty} \frac{\cos(2\pi kR)}{4\pi^2 k^2} = \frac{1}{4\pi^2 N} \eta_{R,N}.$$

We have

$$|\psi_N(R)| = |g_N(R) - h_N(R)| = \left| - \sum_{k=N+1}^{+\infty} \frac{k_R^2 \cos(2\pi k R)}{4\pi^2 k^2 (k_R^2 + 4\pi^2 k^2)} \right| \leq M \sum_{k=N+1}^{+\infty} \frac{1}{16\pi^4 k^4} \leq \frac{M}{48\pi^4 N^3}.$$

Taking the second derivative of h_N in the distribution sense and using Poisson summation formula, we obtain

$$\begin{aligned} h_N''(R) &= \frac{d^2}{dR^2} \left(\sum_{k=N+1}^{+\infty} \frac{e^{2i\pi kR} + e^{-2i\pi kR}}{8\pi^2 k^2} \right) = -\frac{1}{2} \left(\sum_{k \in \mathbb{Z} \,|\, |k| > N} e^{2i\pi kR} \right) \\ &= -\frac{1}{2} \left(\sum_{k \in \mathbb{Z}} e^{2i\pi kR} - \sum_{k=-N}^{N} e^{2i\pi kR} \right) = -\frac{1}{2} \sum_{m \in \mathbb{Z}} \delta_m(R) + \frac{1}{2} \frac{\sin\left((2N+1)\pi R\right)}{\sin(\pi R)}. \end{aligned}$$

Therefore, h_N is smooth on $\mathbb{R} \setminus \mathbb{Z}$. Since it is 1-periodic, it suffices to study it on the open interval $(0,1)$. Since $h_N\left(\frac{1}{2}+t\right) = h_N\left(\frac{1}{2}-t\right)$ for all $|t| < \frac{1}{2}$, we have $h_N'\left(\frac{1}{2}\right) = 0$, so that for all $R \in (0,1)$, and using Taylor formula with integral remainder, we get

$$\begin{aligned} h_N(R) &= h_N\left(\frac{1}{2}\right) + \int_{\frac{1}{2}}^{R} (R-t)\, h_N''(t)\, dt = h_N\left(\frac{1}{2}\right) + \frac{1}{2} \int_{\frac{1}{2}}^{R} (R-t)\, \frac{\sin\left((2N+1)\pi t\right)}{\sin(\pi t)}\, dt \\ &= h_N\left(\frac{1}{2}\right) + \frac{1}{2(2N+1)^2\pi^2} \left((-1)^N - \frac{\sin\left((2N+1)\pi R\right)}{\sin(\pi R)} \right) \\ &\quad - \frac{1}{2(2N+1)^2\pi^2} \int_{\frac{1}{2}}^{R} \left(2\pi \frac{\cos(\pi t)}{\sin(\pi t)} + \frac{(R-t)\pi^2(1+\cos^2(\pi t))}{\sin^2(\pi t)} \right) \frac{\sin\left((2N+1)\pi t\right)}{\sin(\pi t)}\, dt. \end{aligned}$$

Since

$$\left| h_N\left(\frac{1}{2}\right) \right| = \left| \sum_{k=N+1}^{+\infty} \frac{(-1)^k}{4\pi^2 k^2} \right| \leq \frac{1}{4\pi^2 (N+1)^2} \leq \frac{1}{4\pi^2 N^2},$$

and since, for all $R \in (0, 1/2)$,

$$\left| \frac{1}{2(2N+1)^2\pi^2} \left((-1)^N - \frac{\sin\left((2N+1)\pi R\right)}{\sin(\pi R)} \right) \right| \leq \frac{1}{8\pi^2 N^2} \left(1 + \frac{1}{\sin(\pi R)} \right) \leq \frac{1}{4\pi^2 N^2 \sin(\pi R)},$$

$$\left| \int_{\frac{1}{2}}^{R} 2\pi \frac{\cos(\pi t)}{\sin(\pi t)} \frac{\sin\left((2N+1)\pi t\right)}{\sin(\pi t)}\, dt \right| \leq 2\pi \int_{R}^{\frac{1}{2}} \frac{\cos(\pi t)}{\sin^2(\pi t)} dt = 2\left(\frac{1}{\sin(\pi R)} - 1 \right),$$

and, using the inequalities $2t < \sin(\pi t) < \pi t$ for all $0 < t < \frac{1}{2}$,

$$\begin{aligned} \left| \int_{\frac{1}{2}}^{R} \frac{(R-t)\pi^2(1+\cos^2(\pi t))}{\sin^2(\pi t)} \frac{\sin\left((2N+1)\pi t\right)}{\sin(\pi t)}\, dt \right| &\leq 2\pi^2 \int_{R}^{\frac{1}{2}} \frac{t-R}{\sin^3(\pi t)}\, dt \leq \pi^2 \int_{R}^{\frac{1}{2}} \frac{2t}{\sin^3(\pi t)}\, dt \\ &\leq \frac{\pi^2}{4} \int_{R}^{\frac{1}{2}} \frac{1}{t^2} dt \leq \frac{\pi^2}{4R} \leq \frac{\pi^3}{4\sin(\pi R)}, \end{aligned}$$

we finally get

$$\begin{aligned} |\eta_{N,R}| = \left|4\pi^2 N h_N(R)\right| &\leq \frac{1}{N} + \frac{1}{N\sin(\pi R)} + \frac{1}{N}\left(\frac{1}{\sin(\pi R)} - 1 \right) + \frac{\pi^3}{8\sin(\pi R) N} \\ &= \left(2 + \frac{\pi^3}{8} \right) \frac{1}{\sin(\pi R) N}, \end{aligned}$$

which concludes the proof. □

We are now ready to prove Theorem 3.3.2.

Proof of Theorem 3.3.2. Combining Lemmata 3.4.1, 3.4.2, 3.4.3 and 3.4.4, we get that for any $R \in (0,1)$,

$$
\begin{aligned}
E_{R,N} - E_R &= z_1 u_{R,N}(0)(\Pi_N^{\perp} u_R)(0) + z_2 u_{R,N}(R)(\Pi_N^{\perp} u_R)(R) + o\left(\frac{1}{N^{3-\epsilon}}\right) && \text{(Lemma 3.4.2)} \\
&= z_1 u_{R,N}(0)\left(2z_1 u_R(0) f_N(R) + 2z_2 u_R(R) g_N(R)\right) \\
&\quad + z_2 u_{R,N}(R)\left(2z_2 u_R(R) f_N(R) + 2z_1 u_R(0) g_N(R)\right) + o\left(\frac{1}{N^{3-\epsilon}}\right) && \text{(Lemma 3.4.3)} \\
&= \left(2z_1^2 u_{R,N}(0) u_R(0) + 2z_2^2 u_{R,N}(R) u_R(R)\right) f_N(R) \\
&\quad + 2z_1 z_2 \left(u_{R,N}(0) u_R(R) + u_{R,N}(R) u_R(0)\right) g_N(R) + o\left(\frac{1}{N^{3-\epsilon}}\right) \\
&= \left(2z_1^2 u_{R,N}(0) u_R(0) + 2z_2^2 u_{R,N}(R) u_R(R)\right) \frac{1}{4\pi^2 N} a_N \\
&\quad + 2z_1 z_2 \left(u_{R,N}(0) u_R(R) + u_{R,N}(R) u_R(0)\right) \frac{1}{4\pi^2 N} \eta_{R,N} + o\left(\frac{1}{N^{3-\epsilon}}\right) && \text{(Lemma 3.4.4)} \\
&= \frac{\alpha_R}{N} a_N + \frac{\beta_{R,N}^{(1)}}{N} a_N + \frac{\gamma_R}{N^2} \eta_{R,N} + o\left(\frac{1}{N^{3-\epsilon}}\right),
\end{aligned}
$$

where we have used the bounds (3.4.3) and (3.4.11) to obtain the last equality. The proof of (3.3.7) easily follows. □

Acknowledgments

The authors are grateful to Yvon Maday for useful discussions, as well as the anonymous reviewers for interesting suggestions. This work was partially undertaken in the framework of CALSIMLAB, supported by the public grant ANR-11-LABX- 0037-01 overseen by the French National Research Agency (ANR) as part of the Investissements d'avenir program (reference: ANR-11-IDEX-0004-02).

Part III

A posteriori error estimation for the Laplace eigenvalue problem

Chapter 4

Guaranteed and robust a posteriori bounds for Laplace eigenvalues and eigenvectors: conforming approximations

We expose in this chapter the results of [52]. This work was done in collaboration with Eric Cancès, Yvon Maday, Benjamin Stamm and Martin Vohralík.

Abstract

This paper derives a posteriori error estimates for conforming numerical approximations of the Laplace eigenvalue problem with a homogeneous Dirichlet boundary condition. In particular, upper and lower bounds for an arbitrary simple eigenvalue are given. These bounds are guaranteed, fully computable, and converge with optimal speed to the given exact eigenvalue. They are valid without restrictions on the computational mesh or on the approximate eigenvector; we only need to assume that the approximate eigenvalue is separated from the surrounding smaller and larger exact ones, which can be checked in practice. Guaranteed, fully computable, optimally convergent, and polynomial-degree robust bounds on the energy error in the approximation of the associated eigenvector are derived as well, under the same hypotheses. Remarkably, there appears no unknown (solution-, regularity-, or polynomial-degree-dependent) constant in our theory, and no convexity/regularity assumption on the computational domain/exact eigenvector(s) is needed. The multiplicative constant appearing in our estimates depends on (computable estimates of) the gaps to the surrounding exact eigenvalues. Its two improvements are presented. First, it is reduced by a fixed factor under an explicit, a posteriori calculable condition on the mesh and on the approximate eigenvector–eigenvalue pair. Second, when an elliptic regularity assumption on the corresponding source problem is satisfied with known constants, this multiplicative constant can be brought to the optimal value of one. Inexact algebraic solvers are taken into account; the estimates are valid on each iteration and can serve for the design of adaptive stopping criteria. The application of our framework to conforming finite element approximations of arbitrary polynomial degree is provided, along with a numerical illustration on a set of test problems.

4.1 Introduction

Precise numerical approximation of eigenvalues and eigenvectors is crucial in countless applications. Thus, there has been a long-standing interest in answering the question: what is the size of the errors in computed eigenvalues and eigenvectors? This question is usually tackled via a posteriori error estimates. For elliptic source problems such as the Laplace one, conclusive answers are today given by, in particular, the theory of equilibrated fluxes following Prager and Synge [207], see Destuynder and Métivet [87], Braess *et al.* [32], Ern and Vohralík [105], and the references therein. The structure of the *Laplace eigenvalue* problem appears rather richer in comparison with the elliptic source case.

Recently, though, there has been an important progress in obtaining *guaranteed lower bounds* for the *eigenvalues*, especially for the first one: Luo *et al.* [180], Hu *et al.* [140, 141], Carstensen and Gedicke [64], Yang *et al.* [253], or Liu [176] achieve so via the lowest-order nonconforming finite element method, Kuznetsov and Repin [159] and Šebestová and Vejchodský [224, 225] give numerical-method-independent estimates based on flux (functional) estimates, Liu and Oishi [178] elaborate fine a priori approximation estimates for lowest-order conforming finite elements, and, most recently, Xie *et al.* [250] also rely on fluxes. Earlier work comprises Kato [145], Forsythe [109], Weinberger [246], Bazley and Fox [18], Fox and Rheinboldt [110], Moler and Payne [192], Kuttler and Sigillito [155, 156], Still [228], Goerisch and He [118], Plum [205], Behnke *et al.* [21], and Armentano and Durán [10], see also the references therein. Sometimes, though, restrictions may apply. A condition on relative closeness to the (first) eigenvalue is necessary in [159, Remark 3.2], [224, condition (3.6)], and [225, condition (5.23)] (in these references, the bounds actually do not converge with the correct speed); solution of an auxiliary eigenvalue problem for nonconvex domains is requested [178]; potential overestimation on adaptively generated meshes may hamper the bounds of [178, 64, 176], relying on a priori estimates and employing the largest mesh element diameter; an auxiliary global flux problem needs to be solved in [250]; a saturation assumption may be necessary, see the discussion in [140].

The question of precision for both eigenvalues and eigenvectors has also been investigated previously. For conforming finite elements, relying on the a priori error estimates resumed in Babuška and Osborn [14] and Boffi [28], see also the references therein, a posteriori error estimates have been obtained by Verfürth [239], Maday and Patera [183], Larson [162], Heuveline and Rannacher [136], Durán *et al.* [95], Grubišić and Ovall [123], Rannacher *et al.* [210], and Šolín and Giani [227], see also the references therein. These estimates, though, systematically contain uncomputable terms, typically higher order on fine enough meshes. Recently, Wang *et al.* [245] have applied the constitutive relation error methodology to obtain sharp fully computable estimates.

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a polygonal/polyhedral domain with a Lipschitz boundary, and let λ_i, u_i be the eigenvalues and associated eigenvectors of the Laplace operator $-\Delta$ on Ω with Dirichlet boundary conditions. The purpose of the present paper is to derive *guaranteed* and optimally convergent *a posteriori bounds* on both *an arbitrary separated Laplace eigenvalue* and the *associated eigenvector* for conforming (variational) methods. Nonconforming methods including nonconforming, discontinuous Galerkin, or mixed finite elements are treated in Cancès *et al.* [51]. We describe the setting in details in §4.2. §4.3 and §4.4 then contain a collection of equivalence inequalities between respectively the i-th eigenvalue error and the square of the i-th eigenvector energy error, the i-th eigenvector energy error and dual norm of the residual, and between the dual norm of the residual and its computable estimates. These results are valid under one key assumption: λ_{ih} needs to be confined like $\lambda_{i-1} < \lambda_{ih} < \lambda_{i+1}$ (the left inequality of course only needs to hold when $i > 1$), see (4.5.2) below. This can be guaranteed in many cases of practical interest by a domain inclusion argument $\Omega^- \subseteq \Omega \subseteq \Omega^+$ with

known smaller and larger eigenvalues $\lambda_{i-1} \le \lambda_{i-1}(\Omega^-)$ and $\lambda_{i+1}(\Omega^+) \le \lambda_{i+1}$ and by requesting $\overline{\lambda}_{i-1} =: \lambda_{i-1}(\Omega^-) < \lambda_{ih} < \underline{\lambda}_{i+1} := \lambda_{i+1}(\Omega^+)$. Numerical bounds $\overline{\lambda}_{i-1} \ge \lambda_{i-1}$ (typically available during the calculation) and $\underline{\lambda}_{i+1} \le \lambda_{i+1}$ (obtained a on a coarse mesh by the approach of [178, 64, 176]) can also be used, see Remarks 4.5.4 and 4.5.5 below. We also suppose that the approximation spaces consist of appropriate piecewise polynomials. For improved versions of our bounds, we additionally need to check the smallness of the $L^2(\Omega)$-norm of the Riesz representation of the residual, see the *a posteriori calculable* conditions (4.5.6) and (4.5.9) below. These can be always satisfied by refining the computational mesh/increasing the polynomial degree of the approximate solution. Note that no condition of Galerkin orthogonality of the residual to the finite element hat functions needs to be satisfied: the entire analysis is presented in the context of inexact algebraic solvers. Our estimates are valid on each iteration subject to the above inclusion of λ_{ih} and can be used for efficient adaptive stopping criteria of iterative eigenvalue solvers, as promoted in, e.g., Mehrmann and Miedlar [189] or Carstensen and Gedicke [64].

In §4.5, the results of §4.3–§4.4 are turned into actual a posteriori bounds. First, upper and lower bounds for the i-th eigenvalue are given in Theorems 4.5.1 and 4.5.2. For a finite element approximation with an exact algebraic solver for simplicity, we obtain

$$\lambda_{ih} - \eta_i^2 \le \lambda_i \le \lambda_{ih} - \tilde{\eta}_i^2, \tag{4.1.1a}$$

with

$$\eta_i = m_{ih}\|\nabla u_{ih} + \boldsymbol{\sigma}_{ih,\mathrm{dis}}\|, \qquad \tilde{\eta}_i = \tilde{\eta}_i(r_{ih}),$$

being *fully computable quantities*. Here u_{ih} is the approximation of the i-th exact eigenvector u_i, $\|\cdot\|$ is the $L^2(\Omega)$-norm, $\boldsymbol{\sigma}_{ih,\mathrm{dis}}$ is an equilibrated flux reconstruction by mixed finite element local residual problems, and r_{ih} is formed by conforming finite element local residual liftings. The associated eigenvector energy estimates are given next, with Theorem 4.5.7 revealing

$$\|\nabla(u_i - u_{ih})\| \le \eta_i, \qquad \eta_i \le C_i\|\nabla(u_i - u_{ih})\|, \tag{4.1.1b}$$

where C_i is a constant that only depends on λ_1, $\overline{\lambda}_{i-1}$, λ_{ih}, $\underline{\lambda}_{i+1}$, on the space dimension d, and on some Poincaré–Friedrichs-type constant $C_{\mathrm{cont,PF}}$ together with a discrete stability constant C_{st}, both only depending on the shape regularity of the mesh. In particular, C_i is independent of the polynomial degree of u_{ih}, leading to the *polynomial-degree robustness*. Moreover, a computable bound on C_i is given. The constant C_i, however, deteriorates for increasing eigenvalues. We distinguish three different cases. In Cases A and B of Theorems 4.5.1, 4.5.2, and 4.5.7, the multiplicative factor m_{ih} of the estimator η_i contains the factor $\max\{(\frac{\lambda_{ih}}{\overline{\lambda}_{i-1}} - 1)^{-1}, (1 - \frac{\lambda_{ih}}{\underline{\lambda}_{i+1}})^{-1}\}$ and similarly for $\tilde{\eta}_i$; Case B improves the overall size of m_{ih} under the fine-enough-mesh condition (4.5.6). The results of these two cases hold *without any assumption* on the *convexity* of the computational domain Ω and on the *regularity* of the weak solutions. If, additionally, elliptic regularity of the corresponding *source problem* is known, the interpolation and stability constants are computable (typically when $d = 2$ and Ω is convex), and the condition (4.5.9) holds, the factor m_{ih} in front of the principal term $\|\nabla u_{ih} + \boldsymbol{\sigma}_{ih,\mathrm{dis}}\|$ has the optimal behavior $\sqrt{1 + \mathcal{O}(h^2)}$, as summarized in Case C of Theorems 4.5.1, 4.5.2, and 4.5.7.

We show how to apply the above general results to conforming finite elements of arbitrary order in §4.6. Numerical experiments presented in §4.7 fully support the theoretical findings; in particular the necessary conditions hold from quite coarse meshes. We only treat here simple eigenvalues and associated eigenvectors; clustered and multiple eigenvalues will be dealt with in a forthcoming contribution. Finally, building on these results, guaranteed error bounds and fully adaptive strategies with dynamic stopping criteria may become possible for nonlinear eigenvalue problems; some of our first results in this direction are summarized in [49].

4.2 Setting

We denote by $H^1(\Omega)$ the Sobolev space of $L^2(\Omega)$ functions with weak gradients in $[L^2(\Omega)]^d$ and by $V := H^1_0(\Omega)$ its zero-trace subspace. Similarly, $\mathbf{H}(\mathrm{div},\Omega)$ stands for the space of $[L^2(\Omega)]^d$ functions with weak divergences in $L^2(\Omega)$. The notations ∇ and $\nabla\cdot$ are used respectively for the weak gradient and divergence. Moreover, for $\omega \subset \Omega$, $(\nabla u, \nabla v)_\omega$ stands for $\int_\omega \nabla u\cdot\nabla v \,\mathrm{d}\mathbf{x}$ and $(u,v)_\omega$ for $\int_\omega uv\,\mathrm{d}\mathbf{x}$; we also denote $\|\nabla v\|^2_\omega := \int_\omega |\nabla v|^2\,\mathrm{d}\mathbf{x}$ and $\|v\|^2_\omega := \int_\omega v^2\,\mathrm{d}\mathbf{x}$ and drop the index whenever $\omega = \Omega$.

4.2.1 The Laplace eigenvalue problem

We consider here the problem: find eigenvector and eigenvalue pairs (u_k,λ_k), with u_k satisfying a homogeneous Dirichlet boundary condition over $\partial\Omega$ and subject to the constraint $\|u_k\| = 1$, such that $-\Delta u_k = \lambda_k u_k$ in Ω. In a weak form, $(u_k,\lambda_k) \in V\times\mathbb{R}^+$ with $\|u_k\| = 1$ and

$$(\nabla u_k, \nabla v) = \lambda_k (u_k, v) \qquad \forall v \in V. \tag{4.2.1}$$

Actually, cf. Gilbarg and Trudinger [115], Babuška and Osborn [14], Boffi [28], or Strang and Fix [229], u_k, $k\ge 1$, form a *countable orthonormal basis* of $L^2(\Omega)$ consisting of vectors from V, whereas $0<\lambda_1<\lambda_2\le\lambda_3\le\ldots$ going to $+\infty$. The smallest eigenvalue λ_1 is positive and simple and the associated eigenvector u_k to each simple λ_k is unique up to the sign that we fix here by the condition $(u_k,\chi_k)>0$, where $\chi_k\in L^2(\Omega)$ is typically a characteristic function of Ω (for $k=1$) or of its subdomain (for $k>1$). Note that it follows from (4.2.1) and the scaling $\|u_k\|=1$ that $\|\nabla u_k\|^2 = \lambda_k$.

Below, we shall often employ the Parseval identity, giving for any $v\in L^2(\Omega)$

$$\|v\|^2 = \sum_{k\ge1} (v,u_k)^2. \tag{4.2.2}$$

As $(u_k/\sqrt{\lambda_k})_{k\ge1}$ form an orthonormal basis of V, for which one in particular uses that $(\nabla u_k,\nabla u_l) = \lambda_k(u_k,u_l) = 0$ for $k\ne l$, for any $v\in V$, we also obtain

$$\|\nabla v\|^2 = \sum_{k\ge1} \frac{(\nabla v,\nabla u_k)^2}{\lambda_k} = \sum_{k\ge1}\lambda_k (v,u_k)^2. \tag{4.2.3}$$

4.2.2 Residual and its dual norm

The derivation of a posteriori error estimates usually exploits the concept of the *residual* and of its *dual norm*. We will proceed in this way as well. Let V' stand for the dual of V.

Definition 4.2.1 (Residual and its dual norm)**.** *For any pair $(u_{ih},\lambda_{ih})\in V\times\mathbb{R}$, define the residual* $\mathrm{Res}(u_{ih},\lambda_{ih})\in V'$ *by*

$$\langle \mathrm{Res}(u_{ih},\lambda_{ih}), v\rangle_{V',V} := \lambda_{ih}(u_{ih},v) - (\nabla u_{ih},\nabla v) \qquad \forall v\in V. \tag{4.2.4a}$$

Its dual norm is then

$$\|\mathrm{Res}(u_{ih},\lambda_{ih})\|_{-1} := \sup_{v\in V,\,\|\nabla v\|=1} \langle \mathrm{Res}(u_{ih},\lambda_{ih}), v\rangle_{V',V}. \tag{4.2.4b}$$

We will also often work with the *Riesz representation* of the residual $\mathfrak{r}_{(ih)}\in V$,

$$(\nabla \mathfrak{r}_{(ih)},\nabla v) = \langle \mathrm{Res}(u_{ih},\lambda_{ih}), v\rangle_{V',V} \qquad \forall v\in V, \tag{4.2.5a}$$

$$\|\nabla \mathfrak{r}_{(ih)}\| = \|\mathrm{Res}(u_{ih},\lambda_{ih})\|_{-1}. \tag{4.2.5b}$$

4.3 Generic equivalences

In extension of some classical results, see [115, 14, 28, 229], we establish in this section *generic equivalence results* between the following three quantities: the i-th eigenvalue error $\|\nabla u_{ih}\|^2-\lambda_i$, which can potentially be negative, the square of the i-th eigenvector energy error $\|\nabla(u_i-u_{ih})\|^2$, and the square of the dual norm of the residual $\|\mathrm{Res}(u_{ih},\lambda_{ih})\|_{-1}^2$. These equivalences may for the moment contain uncomputable terms like the eigenvalues λ_{i-1}, λ_i, λ_{i+1} or the Riesz representation norm $\|\,\mathscr{r}_{(ih)}\|$, but all such terms will be removed later. To proceed in an abstract way allowing for *inexact algebraic solvers*, we rather work with the eigenvalue error given by $\|\nabla u_{ih}\|^2-\lambda_i$ instead of $\lambda_{ih}-\lambda_i$; of course these coincide when the discrete Rayleigh quotient link $\|\nabla u_{ih}\|^2=\lambda_{ih}$ holds, typically upon solver convergence. A generalization to any self-adjoint operator with compact resolvent can be found in Cancès *et al.* [51].

Our first two lemmas are similar in parts to the developments in [155, 159, 224, 225], giving a computable bound on the $L^2(\Omega)$ error $\|u_i-u_{ih}\|$. Let $i\ge 1$ and define

$$C_{ih}:=\min\left\{\left(1-\frac{\lambda_{ih}}{\lambda_{i-1}}\right)^2,\left(1-\frac{\lambda_{ih}}{\lambda_{i+1}}\right)^2\right\}. \tag{4.3.1}$$

The left term needs to be disregarded for $i=1$.

Lemma 4.3.1 ($L^2(\Omega)$ bound via a quadratic residual inequality)**.** *Let $(u_{ih},\lambda_{ih})\in V\times\mathbb{R}^+$ with $\|u_{ih}\|=1$ and $(u_i,u_{ih})\ge 0$ be the i-th approximate eigenvector-eigenvalue pair, $i\ge 1$. Let λ_i be simple and let $\lambda_{i-1}<\lambda_{ih}$ when $i>1$ and $\lambda_{ih}<\lambda_{i+1}$. Then*

$$\|u_i-u_{ih}\|\le\alpha_{ih}:=\sqrt{2}C_{ih}^{-\frac{1}{2}}\|\,\mathscr{r}_{(ih)}\|. \tag{4.3.2}$$

Proof. Characterizations (4.2.1), (4.2.4a), and (4.2.5a) give

$$(\,\mathscr{r}_{(ih)},u_k)=\frac{(\nabla u_k,\nabla\,\mathscr{r}_{(ih)})}{\lambda_k}=\frac{(\lambda_{ih}(u_{ih},u_k)-(\nabla u_{ih},\nabla u_k))}{\lambda_k}=\left(\frac{\lambda_{ih}}{\lambda_k}-1\right)(u_{ih},u_k). \tag{4.3.3}$$

Consequently, the Parseval equality (4.2.2) with $v=\mathscr{r}_{(ih)}$ yields

$$\|\,\mathscr{r}_{(ih)}\|^2=\sum_{k\ge1}(\,\mathscr{r}_{(ih)},u_k)^2=\sum_{k\ge1}\left(1-\frac{\lambda_{ih}}{\lambda_k}\right)^2(u_{ih},u_k)^2. \tag{4.3.4}$$

Observe that the function $x\in\mathbb{R}^+\mapsto\left(1-\frac{\lambda_{ih}}{x}\right)^2$ reaches its minimum at $x=\lambda_{ih}$ and is decreasing on $(0,\lambda_{ih}]$ and increasing on $[\lambda_{ih},\infty)$. Thus the constant C_{ih} in (4.3.1) equals $\min_{k\ge1,k\ne i}\left(1-\frac{\lambda_{ih}}{\lambda_k}\right)^2$. Further, employing the scalings $\|u_i\|=1$ and $\|u_{ih}\|=1$,

$$(u_{ih}-u_i,u_i)=(u_{ih},u_i)-\|u_i\|^2=(u_{ih},u_i)-\frac{\|u_i\|^2}{2}-\frac{\|u_{ih}\|^2}{2}=-\frac{1}{2}\|u_i-u_{ih}\|^2. \tag{4.3.5}$$

As $(u_i,u_k)=0$ for $k\ge1$, $k\ne i$ from the orthogonality of u_k, elaborating (4.3.4) further while adding and subtracting $C_{ih}(u_{ih}-u_i,u_i)^2$ and using (4.3.1) and (4.3.5) gives

$$\begin{aligned}\|\,\mathscr{r}_{(ih)}\|^2&=\left(\frac{\lambda_{ih}}{\lambda_i}-1\right)^2(u_{ih},u_i)^2+\sum_{k\ge1,k\ne i}\left(1-\frac{\lambda_{ih}}{\lambda_k}\right)^2(u_{ih}-u_i,u_k)^2\\&\ge\left(\frac{\lambda_{ih}}{\lambda_i}-1\right)^2(u_{ih},u_i)^2+C_{ih}\sum_{k\ge1}(u_{ih}-u_i,u_k)^2-C_{ih}(u_{ih}-u_i,u_i)^2\\&=\left(\frac{\lambda_{ih}}{\lambda_i}-1\right)^2(u_{ih},u_i)^2+C_{ih}\|u_i-u_{ih}\|^2-\frac{C_{ih}}{4}\|u_i-u_{ih}\|^4,\end{aligned} \tag{4.3.6}$$

where we have also employed (4.2.2) with $v = u_{ih} - u_i$. Dropping the first (nonnegative and presumably small) term on the right-hand side and denoting $e_{ih} := \|u_i - u_{ih}\|^2$, we conclude the validity of the quadratic residual inequality in e_{ih}

$$\frac{C_{ih}}{4} e_{ih}^2 - C_{ih} e_{ih} + \|\boldsymbol{r}_{(ih)}\|^2 \geq 0. \tag{4.3.7}$$

From the sign assumption $(u_i, u_{ih}) \geq 0$, employing $\|u_i\| = \|u_{ih}\| = 1$,

$$e_{ih} = \|u_i - u_{ih}\|^2 = 2 - 2(u_i, u_{ih}) \leq 2, \tag{4.3.8}$$

so that $C_{ih} e_{ih} \leq 2\|\boldsymbol{r}_{(ih)}\|^2$, i.e., (4.3.2). Note that inspecting more closely the quadratic inequality (4.3.7), the improved bound $e_{ih} \leq 2 - \sqrt{4 - 2\alpha_{ih}^2}$ ($\sqrt{2}$-times better for e_{ih} approaching zero) follows under condition $\|\boldsymbol{r}_{(ih)}\|^2 < C_{ih}$ that we prefer to avoid. □

In addition to (4.3.1), define also (disregarding again the left term for $i = 1$)

$$\widetilde{C}_{ih} := \min \left\{ \lambda_{i-1} \left(1 - \frac{\lambda_{ih}}{\lambda_{i-1}}\right)^2, \lambda_{i+1} \left(1 - \frac{\lambda_{ih}}{\lambda_{i+1}}\right)^2 \right\}. \tag{4.3.9}$$

Lemma 4.3.2 ($L^2(\Omega)$ bound with respect to $\|\nabla \boldsymbol{r}_{(ih)}\|$)**.** *Under the assumptions of Lemma 4.3.1, there also holds*

$$\|u_i - u_{ih}\| \leq \alpha_{ih} := \sqrt{2} \widetilde{C}_{ih}^{-\frac{1}{2}} \|\nabla \boldsymbol{r}_{(ih)}\|. \tag{4.3.10}$$

Proof. Developing (4.2.3) for $v = \boldsymbol{r}_{(ih)}$ via (4.3.3) gives

$$\|\nabla \boldsymbol{r}_{(ih)}\|^2 = \sum_{k \geq 1} \lambda_k (\boldsymbol{r}_{(ih)}, u_k)^2 = \sum_{k \geq 1} \lambda_k \left(1 - \frac{\lambda_{ih}}{\lambda_k}\right)^2 (u_{ih}, u_k)^2. \tag{4.3.11}$$

Next, $\min_{k \geq 1, k \neq i} \lambda_k \left(1 - \frac{\lambda_{ih}}{\lambda_k}\right)^2 = \widetilde{C}_{ih}$. Thus, similarly to (4.3.6)–(4.3.7), with $e_{ih} := \|u_i - u_{ih}\|^2$, $\frac{\widetilde{C}_{ih}}{4} e_{ih}^2 - \widetilde{C}_{ih} e_{ih} + \|\nabla \boldsymbol{r}_{(ih)}\|^2 \geq 0$. One then concludes as in Lemma 4.3.1. □

Recall the sign characterization $(u_i, \chi_i) > 0$ with $\chi_i \in L^2(\Omega)$, $i \geq 1$. The sign condition $(u_i, u_{ih}) \geq 0$ necessary in Lemmas 4.3.1 and 4.3.2 is typically always satisfied; the following lemma can be used for its rigorous verification:

Lemma 4.3.3 (Sign verification)**.** *Let $(u_{ih}, \lambda_{ih}) \in V \times \mathbb{R}^+$ satisfy $\|u_{ih}\| = 1$, $(u_{ih}, \chi_i) > 0$, $\lambda_{i-1} < \lambda_{ih}$ when $i > 1$ and $\lambda_{ih} < \lambda_{i+1}$, and $\alpha_{ih} \leq \|\chi_i\|^{-1}(u_{ih}, \chi_i)$ for α_{ih} given by* (4.3.2) *or* (4.3.10)*. Then the sign condition $(u_i, u_{ih}) \geq 0$ is satisfied.*

Proof. Suppose $-(u_i, u_{ih}) > 0$. Then the bounds of Lemmas 4.3.1 and 4.3.2 hold for $-u_{ih}$ in place of u_{ih}, i.e., $\|u_i + u_{ih}\| \leq \alpha_{ih}$. Consequently, a contradiction follows,

$$(u_{ih}, \chi_i) = -(u_i, \chi_i) + (u_i + u_{ih}, \chi_i) < (u_i + u_{ih}, \chi_i) \leq \|u_i + u_{ih}\| \|\chi_i\| \leq (u_{ih}, \chi_i).$$

□

4.3.1 i-th eigenvalue error equivalences

We first show how to exploit the $L^2(\Omega)$ bound for equivalence between the eigenvalue error and the eigenvector error.

Theorem 4.3.4 (Eigenvalue bounds)**.** *Let $u_{ih} \in V$ with $\|u_{ih}\| = 1$, $i \geq 1$, be arbitrary subject to $\|u_i - u_{ih}\| \leq \alpha_{ih}$ for some $\alpha_{ih} \in \mathbb{R}^+$. Then*

$$\|\nabla(u_i - u_{ih})\|^2 - \lambda_i \alpha_{ih}^2 \leq \|\nabla u_{ih}\|^2 - \lambda_i \leq \|\nabla(u_i - u_{ih})\|^2. \tag{4.3.12}$$

Under the additional assumption $\alpha_{1h}^2 \leq 2$, there also holds, for the first eigenpair,

$$\frac{1}{2}\left(1 - \frac{\lambda_1}{\lambda_2}\right)\left(1 - \frac{\alpha_{1h}^2}{4}\right)\|\nabla(u_1 - u_{1h})\|^2 \leq \|\nabla u_{1h}\|^2 - \lambda_1. \tag{4.3.13}$$

Proof. Using the weak solution characterization (4.2.1) and (4.3.5),

$$\begin{aligned} \|\nabla u_{ih}\|^2 - \lambda_i &= \|\nabla(u_{ih} - u_i)\|^2 + 2(\nabla(u_{ih} - u_i), \nabla u_i) \\ &= \|\nabla(u_{ih} - u_i)\|^2 + 2\lambda_i(u_i, u_{ih} - u_i) \\ &= \|\nabla(u_{ih} - u_i)\|^2 - \lambda_i\|u_i - u_{ih}\|^2. \end{aligned} \tag{4.3.14}$$

Dropping the (nonpositive and presumably small) last term, the upper bound in (4.3.12) follows; estimating it using $\|u_i - u_{ih}\| \leq \alpha_{ih}$, we arrive at the lower bound in (4.3.12).

The bound (4.3.13) only seems to hold for the first eigenpair. To prove it, we use (4.2.2)–(4.2.3) for $v = u_1 - u_{1h}$. First,

$$\|\nabla(u_1 - u_{1h})\|^2 - \lambda_1\|u_1 - u_{1h}\|^2 = \sum_{k\geq 1}(\lambda_k - \lambda_1)(u_1 - u_{1h}, u_k)^2 = \sum_{k\geq 2}(\lambda_k - \lambda_1)(u_1 - u_{1h}, u_k)^2. \tag{4.3.15}$$

Using $\lambda_k \geq \lambda_2$ for $k \geq 2$, $\lambda_2 > \lambda_1$, (4.3.5) for $i = 1$, and the Cauchy–Schwarz inequality,

$$\begin{aligned} \|\nabla(u_1 - u_{1h})\|^2 - \lambda_1\|u_1 - u_{1h}\|^2 &\geq (\lambda_2 - \lambda_1)\sum_{k\geq 1}(u_1 - u_{1h}, u_k)^2 - (\lambda_2 - \lambda_1)(u_1 - u_{1h}, u_1)^2 \\ &= (\lambda_2 - \lambda_1)\|u_1 - u_{1h}\|^2 - \frac{\lambda_2 - \lambda_1}{4}\|u_1 - u_{1h}\|^4. \end{aligned}$$

Using $\|u_1 - u_{1h}\| \leq \alpha_{1h}$ and reemploying (4.2.2) for $v = u_1 - u_{1h}$, we arrive at, second,

$$\begin{aligned} \|\nabla(u_1 - u_{1h})\|^2 - \lambda_1\|u_1 - u_{1h}\|^2 &\geq (\lambda_2 - \lambda_1)\|u_1 - u_{1h}\|^2 - \alpha_{1h}^2\frac{\lambda_2 - \lambda_1}{4}\|u_1 - u_{1h}\|^2 \\ &= \sum_{k\geq 1}(\lambda_2 - \lambda_1)\left(1 - \frac{\alpha_{1h}^2}{4}\right)(u_1 - u_{1h}, u_k)^2. \end{aligned}$$

Summing this with (4.3.15) with weights $\frac{1}{2}$ yields

$$\|\nabla(u_1 - u_{1h})\|^2 - \lambda_1\|u_1 - u_{1h}\|^2 \geq \sum_{k\geq 1}\left\{\frac{\lambda_k - \lambda_1}{2} + \frac{\lambda_2 - \lambda_1}{2}\left(1 - \frac{\alpha_{1h}^2}{4}\right)\right\}(u_1 - u_{1h}, u_k)^2.$$

Now notice that, using (4.2.3) for $v = u_1 - u_{1h}$,

$$\frac{1}{2}\left(1 - \frac{\lambda_1}{\lambda_2}\right)\left(1 - \frac{\alpha_{1h}^2}{4}\right)\|\nabla(u_1 - u_{1h})\|^2 = \sum_{k\geq 1}\frac{\lambda_k}{2}\left(1 - \frac{\lambda_1}{\lambda_2}\right)\left(1 - \frac{\alpha_{1h}^2}{4}\right)(u_1 - u_{1h}, u_k)^2.$$

A simple calculation (note $\frac{1}{2} \le \left(1 - \frac{\alpha_{1h}^2}{4}\right) \le 1$) shows that

$$\frac{\lambda_k - \lambda_1}{2} + \frac{\lambda_2 - \lambda_1}{2}\left(1 - \frac{\alpha_{1h}^2}{4}\right) \ge \frac{\lambda_k}{2}\left(1 - \frac{\lambda_1}{\lambda_2}\right)\left(1 - \frac{\alpha_{1h}^2}{4}\right) \qquad k \ge 1.$$

Thus

$$\|\nabla(u_1 - u_{1h})\|^2 - \lambda_1\|u_1 - u_{1h}\|^2 \ge \frac{1}{2}\left(1 - \frac{\lambda_1}{\lambda_2}\right)\left(1 - \frac{\alpha_{1h}^2}{4}\right)\|\nabla(u_1 - u_{1h})\|^2,$$

and (4.3.13) follows using (4.3.14). □

4.3.2 i-th eigenvector error equivalences

We next investigate the equivalence between the eigenvector error $\|\nabla(u_i - u_{ih})\|$ and the dual norm of the residual $\|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}$. Recall the definition (4.3.1) and also set

$$\overline{C}_{ih} := 1 \text{ if } i = 1, \quad \overline{C}_{ih} := \max\left\{\left(\frac{\lambda_{ih}}{\lambda_1} - 1\right)^2, 1\right\} \text{ if } i > 1. \tag{4.3.16}$$

Furthermore, let

$$\gamma_{ih} := \begin{cases} \|\nabla(u_i - u_{ih})\|^2 & \text{if } \lambda_i \le \|\nabla u_{ih}\|^2 \text{ is known to hold,} \\ \max\{\|\nabla(u_i - u_{ih})\|^2, \lambda_i \alpha_{ih}^2\} & \text{otherwise;} \end{cases} \tag{4.3.17}$$

we refer to Remark 4.5.5 below for the discussion when $\lambda_i \le \|\nabla u_{ih}\|^2$.

Theorem 4.3.5 (Eigenvector bounds)**.** *Let $(u_{ih}, \lambda_{ih}) \in V \times \mathbb{R}^+$ with $\|u_{ih}\| = 1$, $i \ge 1$, be arbitrary subject to $\|u_i - u_{ih}\| \le \alpha_{ih}$ for some $\alpha_{ih} \in \mathbb{R}^+$. Then*

$$\|\nabla(u_i - u_{ih})\|^2 \le \|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}^2 + (\lambda_{ih} + \lambda_i)\alpha_{ih}^2, \tag{4.3.18a}$$

$$\|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}^2 \le \frac{\left(\left|\lambda_{ih} - \|\nabla u_{ih}\|^2\right| + \gamma_{ih}\right)^2}{\lambda_i} + \overline{C}_{ih}\|\nabla(u_i - u_{ih})\|^2. \tag{4.3.18b}$$

Let $\lambda_{i-1} < \lambda_{ih}$ when $i > 1$, $\lambda_{ih} < \lambda_{i+1}$, and $\alpha_{ih}^2 \le 2\frac{\lambda_1}{\lambda_i}$. Then there also holds

$$\|\nabla(u_i - u_{ih})\|^2 \le C_{ih}^{-1}\left(1 - \frac{\lambda_i}{\lambda_1}\frac{\alpha_{ih}^2}{4}\right)^{-1}\|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}^2. \tag{4.3.19}$$

Proof. Starting from (4.3.11), adding and subtracting $C_{ih}\lambda_i(u_{ih} - u_i, u_i)^2$, using $(u_i, u_k) = 0$ for $k \ge 1$, $k \ne i$, (4.3.5), and the Cauchy–Schwarz inequality, we observe

$$\begin{aligned}
\|\nabla \mathscr{r}_{(ih)}\|^2 &\ge \lambda_i\left(\frac{\lambda_{ih}}{\lambda_i} - 1\right)^2 (u_{ih}, u_i)^2 + C_{ih}\sum_{k\ge 1}\lambda_k(u_{ih} - u_i, u_k)^2 - C_{ih}\lambda_i(u_{ih} - u_i, u_i)^2 \\
&= \lambda_i\left(\frac{\lambda_{ih}}{\lambda_i} - 1\right)^2 (u_{ih}, u_i)^2 + C_{ih}\|\nabla(u_i - u_{ih})\|^2 - \frac{C_{ih}}{4}\lambda_i\|u_i - u_{ih}\|^4 \\
&\ge C_{ih}\|\nabla(u_i - u_{ih})\|^2 - \frac{C_{ih}}{4}\lambda_i\|u_i - u_{ih}\|^4.
\end{aligned}$$

Using the Poincaré–Friedrichs inequality $\|u_i - u_{ih}\|^2 \le \frac{1}{\lambda_1}\|\nabla(u_i - u_{ih})\|^2$,

$$\|\nabla \mathscr{r}_{(ih)}\|^2 \ge C_{ih}\|\nabla(u_i - u_{ih})\|^2 - \frac{C_{ih}}{4}\frac{\lambda_i}{\lambda_1}\|\nabla(u_i - u_{ih})\|^2\alpha_{ih}^2,$$

where we have also employed $\|u_i - u_{ih}\| \le \alpha_{ih}$. Thus (4.3.19) follows via (4.2.5b).

The proof of Lemma 4.3.1 gives $\sup_{k\ge1,\,k\neq i}\left(1-\frac{\lambda_{ih}}{\lambda_k}\right)^2 = \overline{C}_{ih}$, recalling (4.3.16). Thus, (4.3.11) together with the Cauchy–Schwarz inequality and $\|u_i\| = \|u_{ih}\| = 1$ give

$$\begin{aligned}\|\nabla \boldsymbol{\mathscr{e}}_{(ih)}\|^2 &\le \lambda_i\left(\frac{\lambda_{ih}}{\lambda_i}-1\right)^2 + \overline{C}_{ih}\sum_{k\ge1,\,k\neq i}\lambda_k(u_{ih}-u_i,u_k)^2\\ &\le \frac{(\lambda_{ih}-\lambda_i)^2}{\lambda_i} + \overline{C}_{ih}\|\nabla(u_i-u_{ih})\|^2.\end{aligned}$$

Using the inequalities (4.3.12) and the definition (4.3.17) of γ_{ih},

$$|\lambda_{ih}-\lambda_i| \le \left|\lambda_{ih}-\|\nabla u_{ih}\|^2\right| + \left|\|\nabla u_{ih}\|^2-\lambda_i\right| \le \left|\lambda_{ih}-\|\nabla u_{ih}\|^2\right| + \gamma_{ih},$$

so that (4.3.18b) is proven.

Finally, (4.3.18a) can be seen as in, e.g., Carstensen and Gedicke [63, Lemma 3.1] combined with $\|u_i - u_{ih}\| \le \alpha_{ih}$. □

4.4 Dual norm of the residual equivalences

We now estimate the dual residual norm $\|\mathrm{Res}(u_{ih},\lambda_{ih})\|_{-1}$, for $u_{ih} \in V$ a piecewise polynomial of degree $p \ge 1$ and $\lambda_{ih} \in \mathbb{R}$. For the upper bound, following [207, 87, 32, 105] and [104, 197, 196] for inexact solvers, see also the references therein, we introduce an *equilibrated flux reconstruction*. This is a vector field $\boldsymbol{\sigma}_{ih}$ constructed from the *local residual* of (u_{ih},λ_{ih}) by solving patchwise mixed finite element problems such that

$$\boldsymbol{\sigma}_{ih} \in \mathbf{V}_h \subset \mathbf{H}(\mathrm{div},\Omega), \tag{4.4.1a}$$

$$\nabla\cdot\boldsymbol{\sigma}_{ih} = \lambda_{ih}u_{ih} - \rho_{ih}, \qquad \lambda_1^{-\frac12}\|\rho_{ih}\| \text{ sufficiently small.} \tag{4.4.1b}$$

Inversely, local conforming residual liftings following [15, §5.1], [213, §4.1.1], [105, §3.3] will allow us to construct $r_{ih} \in X_h \subset V$ leading to a lower bound on $\|\mathrm{Res}(u_{ih},\lambda_{ih})\|_{-1}$.

4.4.1 Meshes and discrete spaces

We first introduce some more notation. Let henceforth $\{\mathcal{T}_h\}_h$ be a family of matching simplicial partitions of the domain Ω, shape regular in the sense that the ratio of each element diameter to the diameter of its largest inscribed ball is uniformly bounded by a constant $\kappa_{\mathcal{T}} > 0$. We denote by K a generic element of $\mathcal{T}_h$. The set of vertices is denoted by $\mathcal{V}_h$, with interior vertices $\mathcal{V}_h^{\mathrm{int}}$, vertices located on the boundary $\mathcal{V}_h^{\mathrm{ext}}$, and a generic vertex $\mathbf{a}$. We call $\mathcal{T}_\mathbf{a}$ the patch of elements of $\mathcal{T}_h$ which share the vertex $\mathbf{a} \in \mathcal{V}_h$, $\omega_\mathbf{a}$ the corresponding subdomain, and $\mathbf{n}_{\omega_\mathbf{a}}$ its outward unit normal. We often tacitly extend functions defined on $\omega_\mathbf{a}$ by zero outside of $\omega_\mathbf{a}$, whereas $V_h(\omega_\mathbf{a})$ stands for the restriction of the space V_h to $\omega_\mathbf{a}$. Next, $\psi_\mathbf{a}$ for $\mathbf{a} \in \mathcal{V}_h$ stands for the piecewise affine "hat" function taking value 1 at the vertex $\mathbf{a}$ and zero at the other vertices. Remarkably, $(\psi_\mathbf{a})_{\mathbf{a}\in\mathcal{V}_h}$ form a partition of unity via $\sum_{\mathbf{a}\in\mathcal{V}_h}\psi_\mathbf{a} = 1|_\Omega$.

Let $\mathbb{P}_s(K)$, $s \ge 0$, stand for polynomials of total degree at most s on $K \in \mathcal{T}_h$, and $\mathbb{P}_s(\mathcal{T}_h)$ for piecewise polynomials on $\mathcal{T}_h$, without any continuity requirement. Let also $\mathbf{V}_h \times Q_h \subset \mathbf{H}(\mathrm{div},\Omega)\times L^2(\Omega)$ stand for the Raviart–Thomas–Nédélec (RTN) mixed finite element spaces of for degree $p+1$, i.e., $\mathbf{V}_h := \{\mathbf{v}_h \in \mathbf{H}(\mathrm{div},\Omega);\, \mathbf{v}_h|_K \in [\mathbb{P}_{p+1}(K)]^d + \mathbb{P}_{p+1}(K)\mathbf{x}\}$ and $Q_h := \mathbb{P}_{p+1}(\mathcal{T}_h)$, see Brezzi and Fortin [36] or Roberts and Thomas [214]. We also denote by Π_{Q_h} the $L^2(\Omega)$-orthogonal projection onto Q_h.

4.4.2 Equilibrated flux reconstruction for inexact solvers

Let $\mathfrak{r}_{ih} \in \mathbb{P}_p(\mathcal{T}_h)$ be a discontinuous piecewise p-degree polynomial that lifts the misfit in the Galerkin orthogonality of the residual $\mathrm{Res}(u_{ih}, \lambda_{ih})$, i.e.

$$\langle \mathrm{Res}(u_{ih}, \lambda_{ih}), v_h \rangle_{V',V} = \lambda_{ih}(u_{ih}, v_h) - (\nabla u_{ih}, \nabla v_h) = (\mathfrak{r}_{ih}, v_h) \qquad \forall v_h \in \mathbb{P}_p(\mathcal{T}_h) \cap V. \quad (4.4.2)$$

A simple elementwise construction of $\mathfrak{r}_{ih}$ is proposed in [197, equation (5.2)]. Typically, $\mathfrak{r}_{ih} = 0$ for an "exact" discrete algebraic solve that we do not suppose here.

We construct $\boldsymbol{\sigma}_{ih}$ in two steps. First, solve the following *homogeneous local Neumann* (Neumann–Dirichlet close to the boundary) discrete *problems* on patches $\omega_{\mathbf{a}}$:

Definition 4.4.1 (Equilibrated flux reconstruction)**.** *For $\mathbf{a} \in \mathcal{V}_h$, set*

$$\begin{aligned} &\mathbf{V}_h^{\mathbf{a}} := \{\mathbf{v}_h \in \mathbf{V}_h(\omega_{\mathbf{a}});\ \mathbf{v}_h{\cdot}\mathbf{n}_{\omega_{\mathbf{a}}} = 0 \text{ on } \partial\omega_{\mathbf{a}}\}, \\ &Q_h^{\mathbf{a}} := \{q_h \in Q_h(\omega_{\mathbf{a}});\ (q_h, 1)_{\omega_{\mathbf{a}}} = 0\}, \end{aligned} \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{int}},$$

$$\begin{aligned} &\mathbf{V}_h^{\mathbf{a}} := \{\mathbf{v}_h \in \mathbf{V}_h(\omega_{\mathbf{a}});\ \mathbf{v}_h{\cdot}\mathbf{n}_{\omega_{\mathbf{a}}} = 0 \text{ on } \partial\omega_{\mathbf{a}} \setminus \partial\Omega\}, \\ &Q_h^{\mathbf{a}} := Q_h(\omega_{\mathbf{a}}), \end{aligned} \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}.$$

Then define $\boldsymbol{\sigma}_{ih,\mathrm{dis}} := \sum_{\mathbf{a}\in\mathcal{V}_h} \boldsymbol{\sigma}_{ih,\mathrm{dis}}^{\mathbf{a}} \in \mathbf{V}_h$, where $\boldsymbol{\sigma}_{ih,\mathrm{dis}}^{\mathbf{a}} \in \mathbf{V}_h^{\mathbf{a}}$ solve

$$\boldsymbol{\sigma}_{ih,\mathrm{dis}}^{\mathbf{a}} := \arg \min_{\substack{\mathbf{v}_h \in \mathbf{V}_h^{\mathbf{a}}, \\ \nabla\cdot\mathbf{v}_h = \Pi_{Q_h}(\lambda_{ih}u_{ih}\psi_{\mathbf{a}} - \nabla u_{ih}\cdot\nabla\psi_{\mathbf{a}} - \mathfrak{r}_{ih}\psi_{\mathbf{a}})}} \|\psi_{\mathbf{a}}\nabla u_{ih} + \mathbf{v}_h\|_{\omega_{\mathbf{a}}} \qquad \forall \mathbf{a} \in \mathcal{V}_h. \quad (4.4.3)$$

Note that the Euler–Lagrange equations for (4.4.3) give the standard *mixed finite element formulation*, cf. [105, Remark 3.7]: find $\boldsymbol{\sigma}_{ih,\mathrm{dis}}^{\mathbf{a}} \in \mathbf{V}_h^{\mathbf{a}}$ and $p_h^{\mathbf{a}} \in Q_h^{\mathbf{a}}$ such that

$$(\boldsymbol{\sigma}_{ih,\mathrm{dis}}^{\mathbf{a}}, \mathbf{v}_h)_{\omega_{\mathbf{a}}} - (p_h^{\mathbf{a}}, \nabla\cdot\mathbf{v}_h)_{\omega_{\mathbf{a}}} = -(\psi_{\mathbf{a}}\nabla u_{ih}, \mathbf{v}_h)_{\omega_{\mathbf{a}}} \qquad \forall \mathbf{v}_h \in \mathbf{V}_h^{\mathbf{a}}, \quad (4.4.4a)$$

$$(\nabla\cdot\boldsymbol{\sigma}_{ih,\mathrm{dis}}^{\mathbf{a}}, q_h)_{\omega_{\mathbf{a}}} = (\lambda_{ih}u_{ih}\psi_{\mathbf{a}} - \nabla u_{ih}\cdot\nabla\psi_{\mathbf{a}} - \mathfrak{r}_{ih}\psi_{\mathbf{a}}, q_h)_{\omega_{\mathbf{a}}} \qquad \forall q_h \in Q_h^{\mathbf{a}}. \quad (4.4.4b)$$

Consequently, $\nabla\cdot\boldsymbol{\sigma}_{ih,\mathrm{dis}} = \lambda_{ih}u_{ih} - \mathfrak{r}_{ih}$, cf., e.g., [196, Lemma 3.6].

Now, proceeding as in [196, Section 3.2], one can construct in a multilevel way a second flux reconstruction $\boldsymbol{\sigma}_{ih,\mathrm{alg}} \in \mathbf{V}_h$ such that $\nabla\cdot\boldsymbol{\sigma}_{ih,\mathrm{alg}} = \mathfrak{r}_{ih}$. Consequently, setting $\boldsymbol{\sigma}_{ih} := \boldsymbol{\sigma}_{ih,\mathrm{dis}} + \boldsymbol{\sigma}_{ih,\mathrm{alg}}$, (4.4.1b) follows with $\rho_{ih} = 0$. Other strategies are pursued in [104, 197], where the algebraic residual is included differently into (4.4.3)/(4.4.4b). These approaches yield

$$\nabla\cdot\boldsymbol{\sigma}_{ih,\mathrm{alg}} = \mathfrak{r}_{ih} - \rho_{ih} \quad (4.4.5)$$

with $\rho_{ih} \neq 0$ and are based on precomputing some algebraic solver iterations in order to ensure that $\|\rho_{ih}\|$ is sufficiently small with respect to the two other contributions in (4.4.9a) below, more precisely verifying (4.4.9b).

4.4.3 Conforming local residual liftings

To estimate $\|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}$ from below, we solve conforming primal counterparts of problems (4.4.4), without the term with $\mathfrak{r}_{ih}$. On each patch $\omega_{\mathbf{a}}$ around the vertex $\mathbf{a} \in \mathcal{V}_h$, define

$$H_*^1(\omega_{\mathbf{a}}) := \{v \in H^1(\omega_{\mathbf{a}});\ (v, 1)_{\omega_{\mathbf{a}}} = 0\}, \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}, \quad (4.4.6a)$$

$$H_*^1(\omega_{\mathbf{a}}) := \{v \in H^1(\omega_{\mathbf{a}});\ v = 0 \text{ on } \partial\omega_{\mathbf{a}} \cap \partial\Omega\}, \quad \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}, \quad (4.4.6b)$$

and let $X_h^{\mathbf{a}}$ be an arbitrary discrete subspace of $H_*^1(\omega_{\mathbf{a}})$, typically $\mathbb{P}_{p+1}(\mathcal{T}_{\mathbf{a}}) \cap H_*^1(\omega_{\mathbf{a}})$.

Definition 4.4.2 (Conforming local Neumann problems)**.** *Define* $r_{ih}^{\mathbf{a}} \in X_h^{\mathbf{a}}$ *by*

$$(\nabla r_{ih}^{\mathbf{a}}, \nabla v_h)_{\omega_{\mathbf{a}}} = \langle \mathrm{Res}(u_{ih}, \lambda_{ih}), \psi_{\mathbf{a}} v_h \rangle_{V',V} \qquad \forall v_h \in X_h^{\mathbf{a}}$$

for each $\mathbf{a} \in \mathcal{V}_h$*. Then set* $r_{ih} := \sum_{\mathbf{a}\in\mathcal{V}_h} \psi_{\mathbf{a}} r_{ih}^{\mathbf{a}}$*.*

The functions $r_{ih}^{\mathbf{a}}$ are discrete Riesz projections of the local residual with hat-weighted test functions. As all $\psi_{\mathbf{a}} r_{ih}^{\mathbf{a}} \in H_0^1(\omega_{\mathbf{a}})$, $r_{ih} \in V$, though $r_{ih}^{\mathbf{a}} \notin V$.

4.4.4 Dual norm of the residual equivalences

Following Carstensen and Funken [61, Theorem 3.1], Braess *et al.* [32, §3], or [105, Lemma 3.12], there exists a constant $C_{\mathrm{cont,PF}}$ only depending on the mesh regularity parameter $\kappa_{\mathcal{T}}$ such that

$$\|\nabla(\psi_{\mathbf{a}} v)\|_{\omega_{\mathbf{a}}} \le C_{\mathrm{cont,PF}} \|\nabla v\|_{\omega_{\mathbf{a}}} \qquad \forall v \in H_*^1(\omega_{\mathbf{a}}),\ \forall \mathbf{a} \in \mathcal{V}_h. \tag{4.4.7}$$

Moreover, the key result of Braess *et al.* [32, Theorem 7], see [106, Corollaries 3.3 and 3.6] for three space dimensions, states that the reconstructions of Definition 4.4.1 satisfy the following *stability* property,

$$\|\psi_{\mathbf{a}} \nabla u_{ih} + \boldsymbol{\sigma}_{ih,\mathrm{dis}}^{\mathbf{a}}\|_{\omega_{\mathbf{a}}} \le C_{\mathrm{st}} \sup_{v\in H_*^1(\omega_{\mathbf{a}});\, \|\nabla v\|_{\omega_{\mathbf{a}}}=1} \{\langle \mathrm{Res}(u_{ih}, \lambda_{ih}), \psi_{\mathbf{a}} v\rangle_{V',V} - (\mathfrak{r}_{ih}, \psi_{\mathbf{a}} v)_{\omega_{\mathbf{a}}}\}. \tag{4.4.8}$$

The constant $C_{\mathrm{st}} > 0$ again only depends on $\kappa_{\mathcal{T}}$, and a computable upper bound on C_{st} is given in [105, Lemma 3.23]. We can summarize the main result of this section:

Theorem 4.4.3 (Residual equivalences)**.** *Let* $(u_{ih}, \lambda_{ih}) \in \mathbb{P}_p(\mathcal{T}_h) \cap V \times \mathbb{R}$ *be arbitrary. Then, for* $\boldsymbol{\sigma}_{ih,\mathrm{dis}}$ *of Definition 4.4.1 and* r_{ih} *of Definition 4.4.2,*

$$\|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1} \le \|\nabla u_{ih} + \boldsymbol{\sigma}_{ih,\mathrm{dis}}\| + \|\boldsymbol{\sigma}_{ih,\mathrm{alg}}\| + \lambda_1^{-\frac{1}{2}} \|\rho_{ih}\|, \tag{4.4.9a}$$

$$\|\nabla u_{ih} + \boldsymbol{\sigma}_{ih,\mathrm{dis}}\| + \|\boldsymbol{\sigma}_{ih,\mathrm{alg}}\| + \lambda_1^{-\frac{1}{2}} \|\rho_{ih}\| \le 3(d+1) C_{\mathrm{st}} C_{\mathrm{cont,PF}} \|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}$$

$$\textit{when } \|\boldsymbol{\sigma}_{ih,\mathrm{alg}}\| + \lambda_1^{-\frac{1}{2}} \|\rho_{ih}\| \le (2(d+1) C_{\mathrm{st}} C_{\mathrm{cont,PF}})^{-1} \|\nabla u_{ih} + \boldsymbol{\sigma}_{ih,\mathrm{dis}}\|, \tag{4.4.9b}$$

$$\frac{\langle \mathrm{Res}(u_{ih}, \lambda_{ih}), r_{ih}\rangle_{V',V}}{\|\nabla r_{ih}\|} \le \|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}. \tag{4.4.9c}$$

Proof. Fix $v \in V$ with $\|\nabla v\| = 1$. Using definition (4.2.4a), adding and subtracting $(\boldsymbol{\sigma}_{ih}, \nabla v)$, and employing the Green theorem and the equilibrium (4.4.1b) yields

$$\langle \mathrm{Res}(u_{ih}, \lambda_{ih}), v\rangle_{V',V} = \lambda_{ih}(u_{ih}, v) - (\nabla u_{ih}, \nabla v) = (\rho_{ih}, v) - (\nabla u_{ih} + \boldsymbol{\sigma}_{ih}, \nabla v).$$

Thus, definition (4.2.4b) of the dual norm of the residual and the Cauchy–Schwarz, Poincaré–Friedrichs, and triangle inequalities yield the bound (4.4.9a). This actually also holds for $\mathbf{V}_h$ being the cheaper RTN space of order p and not $p+1$, as (4.4.1b) still holds. To prove (4.4.9b), we proceed as in [196, Theorem 7.2], while treating the weak norm $\|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1}$ as in Ciarlet and Vohralík [77, Theorems 4.7 and 5.1]. One builds here crucially on inequalities (4.4.7) and (4.4.8) and relies on the choice $p+1$ for $\mathbf{V}_h$. Finally, the bound (4.4.9c) is trivial from (4.2.4b) by taking $v = r_{ih} \in V$. Importantly, this can further be bounded from below by a Hilbertian sum of $\|\nabla r_{ih}^{\mathbf{a}}\|_{\omega_{\mathbf{a}}}$, which can be seen as in [197, proof of Theorem 2]. Thus, this bound is meaningful. □

4.5 Guaranteed and fully computable upper and lower bounds

We combine here the different results of the previous sections to derive the actual guaranteed and fully computable bounds for eigenvalues (in §4.5.1) and eigenvectors (in §4.5.2). A discussion of the results is provided in §4.5.3. We will sometimes use $\zeta_{(ih)} \in V$, the solution of the Laplace *source problem* $-\Delta\zeta_{(ih)} = \imath_{(ih)}$ in Ω, $\zeta_{(ih)} = 0$ on $\partial\Omega$, i.e.,

$$(\nabla\zeta_{(ih)}, \nabla v) = (\imath_{(ih)}, v) \qquad \forall v \in V. \tag{4.5.1}$$

We also denote by $V_h := \mathbb{P}_1(\mathcal{T}_h) \cap V$ the lowest-order conforming finite element space, i.e., the span of $\psi_{\mathbf{a}}$ over all $\mathbf{a} \in \mathcal{V}_h^{\rm int}$, and by h the maximal diameter of all $K \in \mathcal{T}_h$.

4.5.1 Eigenvalues

We first tackle the upper and lower bounds for the i-th eigenvalue λ_i. We discuss the necessary auxiliary bounds below in Remark 4.5.4.

Theorem 4.5.1 (Guaranteed lower bounds for the i-th eigenvalue)**.** *Let the i-th eigenvalue, $i \geq 1$, be simple and suppose the auxiliary bounds $\underline{\lambda}_1 \leq \lambda_1$, $\lambda_i \leq \overline{\lambda}_i$, $\underline{\lambda}_{i+1} \leq \lambda_{i+1}$, as well as $\lambda_{i-1} \leq \overline{\lambda}_{i-1}$ when $i > 1$, for $\underline{\lambda}_1, \overline{\lambda}_i, \underline{\lambda}_{i+1}, \overline{\lambda}_{i-1} > 0$. Let (u_{ih}, λ_{ih}) be any element of $\mathbb{P}_p(\mathcal{T}_h) \cap V \times \mathbb{R}^+$ verifying $\|u_{ih}\| = 1$ and the inequalities*

$$\overline{\lambda}_{i-1} < \lambda_{ih} \text{ when } i > 1, \qquad \lambda_{ih} < \underline{\lambda}_{i+1}. \tag{4.5.2}$$

Let next $\boldsymbol{\sigma}_{ih,\rm dis}$ and r_{ih} be respectively constructed following Definitions 4.4.1 and 4.4.2, let $\boldsymbol{\sigma}_{ih,\rm alg} \in \mathbf{V}_h$ verify (4.4.5) *for an inexact solver, and define*

$$\eta_{i,\rm res} := \|\nabla u_{ih} + \boldsymbol{\sigma}_{ih,\rm dis}\| + \|\boldsymbol{\sigma}_{ih,\rm alg}\| + \underline{\lambda}_1^{-\frac{1}{2}}\|\rho_{ih}\|.$$

Set

$$c_{ih} := \max\left\{\left(\frac{\lambda_{ih}}{\overline{\lambda}_{i-1}} - 1\right)^{-1}, \left(1 - \frac{\lambda_{ih}}{\underline{\lambda}_{i+1}}\right)^{-1}\right\}, \tag{4.5.3a}$$

$$\tilde{c}_{ih} := \max\left\{\overline{\lambda}_{i-1}^{-\frac{1}{2}}\left(\frac{\lambda_{ih}}{\overline{\lambda}_{i-1}} - 1\right)^{-1}, \underline{\lambda}_{i+1}^{-\frac{1}{2}}\left(1 - \frac{\lambda_{ih}}{\underline{\lambda}_{i+1}}\right)^{-1}\right\}, \tag{4.5.3b}$$

with the left terms in the max *disregarded for $i = 1$. Then*

$$\|\nabla u_{ih}\|^2 - \eta_i^2 \leq \lambda_i, \tag{4.5.4}$$

where we distinguish the following three cases:
Case A (No smallness assumption) *If $(u_i, u_{ih}) \geq 0$ is known to hold, define $\overline{\alpha}_{ih} := \sqrt{2}\tilde{c}_{ih}\eta_{i,\rm res}$; if only $(u_{ih}, \chi_i) > 0$ holds, set $\overline{\alpha}_{ih} := \sqrt{2}(1 - \|u_{ih} - \Pi_i u_{ih}\|)^{-\frac{1}{2}}\tilde{c}_{ih}\eta_{i,\rm res}$, where $\Pi_i u_{ih}$ stands for the $L^2(\Omega)$-orthogonal projection of u_{ih} on the span of χ_i. Then* (4.5.4) *holds with*

$$\eta_i^2 := \eta_{i,\rm res}^2 + \big(\lambda_{ih} + \overline{\lambda}_i\big)\overline{\alpha}_{ih}^2. \tag{4.5.5}$$

Case B (Improved estimates under a smallness assumption) *Let $(u_{ih}, \chi_i) > 0$, define $\overline{\alpha}_{ih} := \sqrt{2}\tilde{c}_{ih}\eta_{i,\rm res}$, and request*

$$\overline{\alpha}_{ih} \leq \min\left\{\left(\frac{2\underline{\lambda}_1}{\overline{\lambda}_i}\right)^{\frac{1}{2}}, \|\chi_i\|^{-1}(u_{ih}, \chi_i)\right\}. \tag{4.5.6}$$

Then, (4.5.4) *holds with*

$$\eta_i^2 := c_{ih}^2 \left(1 - \frac{\overline{\lambda}_i}{\underline{\lambda}_1} \frac{\overline{\alpha}_{ih}^2}{4}\right)^{-1} \eta_{i,\mathrm{res}}^2. \tag{4.5.7}$$

Case C (Optimal estimates under elliptic regularity assumption) *Let* $(u_{ih}, \chi_i) > 0$ *and assume that the solution* $\zeta_{(ih)}$ *of problem* (4.5.1) *belongs to the space* $H^{1+\delta}(\Omega)$, $0 < \delta \le 1$, *so that the approximation and stability estimates*

$$\min_{v_h \in V_h} \|\nabla(\zeta_{(ih)} - v_h)\| \le C_{\mathrm{I}} h^{\delta} |\zeta_{(ih)}|_{H^{1+\delta}(\Omega)}, \tag{4.5.8a}$$

$$|\zeta_{(ih)}|_{H^{1+\delta}(\Omega)} \le C_{\mathrm{S}} \|\boldsymbol{\varkappa}_{(ih)}\| \tag{4.5.8b}$$

are satisfied. Define $\overline{\alpha}_{ih} := \sqrt{2} c_{ih} \big[C_{\mathrm{I}} C_{\mathrm{S}} h^{\delta} \eta_{i,\mathrm{res}} + \underline{\lambda}_1^{-\frac{1}{2}} \big(\|\boldsymbol{\sigma}_{ih,\mathrm{alg}}\| + \underline{\lambda}_1^{-\frac{1}{2}} \|\rho_{ih}\|\big)\big]$ *and let*

$$\overline{\alpha}_{ih} \le \|\chi_i\|^{-1} (u_{ih}, \chi_i). \tag{4.5.9}$$

Then (4.5.4) *holds with* η_i^2 *given by* (4.5.5).

Theorem 4.5.2 (Improved guaranteed upper bounds for the i-th eigenvalue)**.** *Let the assumptions of Theorem 4.5.1 be satisfied, with the auxiliary bounds* $\underline{\lambda}_1 \le \lambda_1$, $\underline{\lambda}_i \le \lambda_i \le \overline{\lambda}_i$, *for* $\underline{\lambda}_1, \underline{\lambda}_i, \overline{\lambda}_i > 0$. *Let also* $\lambda_i \le \|\nabla u_{ih}\|^2$, *see Remark 4.5.5 below. Set*

$$\overline{c}_{ih} := 1 \text{ if } i = 1, \quad \overline{c}_{ih} := \max\left\{\left(\frac{\lambda_{ih}}{\underline{\lambda}_1} - 1\right)^2, 1\right\} \text{ if } i > 1,$$

$$d_{ih} := \underline{\lambda}_i^2 \overline{c}_{ih}^2 + 4\underline{\lambda}_i \frac{\langle \mathrm{Res}(u_{ih}, \lambda_{ih}), r_{ih}\rangle_{V',V}^2}{\|\nabla r_{ih}\|^2} + 4\underline{\lambda}_i \overline{c}_{ih} \left|\lambda_{ih} - \|\nabla u_{ih}\|^2\right|.$$

Then

$$\lambda_i \le \|\nabla u_{ih}\|^2 - \tilde{\eta}_i^2, \tag{4.5.10}$$

with, in Cases A and C,

$$\tilde{\eta}_i^2 := \max\left\{-\overline{\lambda}_i \overline{\alpha}_{ih}^2 + \frac{1}{2}\left(\sqrt{d_{ih}} - \left(\underline{\lambda}_i \overline{c}_{ih} + 2\left|\lambda_{ih} - \|\nabla u_{ih}\|^2\right|\right)\right), 0\right\}, \tag{4.5.11}$$

and, in Case B, for $i = 1$ *only,*

$$\tilde{\eta}_1^2 := \max\left\{\frac{1}{4}\left(1 - \frac{\|\nabla u_{1h}\|^2}{\underline{\lambda}_2}\right)\left(1 - \frac{\overline{\alpha}_{1h}^2}{4}\right)\left(\sqrt{d_{1h}} - \left(\underline{\lambda}_1 + 2\left|\lambda_{1h} - \|\nabla u_{1h}\|^2\right|\right)\right), 0\right\}. \tag{4.5.12}$$

Remark 4.5.3 (Exact solvers)**.** *The results of Theorems 4.5.1 and 4.5.2, as well as 4.5.7 below, are presented in a general context of inexact algebraic solvers. For exact solvers, where the algebraic residual representer* $\mathfrak{r}_{ih}$ *in* (4.4.2) *is zero,* $\boldsymbol{\sigma}_{ih,\mathrm{alg}} = \mathbf{0}$, $\rho_{ih} = 0$, *and the condition in* (4.4.9b) *is void. Also, when the Rayleigh quotient link* $\|\nabla u_{ih}\|^2 = \lambda_{ih}$ *holds,* $\|\nabla u_{ih}\|^2$ *can be replaced by* λ_{ih}, *and typically* $\overline{\lambda}_i := \lambda_{ih}$, *see Remark 4.5.5 below.*

Remark 4.5.4 (Auxiliary bounds $\underline{\lambda}_1$, $\underline{\lambda}_i$, and $\underline{\lambda}_{i+1}$)**.** *A straightforward consequence of the* min–max *principle for self-adjoint operators, see, e.g., Gilbarg and Trudinger [115], is that* $\Omega \subseteq \Omega^+ \Rightarrow \lambda_k(\Omega^+) \le \lambda_k$ *and* $\Omega^- \subseteq \Omega \Rightarrow \lambda_k \le \lambda_k(\Omega^-)$ *for all* $k \ge 1$, *where* $\lambda_k(\Omega^{\pm})$ *is the* k*-th eigenvalue on* $\Omega^{\pm}$. *We can then obtain all* $\underline{\lambda}_1$, $\underline{\lambda}_i$, *and* $\underline{\lambda}_{i+1}$ *necessary in Theorem 4.5.1 by this* domain inclusion *for* Ω^+ *with known exact eigenvalues (typically rectangular* d*-parallelepipeds or* d*-spheres, cf. [234]). In what concerns* $\underline{\lambda}_i$, *a very precise choice is to use* $\underline{\lambda}_i := \|\nabla u_{ih}\|^2 - \eta_i^2$,

where η_i^2 was first computed with a rather rough bound $\underline{\lambda}_i$. For $\underline{\lambda}_{i+1}$, if the analytic bounds are too rough to be useful, guaranteed and easily computable numerical bounds *can be used from Liu and Oishi [178] (on convex domains for $d = 2$), Carstensen and Gedicke [64], or Liu [176], typically on a coarse mesh. Finally, as a "practical gratis" strategy for $\underline{\lambda}_{i+1}$, one may simply use $\lambda_{(i+1)h}$ computed by the linear algebra toolbox when solving for (λ_{ih}, u_{ih}), see, e.g., Saad [218] and the references therein. Then Theorems 4.5.1 and 4.5.7 no longer hold* stricto sensu*, but sharp bounds are still observed in practice.*

Remark 4.5.5 (Auxiliary bounds $\overline{\lambda}_{i-1}$ and $\overline{\lambda}_i$)**.** *When (u_{ih}, λ_{ih}) is given by the conforming finite element method of Section 4.6 below, with an exact solver leading to satisfaction of* (4.6.1)*, there holds $\lambda_i \le \lambda_{ih} = \|\nabla u_{ih}\|^2$ and similarly $\lambda_{i-1} \le \lambda_{(i-1)h} = \|\nabla u_{(i-1)h}\|^2$, leading to rather precise auxiliary bounds $\overline{\lambda}_i$ and $\overline{\lambda}_{i-1}$. For the first eigenvalue, there holds $\lambda_1 \le \|\nabla u_{1h}\|^2$ for any $u_{1h} \in H^1_0(\Omega)$. For the i-th eigenvalue, $i > 1$, we in general need to resort to the* min–max *principle giving*

$$\lambda_i \le \max_{\boldsymbol{\xi}\in\mathbb{R}^i,\, \|\boldsymbol{\xi}\|=1} \frac{\|\nabla \sum_{k=1}^i \boldsymbol{\xi}_k u_{kh}\|^2}{\|\sum_{k=1}^i \boldsymbol{\xi}_k u_{kh}\|^2}$$

for an arbitrary linearly independent i-tuple $(u_{1h}, \dots, u_{ih})$, where $\|\boldsymbol{\xi}\|^2 = \sum_{k=1}^i \boldsymbol{\xi}_k^2$.

Remark 4.5.6 (Constants C_{I} and C_{S})**.** *Let Ω be a convex polygon in $\mathbb{R}^2$. Then it is classical that the solution $\zeta_{(ih)}$ of* (4.5.1) *belongs to $H^2(\Omega)$ and $|\zeta_{(ih)}|_{H^2(\Omega)} = \|\Delta\zeta_{(ih)}\| = \|\boldsymbol{\varkappa}_{(ih)}\|$, so that $\delta = 1$ and $C_{\mathrm{S}} = 1$, see Grisvard [122, Theorem 4.3.1.4]. In this situation, calculable bounds on C_{I} can be found in Liu and Kikuchi [177] and Carstensen* et al. *[65], see also Liu and Oishi [178, §2] and the references therein; in particular, for a mesh formed by isosceles right-angled triangles, $C_{\mathrm{I}} \le \frac{0.493}{\sqrt{2}}$.*

We now prove Theorems 4.5.1 and 4.5.2, separately for each case:

Proof (**Case A**). 1) Lower bound of Theorem 4.5.1. If $(u_i, u_{ih}) \ge 0$ is known to hold, we can start from the $L^2(\Omega)$ bound (4.3.10). If this is not the case but $(u_{ih}, \chi_i) > 0$ holds, we first inspect the proof of Lemma 4.3.2 to obtain an alternative $L^2(\Omega)$ estimate. We have $-2(u_i, u_{ih}) = -2(u_i, u_{ih} - \Pi_i u_{ih})\ -2(u_i, \Pi_i u_{ih})$. Note that the second term is negative by the sign assumption $(u_i, \chi_i) > 0$ on u_i. So, instead of (4.3.8), as $\|u_i\| = 1$ and $\|u_{ih} - \Pi_i u_{ih}\| < 1$,

$$\|u_i - u_{ih}\|^2 \le 2 + 2\|u_{ih} - \Pi_i u_{ih}\| =: \delta_{ih} < 4.$$

Consequently, the quadratic inequality in the proof of Lemma 4.3.2 implies $\|u_i - u_{ih}\|^2 \le \|\nabla \boldsymbol{\varkappa}_{(ih)}\|^2 \widetilde{C}_{ih}^{-1}(1 - \delta_{ih}/4)^{-1}$. Thus, the bound (4.4.9a) and assumption (4.5.2) enable us to give a computable upper bound on the $L^2(\Omega)$ error by the estimator $\overline{\alpha}_{ih}$; note that $\min\{a, b\}^{-\frac{1}{2}} = \max\{a^{-\frac{1}{2}}, b^{-\frac{1}{2}}\}$, linking the constant $\widetilde{C}_{ih}$ of (4.3.9) with $\tilde{c}_{ih}$ of (4.5.3b). Consequently, the bound in (4.5.4) follows by combining the upper bounds in (4.3.12), (4.3.18a), and once again (4.4.9a).

2) Upper bound of Theorem 4.5.2. We start from the lower bound in (4.3.12). We then need to bound $\|\nabla(u_i - u_{ih})\|^2$ from below, for which we use (4.3.18b). Relying on the simplifying assumption $\lambda_i \le \|\nabla u_{ih}\|^2$, satisfied namely in cases discussed in Remark 4.5.5, γ_{ih} of (4.3.17) simplifies to $\|\nabla(u_i - u_{ih})\|^2$. Thus (4.3.18b) forms a quadratic inequality for $\|\nabla(u_i - u_{ih})\|^2$, yielding, in combination with (4.4.9c),

$$\|\nabla(u_i - u_{ih})\|^2 \ge \frac{1}{2}\left(\sqrt{d_{ih}} - \left(\underline{\lambda}_i \overline{c}_{ih} + 2\left|\lambda_{ih} - \|\nabla u_{ih}\|^2\right|\right)\right). \tag{4.5.13}$$

Thus (4.5.10) with the estimator (4.5.11) follows. □

Proof (**Case B**). The proof proceeds as above. Note that conditions in (4.5.6) imply that $\alpha_{ih} \leq \sqrt{2\frac{\lambda_1}{\lambda_i}}$ and $\alpha_{ih} \leq \|\chi_i\|^{-1}(u_{ih}, \chi_i)$ for α_{ih} of (4.3.10). We can thus use Lemma 4.3.3 to find that (u_i, u_{ih}) is indeed non-negative, Lemma 4.3.2 for the $L^2(\Omega)$ bound, and the improved estimates (4.3.19) of Theorem 4.3.5 and (4.3.13) of Theorem 4.3.4. For the latter, that seems only to hold for the first eigenpair, we also employ the inequality $1 - \frac{\|\nabla u_{1h}\|^2}{\underline{\lambda}_2} \leq 1 - \frac{\lambda_1}{\lambda_2}$ and (4.5.13) for $i = 1$. □

Proof (**Case C**). The proof is as in Case A (with $(u_i, u_{ih}) \geq 0$), but it relies on Lemma 4.3.1 instead of Lemma 4.3.2. It additionally uses the Aubin–Nitsche trick, cf. [35, Theorem 5.4.8], [122, Theorem 4.3.1.4], or [26]. By (4.5.1), (4.2.5a), and (4.4.2)

$$\|\boldsymbol{\wr}_{(ih)}\|^2 = (\nabla\zeta_{(ih)}, \nabla\boldsymbol{\wr}_{(ih)}) = (\nabla(\zeta_{(ih)} - \zeta_{ih}), \nabla\boldsymbol{\wr}_{(ih)}) + (\mathfrak{r}_{ih}, \zeta_{ih}),$$

where $\zeta_{ih} \in V_h$ is the minimizer in (4.5.8a). Employing (4.4.5), the Green theorem, the Poincaré–Friedrichs inequality $\|\zeta_{ih}\| \leq \underline{\lambda}_1^{-\frac{1}{2}}\|\nabla\zeta_{ih}\|$, and stability $\|\nabla\zeta_{ih}\| \leq \|\nabla\zeta_{(ih)}\|$,

$$(\mathfrak{r}_{ih}, \zeta_{ih}) = -(\boldsymbol{\sigma}_{ih,\mathrm{alg}}, \nabla\zeta_{ih}) + (\rho_{ih}, \zeta_{ih}) \leq \big(\|\boldsymbol{\sigma}_{ih,\mathrm{alg}}\| + \underline{\lambda}_1^{-\frac{1}{2}}\|\rho_{ih}\|\big)\|\nabla\zeta_{(ih)}\|.$$

Noting that (4.5.1) gives $\|\nabla\zeta_{(ih)}\| \leq \underline{\lambda}_1^{-\frac{1}{2}}\|\boldsymbol{\wr}_{(ih)}\|$, the Cauchy–Schwarz inequality, estimates (4.5.8), and the characterization (4.2.5b) altogether give

$$\|\boldsymbol{\wr}_{(ih)}\| \leq C_{\mathrm{I}}C_{\mathrm{S}}h^{\delta}\|\mathrm{Res}(u_{ih}, \lambda_{ih})\|_{-1} + \underline{\lambda}_1^{-\frac{1}{2}}\big(\|\boldsymbol{\sigma}_{ih,\mathrm{alg}}\| + \underline{\lambda}_1^{-\frac{1}{2}}\|\rho_{ih}\|\big).$$

□

4.5.2 Eigenvectors

We now summarize our estimate on the energy error in the approximation of the i-th eigenvector, as well as its efficiency and robustness:

Theorem 4.5.7 (Guaranteed and robust bound for the i-th eigenvector error)**.** *Let the assumptions of Theorem 4.5.1 be verified. Then the energy error can be bounded via*

$$\|\nabla(u_i - u_{ih})\| \leq \eta_i, \tag{4.5.14}$$

where η_i is defined in the Cases A and C by (4.5.5) *and in Case B by* (4.5.7)*, with appropriate $\overline{\alpha}_{ih}$. Under condition* (4.4.9b)*, all these estimators η_i are efficient as*

$$\eta_{i,\mathrm{res}}^2 \leq 3^2(d+1)^2C_{\mathrm{st}}^2C_{\mathrm{cont,PF}}^2\left(\frac{\big(\big|\lambda_{ih} - \|\nabla u_{ih}\|^2\big| + \gamma_{ih}\big)^2}{\lambda_i} + \overline{C}_{ih}\|\nabla(u_i - u_{ih})\|^2\right). \tag{4.5.15}$$

Proof. The guaranteed error bound (4.5.14) follows as in Theorem 4.5.1 upon combining the upper bounds in estimates (4.3.18) or (4.3.19) together with (4.4.9a). The efficiency (4.5.15) is a consequence of (4.4.9b) and of (4.3.18b). □

4.5.3 Comments

We collect here comments about Theorems 4.5.1, 4.5.2, and 4.5.7.

Remark 4.5.8 (Stopping criteria)**.** *The polynomial-degree-robust efficiency* (4.5.15) *holds under the condition* (4.4.9b) *only, which is a typical inexactness (stopping) criterion. For the elliptic regularity Case C, though, it appears wise to rather stop the iterations when $\underline{\lambda}_1^{-\frac{1}{2}}\big(\|\boldsymbol{\sigma}_{ih,\mathrm{alg}}\| + \underline{\lambda}_1^{-\frac{1}{2}}\|\rho_{ih}\|\big)$ is comparable to the first term in $\overline{\alpha}_{ih}$.*

Remark 4.5.9 (Sharpness and comparison of the different bounds of Theorems 4.5.1 and 4.5.7)**.** *The advantage of Case A is that it holds on an* arbitrarily coarse mesh*, provided that only the structural assumption* (4.5.2) *holds. It may, however, lead to a larger overestimation of the error. Case B, under the "fine enough mesh" condition* (4.5.6)*, then significantly improves the multiplicative factor in front of the central term* $\eta_{i,\mathrm{res}} = \|\nabla u_{ih} + \boldsymbol{\sigma}_{ih,\mathrm{dis}}\| + \|\boldsymbol{\sigma}_{ih,\mathrm{alg}}\| + \underline{\lambda}_1^{-\frac{1}{2}}\|\rho_{ih}\|$*, in limit to the factor* c_{ih} *given by* (4.5.3a)*. The bound of Case B still holds without any regularity/convexity/dimension assumption and all the quantities appearing are known. Finally, also the factor* c_{ih} *is asymptotically removed in Case C, when* $\delta > 0$ *and* $h \to 0$*. Here, however, elliptic regularity is needed, see Remark 4.5.6.*

Remark 4.5.10 (Dependence on the maximal element diameter h)**.** *The maximal element diameter* h *is not present at all in Cases A and B of Theorem 4.5.1 and it does not necessarily need to tend to zero in Case C: it only appears as a multiplicative factor of the principal estimator* $\eta_{i,\mathrm{res}}$*. This stands in contrast to previous guaranteed results like [178, Theorem 4.3], [64, Theorem 3.2], or [176, Theorem 2.1].*

Remark 4.5.11 (Polynomial-degree robustness)**.** *The multiplicative factor in the parenthesis in* (4.5.15) *takes the form* $\|\nabla(u_i - u_{ih})\|^2\big(\overline{C}_{ih} + \frac{\|\nabla(u_i-u_{ih})\|^2}{\lambda_i}\big)$ *for an exact algebraic solver in the context of the finite element method* (4.6.1) *below. Noting that* $\frac{\|\nabla(u_i-u_{ih})\|^2}{\lambda_i} \le \frac{2(\lambda_i+\lambda_{ih})}{\lambda_i}$ *(in fact this term becomes negligible with mesh refinement/increasing the polynomial degree), we conclude that the result of Theorem 4.5.7 is fully* robust *with respect to the* polynomial degree p *of* u_{ih}*: all the constants in the comparison between the error* $\|\nabla(u_i - u_{ih})\|$ *and the estimate featuring* $\|\nabla u_{ih} + \boldsymbol{\sigma}_{ih,\mathrm{dis}}\|$ *are independent of* p*. Note, though, that the factor* $\overline{C}_{ih}$ *given by* (4.3.16) *deteriorates for higher eigenvalues.*

Remark 4.5.12 (Error localization and mesh adaptivity)**.** *Since there holds* $\eta_{i,\mathrm{res}}^2 \le 3\sum_{K\in\mathcal{T}_h}\big(\|\nabla u_{ih} + \boldsymbol{\sigma}_{ih,\mathrm{dis}}\|_K^2 + \|\boldsymbol{\sigma}_{ih,\mathrm{alg}}\|_K^2 + \underline{\lambda}_1^{-1}\|\rho_{ih}\|_K^2\big)$*, these local contributions of the estimators of Theorems 4.5.1 and 4.5.7 can directly be used in adaptive mesh refinement based on marking strategies. This is tightly linked to Remark 4.5.10.*

4.6 Application to conforming finite elements

We verify in this section the conditions of the application of our results to the conforming finite element method.

Let $V_h := \mathbb{P}_p(\mathcal{T}_h) \cap V$ for a given polynomial degree $p \ge 1$. In the finite element method, the exact i-th eigenpair $(u_{ih}, \lambda_{ih}) \in V_h \times \mathbb{R}^+$ is such that $(u_{ih}, u_{jh}) = \delta_{ij}$, $1 \le i, j \le \dim V_h$, and

$$(\nabla u_{ih}, \nabla v_h) = \lambda_{ih}(u_{ih}, v_h) \qquad \forall v_h \in V_h, \tag{4.6.1}$$

with the signs ideally fixed by $(u_i, u_{ih}) \ge 0$, practically by $(u_{ih}, \chi_i) > 0$. Thus, upon verifying (4.5.2) and possibly checking (4.5.6) or (4.5.9), all the results of Theorems 4.5.1, 4.5.2, and 4.5.7 hold for any $p \ge 1$. Note that an inexact solution of (4.6.1) in the form (4.4.2) is taken into account. Shall (4.6.1) hold, $\mathfrak{r}_{ih}$ in (4.4.2) vanishes and, moreover, choosing $v_h = u_{ih}$ in (4.6.1) yields $\|\nabla u_{ih}\|^2 = \lambda_{ih}$.

4.7 Numerical experiments

We finally numerically illustrate the estimates of Theorems 4.5.1, 4.5.2, and 4.5.7 on three test cases in $\mathbb{R}^2$, for conforming finite elements (4.6.1) of order $p = 1$. We actually only use the cheaper Raviart–Thomas–Nédélec space of degree $p = 1$ for the flux equilibration instead of

	N	h	ndof	$\underline{\lambda}_2 - \lambda_{1h}$ (4.5.2)	$\|\chi_1\|^{-1}(u_{1h},\chi_1) - \overline{\alpha}_{1h}$ (4.5.9)
$\underline{\lambda}_1{=}1.5\pi^2$ $\underline{\lambda}_2{=}4.5\pi^2$	3	0.4714	16	19.04 (✓)	-0.64 (×)
	4	0.3536	25	21.55 (✓)	0.12 (✓)
	5	0.2828	36	22.69 (✓)	0.40 (✓)
$\underline{\lambda}_1{=}0.5\pi^2$ $\underline{\lambda}_2{=}3\pi^2$	3	0.4714	16	4.233 (✓)	-3.49 (×)
	4	0.3536	25	6.743 (✓)	-0.66 (×)
	5	0.2828	36	7.887 (✓)	0.02 (✓)

Table 4.1 – [Unit square, structured mesh] Validation of assumptions (4.5.2) and (4.5.9)

$p+1$. This still gives guaranteed bounds, see the proof of Theorem 4.4.3, and we do not observe any asymptotic loss of efficiency. The implementation was done in the FreeFem++ code [131]. When we only consider one eigenvalue, it is implicitly assumed that we have chosen $\chi_1 = 1$ for the sign characterization. We consider five test settings with an exact solver and illustrate the use of an inexact solver in a sixth one.

4.7.1 First eigenvalue on the unit square

We start by testing the framework on a unit square $\Omega = (0,1)^2$ and focus on the first eigenvalue. The eigenvalues on a square of size H being $\pi^2(k^2+l^2)/H^2$, $k,l = 1,\dots,\infty$, the first and second eigenvalues are $\lambda_1 = 2\pi^2$ and $\lambda_2 = 5\pi^2$, respectively. In consequence, we can easily choose different $\underline{\lambda}_1 \le \lambda_1$ and $\underline{\lambda}_2 \le \lambda_2$ for the auxiliary eigenvalue bounds and analyze the sensitivity of our results with respect to these choices. The first eigenfunction is given by $u_1(x,y) = \sin(\pi x)\sin(\pi y)$. We focus here on the refined elliptic regularity of Case C, since $d = 2$ and the domain is convex, with constants $C_{\mathrm{S}} = 1$ and $\delta = 1$ given in Remark 4.5.6.

Structured mesh

We first illustrate in Table 4.1 how quickly the computable conditions (4.5.2) and (4.5.9) are satisfied under a uniform refinement of a structured mesh. We take $C_{\mathrm{I}} = \frac{0.493}{\sqrt{2}}$ following Remark 4.5.6 and consider $N = 3,4,5$ subdivisions of each boundary of Ω for the two choices $\underline{\lambda}_1 = 1.5\pi^2$, $\underline{\lambda}_2 = 4.5\pi^2$ and $\underline{\lambda}_1 = 0.5\pi^2$, $\underline{\lambda}_2 = 3\pi^2$, respectively. Note that the finite element space on the coarsest mesh such that all conditions are satisfied contains 25, respectively 36, degrees of freedom only. Indeed, it turns out that our conditions are rather mild.

Next, Figure 4.1 (left) illustrates the convergence of the error $\lambda_{1h} - \lambda_1$ as well as of its lower and upper bounds $\tilde{\eta}_1^2$, η_1^2 given by Case C of Theorems 4.5.1 and 4.5.2. We also plot the eigenfunction energy error $\|\nabla(u_1 - u_{1h})\|$ and its upper bound η_1 of Theorem 4.5.7, Case C. The convergence rates are optimal as expected from the theory.

We present in Table 4.2 precise numbers of the lower and upper bounds $\lambda_{1h} - \eta_1^2 \le \lambda_1 \le \lambda_{1h} - \tilde{\eta}_1^2$ on the exact eigenvalue λ_1, the effectivity indices of the lower and upper bounds $\tilde{\eta}_1^2 \le \lambda_{1h} - \lambda_1 \le \eta_1^2$ of the error $\lambda_{1h} - \lambda_1$, and the effectivity index of the upper bound $\|\nabla(u_1 - u_{1h})\| \le \eta_1$, given respectively by

$$I^{\mathtt{lb}}_{\lambda,\mathtt{eff}} := \frac{\lambda_{1h} - \lambda_1}{\tilde{\eta}_1^2}, \qquad I^{\mathtt{ub}}_{\lambda,\mathtt{eff}} := \frac{\eta_1^2}{\lambda_{1h} - \lambda_1}, \qquad I^{\mathtt{ub}}_{u,\mathtt{eff}} := \frac{\eta_1}{\|\nabla(u_1 - u_{1h})\|}. \tag{4.7.1}$$

We observe rather sharp results, and this also for the relative size of the first eigenvalue confidence interval

$$E_{\lambda,\mathtt{rel}} := 2\frac{(\lambda_{1h} - \tilde{\eta}_1^2) - (\lambda_{1h} - \eta_1^2)}{(\lambda_{1h} - \tilde{\eta}_1^2) + (\lambda_{1h} - \eta_1^2)}. \tag{4.7.2}$$

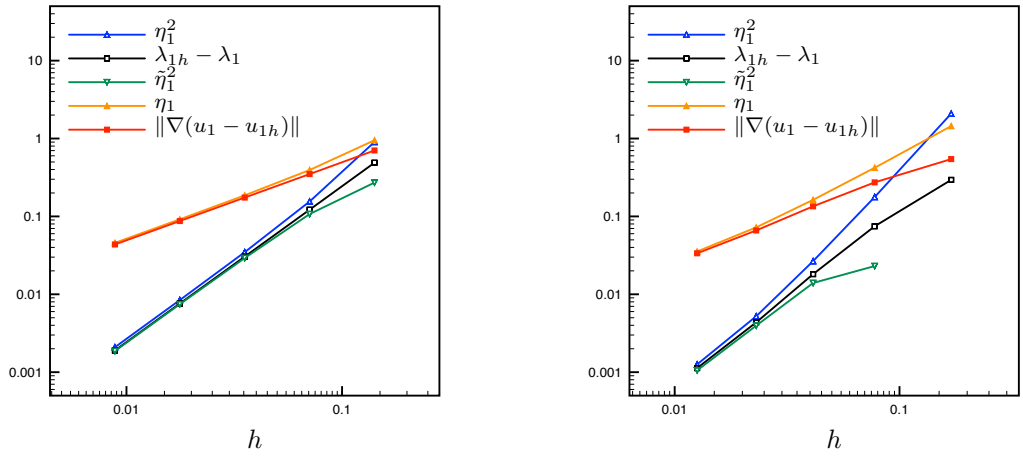

Figure 4.1 – [Unit square] Error in the eigenvalue and eigenvector approximation, its lower bound (eigenvalue only), and its upper bound for the choice $\underline{\lambda}_1 = 1.5\pi^2$, $\underline{\lambda}_2 = 4.5\pi^2$; sequence of structured (left) and unstructured but quasi-uniform (right) meshes; Case C

N	h	ndof	λ_1	λ_{1h}	$\lambda_{1h} - \eta_1^2$	$\lambda_{1h} - \tilde{\eta}_1^2$	$I^{\mathrm{lb}}_{\lambda,\mathrm{eff}}$	$I^{\mathrm{ub}}_{\lambda,\mathrm{eff}}$	$E_{\lambda,\mathrm{rel}}$	$I^{\mathrm{ub}}_{u,\mathrm{eff}}$
10	0.1414	121	19.7392	20.2284	19.3256	19.9566	1.80	1.85	3.21e-02	1.35
20	0.0707	441	19.7392	19.8611	19.7058	19.7539	1.14	1.27	2.44e-03	1.13
40	0.0354	1681	19.7392	19.7697	19.7349	19.7404	1.04	1.14	2.79e-04	1.07
80	0.0177	6561	19.7392	19.7468	19.7384	19.7394	1.02	1.11	4.91e-05	1.05
160	0.0088	25921	19.7392	19.7411	19.7390	19.7392	1.02	1.10	1.14e-05	1.05

Table 4.2 – [Unit square, structured mesh] Lower and upper bounds on the exact eigenvalue λ_1, the effectivity indices, and size of the relative λ_1 confidence interval; $\underline{\lambda}_1 = 1.5\pi^2$, $\underline{\lambda}_2 = 4.5\pi^2$; Case C

Unstructured mesh

Consider now a sequence of unstructured quasi-uniform meshes, obtained by an initial partition of each boundary edge into N intervals. Conditions (4.5.2) and (4.5.9) turn here to be satisfied similarly as in Table 4.1.

The convergence plots for this case are presented in Figure 4.1 (right), showing a similar behavior as for the structured meshes. This time, we use the upper bound on C_{I} according to [177, Eqn. (46)]: $C_{\mathrm{I}} = 0.493 \max_{K\in\mathcal{T}_h} \frac{1+|\cos(\theta_K)|}{\sin(\theta_K)} \sqrt{\frac{\nu_+(\alpha_K,\theta_K)}{2}} \frac{h_K^{[177]}}{h_K}$. We refer to [177] for the definition of $h_K^{[177]}$ and other notation. We observe in Table 4.3 that the results are similar to structured meshes; in particular the case of $\underline{\lambda}_1 = 0.5\pi^2$, $\underline{\lambda}_2 = 3\pi^2$ is less sensitive to the unstructured mesh (not presented).

4.7.2 First eigenvalue on an L-shaped domain: mesh adaptivity

We next consider the L-shaped domain $\Omega := (-1,1)^2 \setminus [0,1]\times[-1,0]$, where $\lambda_1 \approx 9.6397238440$ is known to high accuracy [234]. Including Ω into the square $\Omega^+ = (-1,1)^2$, cf. Remark 4.5.4, we take $\underline{\lambda}_1 = \lambda_1(\Omega^+) = \pi^2/2$, whereas $\underline{\lambda}_2 = 15.1753$ from Table 1 of [176] is employed. We test here the Cases A and B within an adaptive refinement strategy. To do so, we use the local character of our estimators, see Remark 4.5.12. We employ the Dörfler marking with $\theta = 0.6$ and the newest vertex bisection mesh refinement.

Table 4.4 illustrates whether the conditions (4.5.2) and (4.5.6) are satisfied under this adap-

N	h	ndof	λ_1	λ_{1h}	$\lambda_{1h}-\eta_1^2$	$\lambda_{1h}-\tilde{\eta}_1^2$	$I^{\mathrm{lb}}_{\lambda,\mathrm{eff}}$	$I^{\mathrm{ub}}_{\lambda,\mathrm{eff}}$	$E_{\lambda,\mathrm{rel}}$	$I^{\mathrm{ub}}_{u,\mathrm{eff}}$
10	0.1698	143	19.7392	20.0336	17.9458	20.6491	–	7.09	1.40e-01	2.65
20	0.0776	523	19.7392	19.8139	19.6366	19.7909	3.24	2.37	7.83e-03	1.54
40	0.0413	1975	19.7392	19.7573	19.7307	19.7434	1.30	1.47	6.42e-04	1.21
80	0.0230	7704	19.7392	19.7436	19.7383	19.7396	1.10	1.20	6.41e-05	1.09
160	0.0126	30666	19.7392	19.7403	19.7391	19.7393	1.07	1.12	1.04e-05	1.06

Table 4.3 – [Unit square, unstructured mesh] Lower and upper bounds on the exact eigenvalue λ_1, the effectivity indices, and size of the relative λ_1 confidence interval; $\underline{\lambda}_1 = 1.5\pi^2$, $\underline{\lambda}_2 = 4.5\pi^2$; Case C

Level	h	ndof	$\underline{\lambda}_2 - \lambda_{1h}$ (4.5.2)	$\overline{\alpha}_{1h}\sqrt{\lambda_{1h}/2\underline{\lambda}_1}$ (4.5.6)	$\|\chi_1\|^{-1}(u_{1h},\chi_1) - \overline{\alpha}_{1h}$ (4.5.6)
1	0.7500	22	1.8223 (×)	2.97 (×)	-6.17 (×)
4	0.7071	34	3.8799 (✓)	0.94 (✓)	-1.27 (×)
10	0.5000	140	5.2053 (✓)	0.33 (✓)	0.13 (✓)

Table 4.4 – [L-shaped domain, adaptive mesh refinement] Validation of the assumptions (4.5.2) and (4.5.6) for $\underline{\lambda}_1 = \pi^2/2$ and $\underline{\lambda}_2 = 15.1753$

tive refinement. Figure 4.2 (right) illustrates the error in the eigenvalue and the eigenvector and their bounds (4.5.4), (4.5.10), and (4.5.14). Optimal convergence rates are indicated by dashed lines. The initial mesh is structured with 22 degrees of freedom and the conditions (4.5.2) and (4.5.6) are all satisfied starting from 140 degrees of freedom. The transition from Case A to Case B in Theorems 4.5.1, 4.5.2, and 4.5.7 is marked by a dotted line. Figure 4.2 (left) then depicts an adaptively refined mesh and Table 4.5 presents more details on the errors and efficiencies.

4.7.3 First eigenvalue on a domain with a hole: mesh adaptivity

We next consider a domain with a polygonal hole, see Figure 4.3 (left) illustrating the mesh used at iteration 20 of our adaptive mesh refinement strategy. The lower bounds $\underline{\lambda}_1$ and $\underline{\lambda}_2$ on the first and second eigenvalue have been obtained once and for all before starting the adaptive algorithm following the estimates derived in [176], on a uniform mesh with 1143 nodes. Figure 4.3 (right) shows the interval between our lower $(\lambda_{1h}-\eta_1^2)$ and upper $(\lambda_{1h}-\tilde{\eta}_1^2)$ bounds on the first eigenvalue, relying on Case B of Theorems 4.5.1 and 4.5.2, whose assumptions hold starting from 2494 degrees of freedom; Table 4.6 states the numbers. Note that the interval size $(\lambda_{1h}-\tilde{\eta}_1^2)-(\lambda_{1h}-\eta_1^2) = \eta_1^2-\tilde{\eta}_1^2$ behaves like $1/\texttt{ndof}$.

4.7.4 Higher eigenvalues

We now test the upper and lower bounds for higher eigenvalues. First we consider the unit triangle with vertices $(0,0)$, $(1,0)$, $(0,1)$ and a family of structured meshes. The auxiliary lower bounds are obtained by a computation on a fixed coarse mesh with 2145 triangles following [176], which results in

$$\underline{\lambda}_1 = 49.2883, \quad \underline{\lambda}_2 = 98.4296, \quad \underline{\lambda}_3 = 127.937, \quad \underline{\lambda}_4 = 166.975, \quad \underline{\lambda}_5 = 196.439.$$

Figure 4.4 gives the convergence plots for the first four eigenvalues and Table 4.7 provides more details on absolute numbers and efficiency. As the domain is convex (case C), we obtain excellent upper bounds for the error in all four eigenvalue/eigenvector pairs. The lower bound of the eigenvalue error (the improved eigenvalue upper bound of Theorem 4.5.2) is, however, degrading for higher eigenvalues.

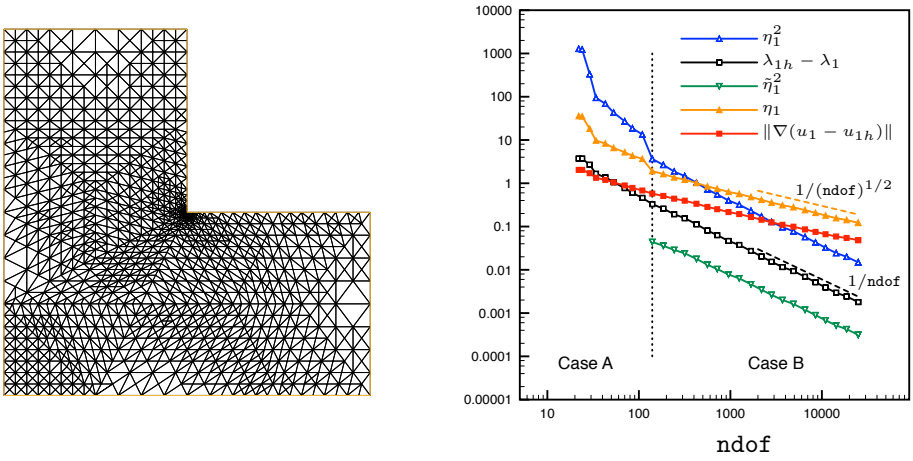

Figure 4.2 – [L-shaped domain, adaptive mesh refinement] Mesh of the adaptive algorithm on step 18 (left) and error in the first eigenvalue and eigenvector approximation, its lower bound (eigenvalue only), and its upper bound (right); Cases A and B

Level	ndof	λ_1	λ_{1h}	$\lambda_{1h} - \eta_1^2$	$\lambda_{1h} - \tilde{\eta}_1^2$	$I^{\mathrm{lb}}_{\lambda,\mathrm{eff}}$	$I^{\mathrm{ub}}_{\lambda,\mathrm{eff}}$	$E_{\lambda,\mathrm{rel}}$	$I^{\mathrm{ub}}_{u,\mathrm{eff}}$
10	140	9.6397	9.9700	6.3175	9.9260	7.50	11.06	4.44e-01	3.31
15	561	9.6397	9.7207	9.0035	9.7075	6.17	8.86	7.53e-02	2.98
20	2188	9.6397	9.6601	9.4887	9.6566	5.88	8.43	1.75e-02	2.88
25	8513	9.6397	9.6449	9.6019	9.6440	5.77	8.31	4.37e-03	2.75
30	24925	9.6397	9.6415	9.6266	9.6412	5.73	8.26	1.51e-03	2.51

Table 4.5 – [L-shaped domain, adaptive mesh refinement] Lower and upper bounds on the first exact eigenvalue λ_1, the effectivity indices, and the size of the relative λ_1 confidence interval; $\underline{\lambda}_1 = \pi^2/2$ and $\underline{\lambda}_2 = 15.1753$, Case B

We now apply the same setting to the L-shaped domain where we obtain again the auxiliary lower bounds by the method presented in [176] for a coarse structured mesh with 3201 triangles resulting in

$$\underline{\lambda}_1 = 9.60692, \quad \underline{\lambda}_2 = 15.1695, \quad \underline{\lambda}_3 = 19.6932, \quad \underline{\lambda}_4 = 29.4166, \quad \underline{\lambda}_5 = 31.7363.$$

Figure 4.5 plots the convergence of the errors the estimators whereas Table 4.8 provides more details on the efficiency. We now observe that the efficiency also degrades for the upper bound of the eigenvalue and eigenvector error. Further, improved lower bounds of the eigenvalue error are not available for the considered meshes for $i > 1$. This appears as the resulting $\tilde{\eta}_i$ are all equal to zero, see (4.5.11), respectively (4.5.12), so that our eigenvalue upper bound stays that of the finite element method. For all meshes and all considered eigenvalues, though, our estimates still give a rather tight guaranteed eigenvalue confidence interval and quite reasonable eigenvector effectivity indices. We can also observe by a jump of the blue curve (η_i^2) the change between the cases A and B. The critical mesh size where this change occurs seems to degrade with increasing eigenvalues.

4.7.5 Inexact algebraic eigenvalue solvers

We finally consider inexact eigenvalue solvers. Since we are using FreeFem++, we rely on an algebraic eigenvalue solver based on the ARPACK package that is built in FreeFem++. Here a user-specified tolerance can be provided and we choose it in a mesh-dependent way as

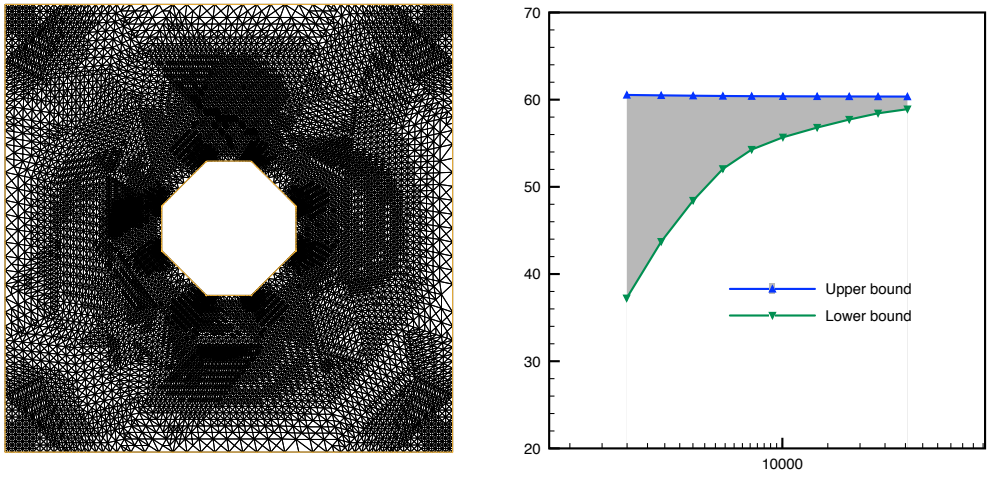

Figure 4.3 – [Domain with a hole, adaptive mesh refinement] Mesh of the adaptive algorithm at iteration 20 (left) and the lower and upper bounds for the exact eigenvalue λ_1 (right); Case B

`ndof`	2494	3390	4508	5879	7602	10047	13640	18163	23494	30533
$\lambda_{1h} - \tilde{\eta}_1^2$	60.541	60.494	60.455	60.422	60.401	60.387	60.376	60.367	60.359	60.354
$\lambda_{1h} - \eta_1^2$	37.223	43.710	48.428	52.058	54.275	55.680	56.799	57.719	58.436	58.910

Table 4.6 – [Domain with a hole, adaptive mesh refinement] Lower and upper bounds on the exact eigenvalue λ_1 as a function of the degrees of freedom; Case B

$\texttt{tol}(h) = h^2$ to materialize an inexact solver. We set $\boldsymbol{\sigma}_{ih,\mathrm{dis}}$ following Definition 4.4.1. In order to compute $\boldsymbol{\sigma}_{ih,\mathrm{alg}}$ in (4.4.5), we proceed as in [197] and the references therein and first compute a second reconstructed flux $\hat{\boldsymbol{\sigma}}_{ih,\mathrm{dis}}$ corresponding to some additional algebraic iterations (here corresponding to the tolerance $h^2/100$ in ARPACK); then $\boldsymbol{\sigma}_{ih,\mathrm{alg}} := \hat{\boldsymbol{\sigma}}_{ih,\mathrm{dis}} - \boldsymbol{\sigma}_{ih,\mathrm{dis}}$. Figure 4.6 demonstrates that we still obtain excellent lower and upper bounds. Adaptive stopping criteria of the form (4.4.9b), leading to savings in algebraic solver iterations, are not investigated here.

4.7.6 Comparison with existing results

We finally compare our results with some existing ones from [64, 178, 176]. In what concerns the unit square and the first eigenvalue of Section 4.7.1, our estimates appear sharper while comparing Table 4.2 with the estimates presented in [64], see Figure 6.2 therein. For the L-shaped domain and uniformly refined meshes of Section 4.7.4 for the first eigenvalue, we also obtain better results than those presented in [64, Figure 6.4], where an efficiency issue appears; compared to the results presented in [178, Table 5.5], we observe that our lower bound $\lambda_{1h} - \eta_1^2$ of the exact eigenvalue is a little less sharp, whereas the upper bound $\lambda_{1h} - \tilde{\eta}_1^2$ is not present in [178]. Recall also from §4.1 that our estimates are much cheaper here than those of [178] (there is no auxiliary eigenvalue problem to solve). For adaptive meshes, we observe that our efficiency of the confidence interval for the first eigenvalue as measured in [64] by $\frac{1}{2}(\eta_1^2 - \tilde{\eta}_1^2)/|\lambda_1 - \lambda_{1h} + \frac{1}{2}(\tilde{\eta}_1^2 + \eta_1^2)|$ is approaching 1.086 which is much better than in [64, Figure 6.5].

To facilitate the comparisons, we finally present in Tables 4.9 and 4.10 several methods for the tests of [178, Table 5.2 ($h = 1/64$) and Table 5.3 ($h = 1/32$)]. We compare in particular the approach presented in this article, lowest-order conforming finite elements from [178], and the lowest-order Crouzeix–Raviart (CR) method presented in [64], with explicit upper bound of the interpolation constants derived in either [62] or [176]. For the eigenvalue upper bounds

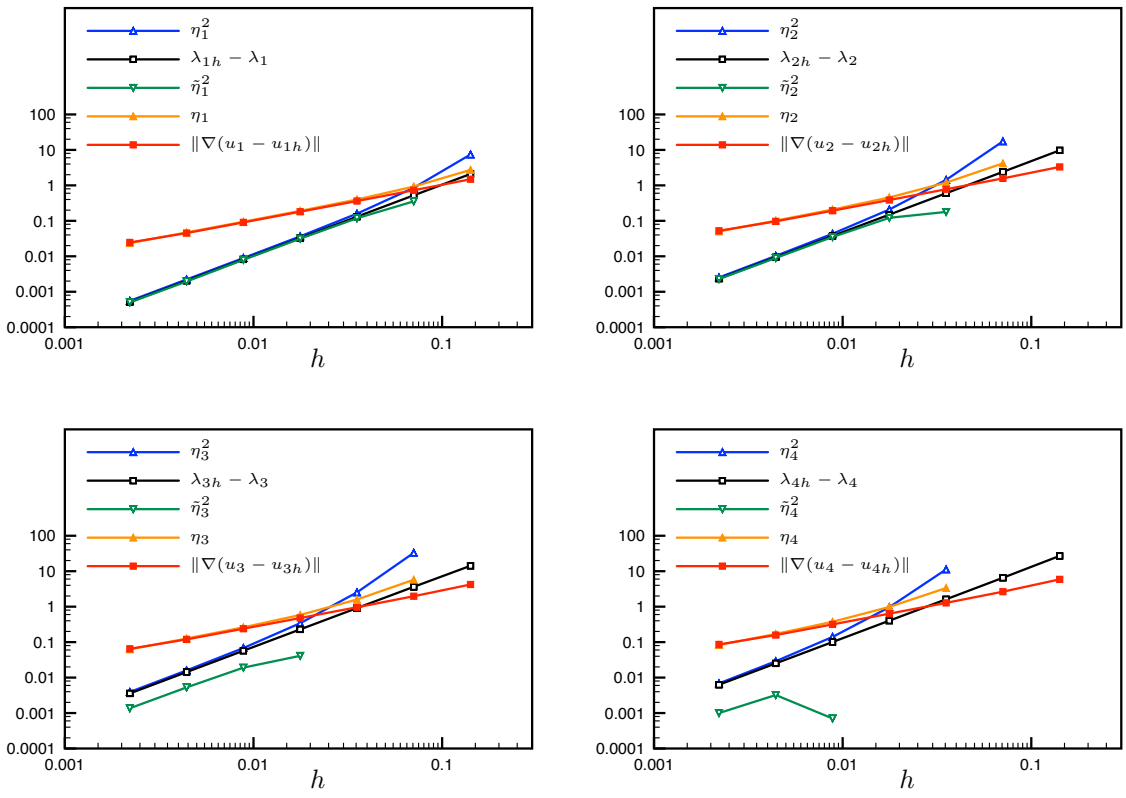

Figure 4.4 – [Triangular domain, structured meshes] Errors in the first four eigenvalue and eigenvector approximations, their lower bounds (eigenvalues only), and their upper bounds

in the CR case, we evaluate the Rayleigh quotient on the $\mathbb{P}_1$ conforming nodal averaging of the original eigenvectors.

On the convex triangle, the present approach seems to give the sharpest results, whereas on the L-shaped domain, the method based on the CR finite elements with the constant from [176] is better for the lower bound. Recall, though, that important advantages of the present theory are that it additionally gives a guaranteed control of the eigenvector error by the same estimators, is not specific to a particular scheme but yields general results that are here applied to any order conforming finite element method and extended in [51] to basically any numerical scheme, and achieves polynomial-degree robustness. It can also be noted that the present estimators take elementwise form immediately suitable for adaptive mesh refinement.

N	h	ndof	λ_i	λ_{ih}	$\lambda_{ih} - \eta_i^2$	$\lambda_{ih} - \tilde{\eta}_i^2$	$I^{\mathtt{lb}}_{\lambda,\mathtt{eff}}$	$I^{\mathtt{ub}}_{\lambda,\mathtt{eff}}$	$E_{\lambda,\mathtt{rel}}$	$I^{\mathtt{ub}}_{u,\mathtt{eff}}$
40	0.0354	861	49.3480	49.4789	49.3197	49.3607	1.11	1.22	8.29e-04	1.10
80	0.0177	3321	49.3480	49.3807	49.3442	49.3493	1.04	1.12	1.03e-04	1.06
160	0.0088	13041	49.3480	49.3562	49.3473	49.3482	1.03	1.09	1.94e-05	1.05
320	0.0044	51681	49.3480	49.3501	49.3478	49.3481	1.05	1.08	5.49e-06	1.04
640	0.0022	205761	49.3480	49.3485	49.3480	49.3480	1.02	1.08	1.07e-06	1.02
40	0.0354	861	98.6960	99.2953	97.8659	99.1171	3.36	2.39	1.27e-02	1.54
80	0.0177	3321	98.6960	98.8457	98.6376	98.7242	1.23	1.39	8.77e-04	1.18
160	0.0088	13041	98.6960	98.7335	98.6903	98.6985	1.07	1.15	8.29e-05	1.08
320	0.0044	51681	98.6960	98.7054	98.6952	98.6964	1.04	1.10	1.29e-05	1.05
640	0.0022	205761	98.6960	98.6984	98.6959	98.6961	1.03	1.08	2.54e-06	1.02
40	0.0354	861	128.3049	129.2175	126.6899	129.2175	–	2.77	2.30e-02	1.65
80	0.0177	3321	128.3049	128.5334	128.1923	128.4923	5.56	1.49	2.34e-03	1.22
160	0.0088	13041	128.3049	128.3620	128.2940	128.3429	3.00	1.19	3.81e-04	1.09
320	0.0044	51681	128.3049	128.3191	128.3032	128.3139	2.70	1.12	8.30e-05	1.06
640	0.0022	205761	128.3049	128.3084	128.3045	128.3071	2.62	1.10	1.99e-05	1.03
40	0.0354	861	167.7833	169.3980	158.1506	169.3980	–	6.97	9.48e-02	2.62
80	0.0177	3321	167.7833	168.1858	167.2205	168.1858	–	2.40	6.94e-03	1.55
160	0.0088	13041	167.7833	167.8838	167.7437	167.8831	142.86	1.39	8.31e-04	1.18
320	0.0044	51681	167.7833	167.8084	167.7795	167.8052	7.80	1.15	1.53e-04	1.07
640	0.0022	205761	167.7833	167.7896	167.7827	167.7886	6.29	1.09	3.49e-05	1.02

Table 4.7 – [Triangular domain, uniform mesh refinement] Lower and upper bounds on the first four exact eigenvalues λ_i, the effectivity indices, and the sizes of the relative λ_i confidence intervals

N	h	ndof	λ_i	λ_{ih}	$\lambda_{ih} - \eta_i^2$	$\lambda_{ih} - \tilde{\eta}_i^2$	$I^{\mathtt{lb}}_{\lambda,\mathtt{eff}}$	$I^{\mathtt{ub}}_{\lambda,\mathtt{eff}}$	$E_{\lambda,\mathtt{rel}}$	$I^{\mathtt{ub}}_{u,\mathtt{eff}}$
40	0.0839	1437	9.6397	9.6955	9.1450	9.6870	6.59	9.87	5.76e-02	3.16
80	0.0459	5674	9.6397	9.6588	9.4719	9.6559	6.52	9.78	1.92e-02	3.15
160	0.0234	21878	9.6397	9.6467	9.5779	9.6456	6.58	9.86	7.04e-03	3.16
320	0.0125	86810	9.6397	9.6423	9.6167	9.6419	6.63	9.93	2.62e-03	3.14
640	0.0059	352256	9.6397	9.6407	9.6310	9.6406	6.73	9.98	9.94e-04	2.77
40	0.0839	1437	15.1973	15.2440	14.2080	15.2440	–	22.17	1.64e-01	4.70
80	0.0459	5674	15.1973	15.2092	14.9577	15.2092	–	21.11	4.09e-02	4.60
160	0.0234	21878	15.1973	15.2002	15.1378	15.2002	–	20.87	1.02e-02	4.57
320	0.0125	86810	15.1973	15.1980	15.1825	15.1980	–	20.81	2.55e-03	4.55
640	0.0059	352256	15.1973	15.1974	15.1936	15.1974	–	20.81	6.36e-04	4.09
40	0.0839	1437	19.7392	19.8216	18.7524	19.8216	–	12.97	1.75e-01	3.59
80	0.0459	5674	19.7392	19.7597	19.5056	19.7597	–	12.38	4.44e-02	3.52
160	0.0234	21878	19.7392	19.7444	19.6805	19.7444	–	12.23	1.14e-02	3.50
320	0.0125	86810	19.7392	19.7405	19.7246	19.7405	–	12.19	2.84e-03	3.48
640	0.0059	352256	19.7392	19.7395	19.7356	19.7395	–	12.20	7.01e-04	3.07
40	0.0839	1437	29.5215	29.7057	-154.6818	29.7057	–	1000.68	–	31.53
80	0.0459	5674	29.5215	29.5675	-10.5379	29.5675	–	871.81	3.08e+00	29.51
160	0.0234	21878	29.5215	29.5331	26.5255	29.5331	–	258.67	2.59e-01	16.08
320	0.0125	86810	29.5215	29.5244	28.8467	29.5244	–	231.37	6.37e-02	15.16
640	0.0059	352256	29.5215	29.5222	29.3595	29.5222	–	225.32	1.56e-02	13.45

Table 4.8 – [L-shaped domain, uniform mesh refinement] Lower and upper bounds on the first four exact eigenvalues λ_i, the effectivity indices, and the sizes of the relative λ_i confidence intervals

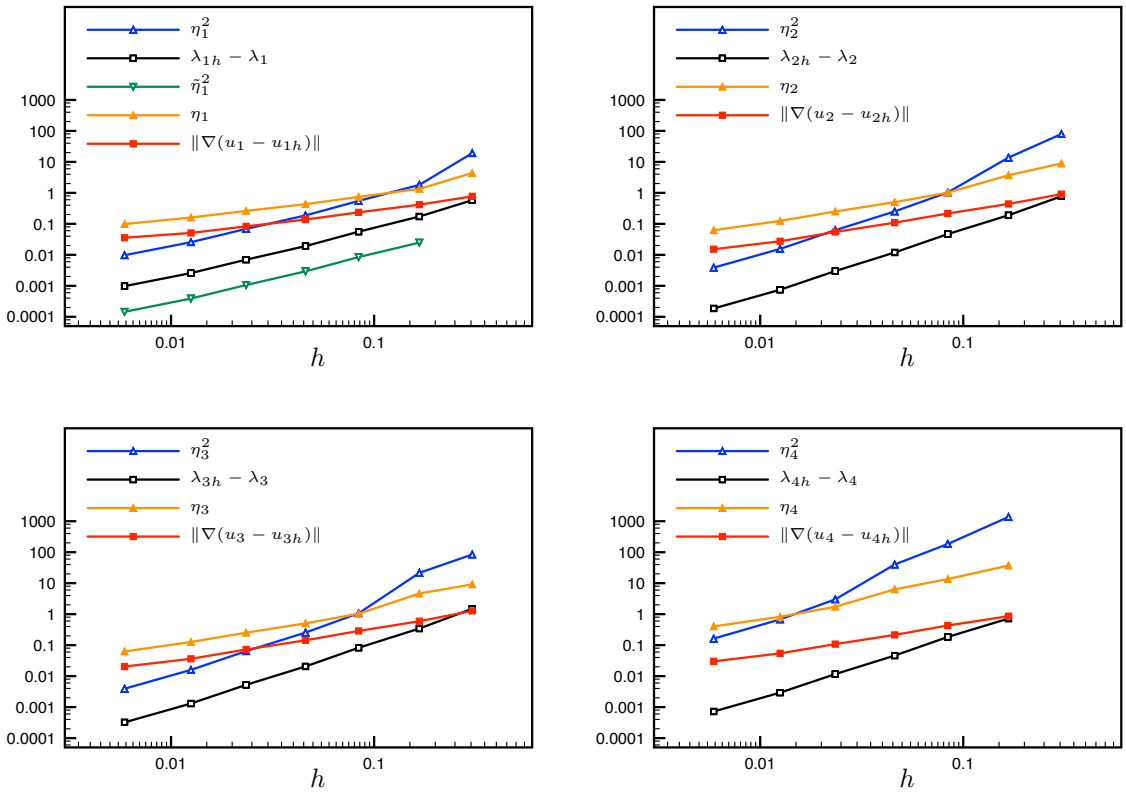

Figure 4.5 – [L-shaped domain, unstructured meshes] Error in the first four eigenvalue and eigenvector approximations, their lower bounds (eigenvalues only), and their upper bounds

$\lambda_1 = 49.348$	in this work	Liu&Oishi [178]	CR with [176]	CR with [62]
Lower bound:	49.341	49.254	49.288	49.225
Upper bound:	49.351	49.400	49.402	

$\lambda_2 = 98.696$	in this work	Liu&Oishi [178]	CR with [176]	CR with [62]
Lower bound:	98.562	98.352	98.430	98.179
Upper bound:	98.762	98.931	98.944	

Table 4.9 – [Triangular domain, structured meshes] Comparison of different methods; CR is the Crouzeix–Raviart method based approach presented in [64] and the constants indicated in the reference

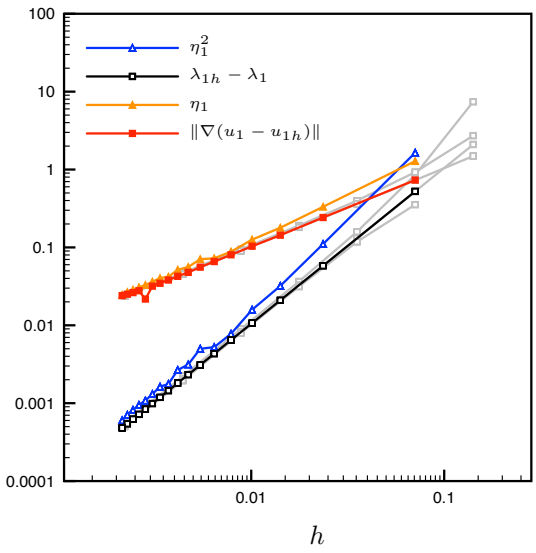

Figure 4.6 – [Triangular domain, structured meshes, inexact solver] Error in the first eigenvalue and eigenvector approximation, its lower bound (eigenvalues only), and its upper bound for a uniform refinement; the convergence plots for an exact solver are indicated in gray

$\lambda_1 = 9.6380$	in this work	Liu&Oishi [178]	CR with [176]	CR with [65]
Lower bound:	9.380	9.559	9.609	9.600
Upper bound:	9.665	9.670	9.682	

$\lambda_2 = 15.197$	in this work	Liu&Oishi [178]	CR with [176]	CR with [65]
Lower bound:	14.632	14.950	15.175	15.152
Upper bound:	15.225	15.225	15.226	

Table 4.10 – [L-shaped domain, structured meshes] Comparison of different methods; CR is the Crouzeix–Raviart method based approach presented in [64] and the constants indicated in the reference

Chapter 5

Guaranteed and robust a posteriori bounds for Laplace eigenvalues and eigenvectors: a unified framework

We expose in this chapter the results of [51]. This work was done in collaboration with Eric Cancès, Yvon Maday, Benjamin Stamm and Martin Vohralík.

Abstract

This paper develops a general framework for a posteriori error estimates in numerical approximations of the Laplace eigenvalue problem, applicable to all standard numerical methods. Guaranteed and computable upper and lower bounds on an arbitrary simple eigenvalue are given, as well as on the energy error in the approximation of the associated eigenvector. The bounds are valid under the sole condition that the approximate i-th eigenvalue lies between the exact $(i-1)$-th and $(i+1)$-th eigenvalue, where the relative gaps are sufficiently large. We give a practical way how to check this; the precision of the resulting estimates depends on these relative gaps. Our bounds feature no unknown (solution-, regularity-, or polynomial-degree-dependent) constant, are optimally convergent (efficient), and polynomial-degree robust. Under a further explicit, a posteriori, minimal resolution condition, the multiplicative constant in our estimates can be reduced by a fixed factor; moreover, when an elliptic regularity assumption is satisfied with known constants, this multiplicative constant can be brought to the optimal value of 1 with mesh refinement. Applications of our framework to nonconforming, discontinuous Galerkin, and mixed finite element approximations of arbitrary polynomial degree are provided, along with numerical illustrations.

5.1 Introduction

Precise numerical approximation of eigenvalues and eigenvectors of elliptic operators on general domains is crucial in countless applications. In addition to standard conforming Galerkin (variational) approximations, *nonconforming methods* such as nonconforming finite elements, discontinuous Galerkin elements, or mixed finite elements are very popular, and one naturally asks the question of the size of the error in their eigenvalue and eigenvector approximations.

The issue of error control is usually tackled via the a posteriori estimates theory. Recently, powerful estimates were obtained for nonconforming approximations of the Laplace source problem, see Destuynder and Métivet [86], Ainsworth [2, 3], Kim [149, 150], Vohralík [242], Carstensen and Merdon [66], or Ern *et al.* [102, 104, 105], see also the references therein. The Laplace eigenvalue problem seems to be structurally more difficult. Recently, though, *guaranteed* a posteriori estimates on the error in the *i-th eigenvalue* have been obtained in Carstensen and Gedicke [64] and Liu [176]. The theory of [64, 176] applies for arbitrarily coarse meshes and gives convincing numerical results in many test cases. One could, however, comment that these results only seem to apply to the lowest-order nonconforming finite element method, the arguments used are of a priori nature that relies on an interpolation estimate, and an overestimation in presence of singularities may appear as the bounds feature the diameter of the largest mesh element, see [64, Section 6.3]. Armentano and Durán [10], Luo *et al.* [180], Hu *et al.* [140, 141], or Yang *et al.* [253] also derived (guaranteed) eigenvalue estimates for the nonconforming finite element method, where, however, a saturation assumption may be necessary and/or the results are valid only asymptotically. Errors in both eigenvalue and eigenvector approximations in nonconforming methods have also been studied previously, although rather seldom. We refer in particular to Dari *et al.* [82] for nonconforming finite elements, to Giani and Hall [113] for discontinuous Galerkin elements, and to Durán *et al.* [94] and Jia *et al.* [143] for mixed finite elements. Unfortunately, these estimates systematically contain solution-independent but unknown constants as well as solution-dependent, uncomputable terms, claimed higher order on fine enough meshes via a priori arguments.

The purpose of the present paper is to extend our conforming theory of Cancès *et al.* [52] to a *general framework* for guaranteed and optimally convergent *a posteriori bounds* for both arbitrary simple *i-th eigenvalue* and the associated *eigenvector* of the Laplace eigenvalue problem. Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a polygonal/polyhedral domain with a Lipschitz boundary. Let the exact eigenvector and eigenvalue pairs (u_i, λ_i) satisfy

$$-\Delta u_i = \lambda_i u_i \quad \text{in } \Omega, \tag{5.1.1}$$

with the condition $\|u_i\| = 1$, see (5.2.1) below for the precise weak formulation. We denote by $(\cdot,\cdot)$ the $L^2(\Omega)$ or $[L^2(\Omega)]^d$ scalar product over Ω and by $\|\cdot\|$ the associated norm and let

$$H^1(\mathcal{T}_h) := \{v \in L^2(\Omega);\, v|_K \in H^1(K) \quad \forall K \in \mathcal{T}_h\} \tag{5.1.2}$$

be the so-called *broken Sobolev space*, where the traces on mesh faces do not need to coincide. It is defined over a computational mesh $\mathcal{T}_h$ of the domain Ω; details of the setting are given in Section 5.2. On $H^1(\mathcal{T}_h)$, we generalize the usual weak gradient ∇ of $H^1(\Omega)$ to the discrete gradient ∇_θ featuring a parameter $\theta \in \{-1, 0, 1\}$, see (5.2.5) below. We consider here an abstract setting where the approximate eigenvector–eigenvalue pair (u_{ih}, λ_{ih}) is not necessarily linked to any particular numerical method, $u_{ih} \in H^1(\mathcal{T}_h)$ is a piecewise polynomial possibly nonconforming in the sense that $u_{ih} \notin H^1_0(\Omega)$ and not necessarily scaled to $\|u_{ih}\| = 1$, $\lambda_{ih} \in \mathbb{R}^+$, and the relation $\|\nabla_\theta u_{ih}\|^2 = \lambda_{ih}$ typically does not hold at the discrete level in contrast to the continuous one. Concrete examples of numerical methods fitting to our setting can be found later in Section 5.7.

Our main tools are an *equilibrated flux reconstruction* $\boldsymbol{\sigma}_{ih} \in \mathbf{H}(\mathrm{div}, \Omega)$ satisfying $\nabla{\cdot}\boldsymbol{\sigma}_{ih} = \lambda_{ih} u_{ih}$ and an *eigenvector reconstruction* $s_{ih} \in H^1_0(\Omega)$, both defined in Section 5.3. These are piecewise polynomials such that $-\boldsymbol{\sigma}_{ih}$ and ∇s_{ih} are as close as possible to the discrete gradient $\nabla_\theta u_{ih}$. They are constructed over patches of mesh elements following Destuynder and Métivet [87], Braess and Schöberl [33], Carstensen and Merdon [66], and Ern and Vohralík [105], see also the references therein. We employ them in Section 5.4 to show in particular how the dual norm of the residual of the pair (s_{ih}, λ_{ih}) can be bounded in a computable way. Section 5.5 then bounds the $L^2(\Omega)$-norm of the Riesz representation of the residual of (s_{ih}, λ_{ih}) under an assumption of elliptic regularity on the corresponding Laplace source problem. It enables later to give improved computable estimates in the considered nonconforming setting.

Our main results are collected in Section 5.6 and crucially rely on [52, Section 3], where mutual relations between the i-th eigenvalue error, the associated eigenvector energy error, and the dual norm of the residual in terms of an arbitrary pair $(\tilde{s}_{ih}, \lambda_{ih}) \in H^1_0(\Omega) \times \mathbb{R}^+$ such that $\|\tilde{s}_{ih}\| = 1$ are given. Using $\tilde{s}_{ih} := s_{ih}/\|s_{ih}\|$, where s_{ih} is the above eigenvector reconstruction, inequality (5.6.4) of Theorem 5.6.1 in particular gives

$$\|\nabla \tilde{s}_{ih}\|^2 - \eta_i^2 \le \lambda_i, \tag{5.1.3a}$$

where η_i is an a posteriori error estimator with a typical structure

$$\eta_i = m_{ih}(\lambda_{ih}\|u_{ih} - s_{ih}\|/\sqrt{\underline{\lambda}_1} + \|\nabla s_{ih} + \boldsymbol{\sigma}_{ih}\|)/\|s_{ih}\|. \tag{5.1.3b}$$

Thus (5.1.3a) gives a *guaranteed* and computable lower bound for the i-th exact *eigenvalue* λ_i. An upper bound on λ_i is recalled in inequality (5.6.10) of Theorem 5.6.3. Then a *guaranteed* and computable a posteriori estimate on the associated *eigenvector energy error* is given next, see in particular estimate (5.6.16) of Theorem 5.6.4 revealing

$$\|\nabla_\theta(u_i - u_{ih})\| \le \eta_i + \|\nabla_\theta(u_{ih} - \tilde{s}_{ih})\|. \tag{5.1.3c}$$

The eigenvalue and eigenvector error bounds (5.1.3) are *efficient* (optimally convergent) in the sense that

$$\begin{aligned}\eta_i + \|\nabla_\theta(u_{ih} - \tilde{s}_{ih})\| \le C_i(&\|\nabla_\theta(u_i - u_{ih})\| + \text{ consistency terms}\\ &+ \text{norm of mean values of jumps of } u_{ih}),\end{aligned} \tag{5.1.4}$$

see inequality (5.6.17) of Theorem 5.6.4. Here C_i is a generic constant that only depends on λ_1, λ_{ih}, on the lower bound $\underline{\lambda}_{i+1}$ of λ_{i+1}, possibly on the upper bound $\overline{\lambda}_{i-1}$ of λ_{i-1}, on the shape of Ω, and on some (broken) Poincaré–Friedrichs constants over patches of elements and a stability constant of mixed finite elements (both only depending on the shape regularity of the mesh and on the space dimension d). The constant C_i is in particular independent of the polynomial degree of u_{ih}, leading to *polynomial-degree robustness*. The consistency terms above may not be present, typically for nonconforming finite elements. Similarly, the jump mean values of u_{ih} are zero in nonconforming and mixed finite elements, and this term also vanishes in our developments for the symmetric variant of the discontinuous Galerkin finite element method when the parameter θ equals 1.

The above results are valid under the condition (5.6.1), requesting λ_{ih} to lie between computable bounds $\underline{\lambda}_{i+1}$ and $\overline{\lambda}_{i-1}$ (when $i > 1$) on the surrounding exact eigenvalues λ_{i+1} and λ_{i-1}, see the discussion in [52, Remark 5.4] for its practical verification. We also need the residual orthogonality condition of Assumption 5.3.1 in order to reconstruct the equilibrated flux. Then, for u_{ih} a piecewise polynomial of degree p, the reconstructions s_{ih} and $\boldsymbol{\sigma}_{ih}$ are prescribed in discrete spaces of order $p+1$. There is no specific condition on the fineness of the mesh in Case A of Theorems 5.6.1 and 5.6.4, but the multiplicative factor m_{ih} in (5.1.3b)

contains the relative gap of the form $\max\{(\frac{\lambda_{ih}}{\lambda_{i-1}}-1)^{-1},(1-\frac{\lambda_{ih}}{\lambda_{i+1}})^{-1}\}$. Two improvements are possible. Under the computable minimal resolution criterion (5.6.6b) (satisfied for fine enough meshes), m_{ih} is reduced by a fixed factor, see Case B of Theorems 5.6.1 and 5.6.4. If an elliptic regularity assumption on the corresponding source problem is satisfied and if the minimal resolution condition (5.6.8b) holds, the relative gap in m_{ih} is multiplied by a power of the mesh size h, so that m_{ih} can be brought to the optimal value of 1 in the limit of mesh size tending to zero, which we show in Case C of Theorems 5.6.1 and 5.6.4. The efficiency constant C_i from (5.1.4) or (5.6.17) can be fully traced from the detailed estimates of Theorem 5.6.4; it in particular deteriorates for increasing eigenvalues.

The application of our abstract results to a given numerical method merely requires the verification of the setting and of Assumption 5.3.1. We undertake this in Section 5.7 for nonconforming finite elements, discontinuous Galerkin elements, and mixed finite elements of arbitrary polynomial degree. For mixed finite elements, elementwise postprocessings of the approximate eigenvector and of its flux need to be performed first. Numerical experiments are presented in Section 5.8 for the nonconforming and discontinuous Galerkin methods and fully support our theoretical findings for a couple of model problems. In particular, the a posteriori applicability conditions (5.6.6b) and (5.6.8b) for cases B and C are satisfied here already on very coarse meshes. Some concluding remarks and an outlook are given in Section 5.9. In particular, inexact algebraic eigenvalue solvers promoted in Mehrmann and Miedlar [189] or Carstensen and Gedicke [64] can be treated as in [52] and the references therein, allowing to generalize the present estimates to an arbitrary iterative solver step where Assumption 5.3.1 typically does not hold.

We finish our paper by two extensions. We first show in Appendix 5.10.1 that the key relations between the i-th eigenvalue error, the associated eigenvector energy error, and the dual norm of the residual, when u_{ih} is conforming, are in fact not restricted to the Laplace operator; we extend them to the generic class of bounded-below self-adjoint operators with compact resolvent in Theorems 5.10.1 and 5.10.2. Appendix 5.10.2 then gives a further possible improvement of the first eigenvalue upper bound: from $\lambda_1 \le \|\nabla \tilde{s}_{1h}\|^2$ of (5.6.11) in Theorem 5.6.3 to $\lambda_1 \le \|\nabla \tilde{s}_{1h}\|^2 - \tilde{\eta}_1$ of Theorem 5.10.5.

5.2 Setting

Let $H^1(\Omega)$ be the Sobolev space of $L^2(\Omega)$ functions with weak gradients ∇ in $[L^2(\Omega)]^d$. We denote henceforth by $V := H^1_0(\Omega)$ its zero-trace subspace. Later, we will also employ the space $\mathbf{H}(\mathrm{div},\Omega)$ of $[L^2(\Omega)]^d$ functions with weak divergences $\nabla\cdot$ in $L^2(\Omega)$.

5.2.1 The Laplace eigenvalue problem

The weak formulation of (5.1.1) looks for $(u_i,\lambda_i)\in V\times\mathbb{R}^+$ with $\|u_i\|=1$ and

$$(\nabla u_i,\nabla v)=\lambda_i(u_i,v)\qquad \forall v\in V. \tag{5.2.1}$$

It is well-known (see, e.g., Babuška and Osborn [14] or Boffi [28] and the references therein) that u_i, $i\ge 1$, form a countable orthonormal basis of $L^2(\Omega)$ consisting of eigenvectors from V; we assume that the sequence of eigenvalues is such that $0<\lambda_1<\lambda_2\le\ldots\le\lambda_i\to\infty$. We will actually suppose that the eigenvalue λ_i that we study is simple, which is always the case for $i=1$. The associated eigenvector u_i is then uniquely defined upon fixing its sign by the condition $(u_i,\chi_i)>0$, where $\chi_i\in L^2(\Omega)$ is typically a characteristic function of Ω (for $i=1$) or of its subdomain (for $i>1$). The setting for $i=1$ in particular implies the Poincaré–Friedrichs

inequality

$$\|v\|^2 \le \frac{1}{\lambda_1}\|\nabla v\|^2 \qquad \forall v \in V. \tag{5.2.2}$$

5.2.2 Meshes and generic piecewise polynomial spaces

We denote by $\mathcal{T}_h$ a matching simplicial mesh in the sense of Ciarlet [76], shape-regular with a parameter $\kappa_{\mathcal{T}} > 0$: the ratio of the element diameter h_K and of the diameter of the inscribed ball to $K \in \mathcal{T}_h$ is bounded by $\kappa_{\mathcal{T}}$ (uniformly for a sequence of meshes). Denote also by h the maximal element diameter over all $K \in \mathcal{T}_h$. The mesh $(d-1)$-dimensional faces are collected in the set $\mathcal{E}_h$, with interior faces $\mathcal{E}_h^{\rm int}$ and boundary faces $\mathcal{E}_h^{\rm ext}$. A generic face is denoted by e, its diameter by h_e, and its unit normal vector (the direction is arbitrary but fixed) by $\mathbf{n}_e$. We will often employ the jump operator $[\![\cdot]\!]$ yielding the difference of the traces of the argument from the two mesh elements that share $e \in \mathcal{E}_h^{\rm int}$ along $\mathbf{n}_e$ and the actual trace for $e \in \mathcal{E}_h^{\rm ext}$. Similarly, the average operator $\{\!\{\cdot\}\!\}$ yields the mean value of the traces from adjacent mesh elements on interior faces and the actual trace on boundary faces. The set of vertices will be denoted by $\mathcal{V}_h$; it is composed of interior vertices $\mathcal{V}_h^{\rm int}$ and vertices located on the boundary $\mathcal{V}_h^{\rm ext}$, with a generic vertex denoted by $\mathbf{a}$.

Let $\mathbb{P}_q(K)$ stand for polynomials of total degree at most $q \ge 0$ on $K \in \mathcal{T}_h$, and $\mathbb{P}_q(\mathcal{T}_h) \subset H^1(\mathcal{T}_h)$ for piecewise q-th order polynomials on $\mathcal{T}_h$. For a given $q \ge 1$, we denote by $V_h^q := \mathbb{P}_q(\mathcal{T}_h) \cap V$ the q-th order conforming finite element space. Similarly, for $q \ge 0$, $\mathbf{V}_h^q := \{\mathbf{v}_h \in \mathbf{H}(\mathrm{div},\Omega); \mathbf{v}_h|_K \in [\mathbb{P}_q(K)]^d + \mathbb{P}_q(K)\mathbf{x}\}$ and $Q_h^q := \mathbb{P}_q(\mathcal{T}_h)$ stand for the Raviart–Thomas–Nédélec mixed finite element spaces of order q, cf. Brezzi and Fortin [36] or Roberts and Thomas [214]. We will also use the lowest-order broken space $\mathbf{V}^0(\mathcal{T}_h) := \{\mathbf{v}_h \in [L^2(\Omega)]^d;\ \mathbf{v}_h|_K \in [\mathbb{P}_0(K)]^d + \mathbb{P}_0(K)\mathbf{x}\}$, where in contrast to $\mathbf{V}_h^q$, no normal trace continuity is imposed via the inclusion in $\mathbf{H}(\mathrm{div},\Omega)$.

5.2.3 Broken and discrete gradients

On the broken Sobolev space $H^1(\mathcal{T}_h)$ defined in (5.1.2), the usual weak gradient ∇ is not defined. We will in this paper use two successive generalizations of the notion of the weak gradient. We will first denote by $\nabla_h v \in [L^2(\Omega)]^d$ the *broken gradient* of $v \in H^1(\mathcal{T}_h)$ given by

$$(\nabla_h v)|_K := \nabla(v|_K) \qquad \forall K \in \mathcal{T}_h. \tag{5.2.3}$$

We will need the following generalization of (5.2.2), the so-called broken Poincaré–Friedrichs inequality, see Brenner [34, Remark 1.1] or Vohralík [241, Theorem 5.4] and the references therein:

$$\|v\| \le C_{\rm bF}\left(\|\nabla_h v\|^2 + \sum_{e\in\mathcal{E}_h^{\rm int}} h_e^{-1}\|\Pi_e^0[\![v]\!]\|_e^2 + \langle v,1\rangle_{\partial\Omega}^2\right)^{\frac12} \qquad \forall v \in H^1(\mathcal{T}_h), \tag{5.2.4}$$

where Π_e^0 stands for the $L^2(e)$-orthogonal projection onto constants on the face e, $\langle\cdot,\cdot\rangle$ denotes the $L^2(\Omega)$ scalar product over $\partial\Omega$, and the constant $C_{\rm bF}$ only depends on the domain Ω, the space dimension d, and the mesh shape regularity parameter $\kappa_{\mathcal{T}}$.

In order to prove the elliptic regularity bound of Theorem 5.5.4 below in a very general setting, we are lead to a further generalization. It is motivated by the lifting operators used in the discontinuous Galerkin finite element method, see Di Pietro and Ern [88, Section 4.3] and the references therein, but we crucially rely here on the space $\mathbf{V}^0(\mathcal{T}_h)$ of the lowest-order broken Raviart–Thomas–Nédélec polynomials defined above. Let $v \in H^1(\mathcal{T}_h)$. For each face $e \in \mathcal{E}_h$, we define the lifting operator $\mathfrak{l}_e : L^2(e) \to \mathbf{V}^0(\mathcal{T}_e)$, where $\mathcal{T}_e$ regroups the mesh elements sharing

the face e and $\mathbf{V}^0(\mathcal{T}_e)$ is the restriction of $\mathbf{V}^0(\mathcal{T}_h)$ thereon. The lifting $\mathfrak{l}_e(\llbracket v \rrbracket)$ is prescribed by $(\mathfrak{l}_e(\llbracket v \rrbracket), \mathbf{v}_h)_{\mathcal{T}_e} = \langle \{\!\{\mathbf{v}_h\}\!\}\cdot\mathbf{n}_e, \llbracket v \rrbracket\rangle_e$ for all $\mathbf{v}_h \in \mathbf{V}^0(\mathcal{T}_e)$. We then extend $\mathfrak{l}_e(\llbracket v \rrbracket)$ by zero outside of $\mathcal{T}_e$ to form an element of $\mathbf{V}^0(\mathcal{T}_h)$. For a parameter $\theta \in \{-1, 0, 1\}$, the *discrete gradient* $\nabla_\theta v \in [L^2(\Omega)]^d$ is then given by

$$\nabla_\theta v := \nabla_h v - \theta \sum_{e\in\mathcal{E}_h} \mathfrak{l}_e(\llbracket v \rrbracket). \tag{5.2.5}$$

We observe that $\nabla_\theta v = \nabla_h v$ when $\theta = 0$ or when the jumps of v are of mean value 0, i.e., $\langle \llbracket v \rrbracket, 1\rangle_e = 0$ for all $e \in \mathcal{E}_h$; indeed, this follows from the fact that $\mathbf{v}_h\cdot\mathbf{n}_e$ are constants for $\mathbf{v}_h \in \mathbf{V}^0(\mathcal{T}_h)$. Consistently,

$$\nabla_\theta v = \nabla_h v = \nabla v \qquad \forall v \in H^1_0(\Omega). \tag{5.2.6}$$

5.2.4 Residual and its dual norm

The derivation of a posteriori error estimates usually exploits the concept of the *residual* and of its *dual norm*. We will proceed in this way as well. Throughout the paper, it will reveal convenient to employ the residual of different pairs $(w_i, \lambda_{ih}) \in H^1(\mathcal{T}_h) \times \mathbb{R}$, with for w_i the approximate solution u_{ih}, the eigenvector reconstruction s_{ih} of Definition 5.3.3 below, or a generic function in V. Let V' stand for the dual of V.

Definition 5.2.1 (Residual and its dual norm)**.** *For any pair $(w_i, \lambda_{ih}) \in H^1(\mathcal{T}_h) \times \mathbb{R}$, define the residual $\mathrm{Res}_\theta(w_i, \lambda_{ih}) \in V'$ by*

$$\langle \mathrm{Res}_\theta(w_i, \lambda_{ih}), v\rangle_{V',V} := \lambda_{ih}(w_i, v) - (\nabla_\theta w_i, \nabla v) \qquad \forall v \in V. \tag{5.2.7a}$$

Its dual norm is then

$$\|\mathrm{Res}_\theta(w_i, \lambda_{ih})\|_{-1} := \sup_{v\in V,\, \|\nabla v\|=1} \langle \mathrm{Res}_\theta(w_i, \lambda_{ih}), v\rangle_{V',V}. \tag{5.2.7b}$$

We will also often work with the *Riesz representation* of the residual $\mathfrak{r}_{w_i} \in V$, given by

$$(\nabla \mathfrak{r}_{w_i}, \nabla v) = \langle \mathrm{Res}_\theta(w_i, \lambda_{ih}), v\rangle_{V',V} \qquad \forall v \in V. \tag{5.2.8a}$$

Then

$$\|\nabla \mathfrak{r}_{w_i}\| = \|\mathrm{Res}_\theta(w_i, \lambda_{ih})\|_{-1}. \tag{5.2.8b}$$

5.3 Eigenvector and equilibrated flux reconstructions

We introduce in this section two key reconstructions, following [207, 160, 86, 2, 149, 150, 242, 102, 3, 33, 104, 105] and the references therein. To motivate, note that from (5.2.1), it is straightforward that $-\nabla u_i \in \mathbf{H}(\mathrm{div}, \Omega)$, with the weak divergence equal to $\lambda_i u_i$. On the discrete level, however, $-\nabla_\theta u_{ih} \notin \mathbf{H}(\mathrm{div}, \Omega)$ in general, and, a fortiori, $\nabla\cdot(-\nabla_\theta u_{ih}) \neq \lambda_{ih} u_{ih}$. We will thus introduce an *equilibrated flux reconstruction*, a vector-valued function $\boldsymbol{\sigma}_{ih}$ constructed from the given pair (u_{ih}, λ_{ih}), satisfying

$$\boldsymbol{\sigma}_{ih} \in \mathbf{H}(\mathrm{div}, \Omega), \tag{5.3.1a}$$

$$\nabla\cdot\boldsymbol{\sigma}_{ih} = \lambda_{ih} u_{ih}. \tag{5.3.1b}$$

Similarly, as we treat here cases where $u_{ih} \notin V$, possibly jumping between the mesh elements, we will employ an *eigenvector reconstruction*, a scalar-valued function s_{ih} constructed from u_{ih} and satisfying

$$s_{ih} \in V. \tag{5.3.2}$$

Actually, both $\boldsymbol{\sigma}_{ih}$ and s_{ih} will be piecewise polynomials defined in standard finite element subspaces of $\mathbf{H}(\mathrm{div}, \Omega)$ and V, respectively.

5.3.1 Orthogonality of the residual

Let $\psi_{\mathbf{a}}$ for $\mathbf{a} \in \mathcal{V}_h$ stand for the piecewise affine function taking value 1 at the vertex $\mathbf{a}$ and zero at the other vertices. Remarkably, these functions form a partition of unity via $\sum_{\mathbf{a}\in\mathcal{V}_h} \psi_{\mathbf{a}} = 1|_\Omega$. Denote by $\mathcal{T}_{\mathbf{a}}$ the patch of elements of $\mathcal{T}_h$ which share the vertex $\mathbf{a} \in \mathcal{V}_h$ and by $\omega_{\mathbf{a}}$ the corresponding subdomain of Ω. Our key assumption will be:

Assumption 5.3.1 (Orthogonality of (u_{ih}, λ_{ih}) residual to $\psi_{\mathbf{a}}$)**.** *There holds*

$$\lambda_{ih}(u_{ih}, \psi_{\mathbf{a}})_{\omega_{\mathbf{a}}} - (\nabla_\theta u_{ih}, \nabla\psi_{\mathbf{a}})_{\omega_{\mathbf{a}}} = \langle \mathrm{Res}_\theta(u_{ih}, \lambda_{ih}), \psi_{\mathbf{a}}\rangle_{V',V} = 0 \qquad \forall \mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}.$$

Assumption 5.3.1 can typically be verified for an exact algebraic solver; Section 5.7 below shows how to check it for some standard numerical methods. Inexact solvers, where Assumption 5.3.1 does not hold, can be treated as in [52] and the references therein.

5.3.2 Reconstruction spaces

In practice, the approximate eigenvector u_{ih} is a piecewise polynomial, $u_{ih} \in \mathbb{P}_p(\mathcal{T}_h)$ for some $p \geq 1$. To define the reconstructions in this setting, we will, for each vertex $\mathbf{a} \in \mathcal{V}_h$, work with restrictions $\mathbf{V}_h^{p+1}(\omega_{\mathbf{a}})$ and $Q_h^{p+1}(\omega_{\mathbf{a}})$ of the spaces from Section 5.2.2 to the patch subdomain $\omega_{\mathbf{a}}$; conversely, we will often tacitly extend functions defined on $\omega_{\mathbf{a}}$ by zero outside of $\omega_{\mathbf{a}}$. With $\mathbf{n}_{\omega_{\mathbf{a}}}$ standing for the outward unit normal of $\omega_{\mathbf{a}}$, we define

$$\begin{aligned}
&\mathbf{V}_h^{\mathbf{a}} := \{\mathbf{v}_h \in \mathbf{V}_h^{p+1}(\omega_{\mathbf{a}});\ \mathbf{v}_h{\cdot}\mathbf{n}_{\omega_{\mathbf{a}}} = 0 \text{ on } \partial\omega_{\mathbf{a}}\},\\
&Q_h^{\mathbf{a}} := \{q_h \in Q_h^{p+1}(\omega_{\mathbf{a}});\ (q_h, 1)_{\omega_{\mathbf{a}}} = 0\},
&&\mathbf{a} \in \mathcal{V}_h^{\mathrm{int}},\\
&\mathbf{V}_h^{\mathbf{a}} := \{\mathbf{v}_h \in \mathbf{V}_h^{p+1}(\omega_{\mathbf{a}});\ \mathbf{v}_h{\cdot}\mathbf{n}_{\omega_{\mathbf{a}}} = 0 \text{ on } \partial\omega_{\mathbf{a}} \setminus \partial\Omega\},\\
&Q_h^{\mathbf{a}} := Q_h^{p+1}(\omega_{\mathbf{a}}),
&&\mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}},\\
&W_h^{\mathbf{a}} := \mathbb{P}_{p+1}(\mathcal{T}_{\mathbf{a}}) \cap H_0^1(\omega_{\mathbf{a}})
&&\mathbf{a} \in \mathcal{V}_h.
\end{aligned}$$

5.3.3 Equilibrated flux reconstruction

We construct $\boldsymbol{\sigma}_{ih}$ satisfying (5.3.1) by *local constrained minimizations*:

Definition 5.3.2 (Equilibrated flux reconstruction)**.** *Let $(u_{ih}, \lambda_{ih}) \in \mathbb{P}_p(\mathcal{T}_h) \times\mathbb{R}$ be arbitrary but satisfying Assumption 5.3.1. Prescribe $\boldsymbol{\sigma}_{ih}^{\mathbf{a}} \in \mathbf{V}_h^{\mathbf{a}}$ by solving*

$$\boldsymbol{\sigma}_{ih}^{\mathbf{a}} := \arg \min_{\mathbf{v}_h\in\mathbf{V}_h^{\mathbf{a}},\, \nabla\cdot\mathbf{v}_h = \Pi_{Q_h^{\mathbf{a}}}(\psi_{\mathbf{a}}\lambda_{ih}u_{ih} - \nabla_\theta u_{ih}\cdot\nabla\psi_{\mathbf{a}})} \|\psi_{\mathbf{a}}\nabla_\theta u_{ih} + \mathbf{v}_h\|_{\omega_{\mathbf{a}}} \qquad \forall \mathbf{a} \in \mathcal{V}_h \tag{5.3.3a}$$

and define

$$\boldsymbol{\sigma}_{ih} := \sum_{\mathbf{a}\in\mathcal{V}_h} \boldsymbol{\sigma}_{ih}^{\mathbf{a}}. \tag{5.3.3b}$$

In (5.3.3a), $\Pi_{Q_h^{\mathbf{a}}}$ stands for the $L^2(\Omega)$-orthogonal projection onto the local space $Q_h^{\mathbf{a}}$. It is actually only needed for the simplification of Remark 5.6.10 below; otherwise, $\psi_{\mathbf{a}}\lambda_{ih}u_{ih} - \nabla_\theta u_{ih}\cdot\nabla\psi_{\mathbf{a}}$ is a piecewise polynomial of degree $p+1$ on the patch $\mathcal{T}_{\mathbf{a}}$, with mean value zero thanks to Assumption 5.3.1, so that $\Pi_{Q_h^{\mathbf{a}}}(\psi_{\mathbf{a}}\lambda_{ih}u_{ih} - \nabla_\theta u_{ih}\cdot\nabla\psi_{\mathbf{a}}) = \psi_{\mathbf{a}}\lambda_{ih}u_{ih} - \nabla_\theta u_{ih}\cdot\nabla\psi_{\mathbf{a}}$. Imposing the divergence of $\boldsymbol{\sigma}_{ih}^{\mathbf{a}}$ in this way and defining $\boldsymbol{\sigma}_{ih}$ via (5.3.3b) leads to (5.3.1b), see, e.g., [105, Lemma 3.5].

It is easy to verify that problems (5.3.3a) are equivalent (see [105, Remark 3.7]) to the mixed finite element approximation to the homogeneous Neumann (Neumann–Dirichlet close

to the boundary) problem posed on the patch $\mathcal{T}_\mathbf{a}$: find $\boldsymbol{\sigma}_{ih}^\mathbf{a} \in \mathbf{V}_h^\mathbf{a}$ and $p_h^\mathbf{a} \in Q_h^\mathbf{a}$ such that

$$\begin{aligned}(\boldsymbol{\sigma}_{ih}^\mathbf{a}, \mathbf{v}_h)_{\omega_\mathbf{a}} - (p_h^\mathbf{a}, \nabla\cdot\mathbf{v}_h)_{\omega_\mathbf{a}} &= -(\psi_\mathbf{a}\nabla_\theta u_{ih}, \mathbf{v}_h)_{\omega_\mathbf{a}} && \forall \mathbf{v}_h \in \mathbf{V}_h^\mathbf{a},\\ (\nabla\cdot\boldsymbol{\sigma}_{ih}^\mathbf{a}, q_h)_{\omega_\mathbf{a}} &= (\psi_\mathbf{a}\lambda_{ih}u_{ih} - \nabla_\theta u_{ih}\cdot\nabla\psi_\mathbf{a}, q_h)_{\omega_\mathbf{a}} && \forall q_h \in Q_h^\mathbf{a}.\end{aligned}$$

It follows from the standard references [36, 214] that the discrete inf–sup condition for the pair $\mathbf{V}_h^\mathbf{a} \times Q_h^\mathbf{a}$ is satisfied.

5.3.4 Eigenvector reconstruction

For nonconforming eigenvectors u_{ih}, i.e., u_{ih} is a piecewise polynomial not included in $V = H_0^1(\Omega)$ but merely in $H^1(\mathcal{T}_h)$, the eigenvector reconstruction complying with requirement (5.3.2) is obtained via *local unconstrained minimizations* employing the broken gradient (5.2.3):

Definition 5.3.3 (Eigenvector reconstruction)**.** *Let $u_{ih} \in \mathbb{P}_p(\mathcal{T}_h)$ be arbitrary. Prescribe $s_{ih}^\mathbf{a} \in W_h^\mathbf{a}$ by solving*

$$s_{ih}^\mathbf{a} := \arg\min_{v_h \in W_h^\mathbf{a}} \|\nabla_h(\psi_\mathbf{a} u_{ih} - v_h)\|_{\omega_\mathbf{a}} \qquad \forall \mathbf{a} \in \mathcal{V}_h \tag{5.3.4}$$

and define

$$s_{ih} := \sum_{\mathbf{a}\in\mathcal{V}_h} s_{ih}^\mathbf{a}.$$

Immediately, problems (5.3.4) are equivalently described by their Euler–Lagrange conditions; these request to find the conforming finite element approximation $s_{ih}^\mathbf{a} \in W_h^\mathbf{a}$ to the homogeneous Dirichlet problem posed over the patch $\mathcal{T}_\mathbf{a}$ such that

$$(\nabla s_{ih}^\mathbf{a}, \nabla v_h)_{\omega_\mathbf{a}} = (\nabla_h(\psi_\mathbf{a} u_{ih}), \nabla v_h)_{\omega_\mathbf{a}} \qquad \forall v_h \in W_h^\mathbf{a}.$$

5.4 Dual norm of the residual and nonconformity bounds

We summarize here bounds on the dual norm of the residual and on nonconformity that are available from the context of source problems. They will be crucial later in Section 5.6.

5.4.1 Some additional notation and useful inequalities

We first need to introduce some more background. Consider a vertex $\mathbf{a} \in \mathcal{V}_h$ and on the patch domain $\omega_\mathbf{a}$ define

$$H_*^1(\omega_\mathbf{a}) := \{v \in H^1(\omega_\mathbf{a});\ (v,1)_{\omega_\mathbf{a}} = 0\}, \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}, \tag{5.4.1a}$$

$$H_*^1(\omega_\mathbf{a}) := \{v \in H^1(\omega_\mathbf{a});\ v = 0 \text{ on } \partial\omega_\mathbf{a} \cap \partial\Omega\}, \quad \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}. \tag{5.4.1b}$$

Then the *Poincaré–Friedrichs* inequality, corresponding to (5.2.2) on the patches $\omega_\mathbf{a}$, states

$$\|v\|_{\omega_\mathbf{a}} \le C_{\mathrm{PF},\omega_\mathbf{a}} h_{\omega_\mathbf{a}} \|\nabla v\|_{\omega_\mathbf{a}} \qquad \forall v \in H_*^1(\omega_\mathbf{a}), \tag{5.4.2a}$$

where $C_{\mathrm{PF},\omega_\mathbf{a}}$ depends only on the mesh regularity parameter $\kappa_\mathcal{T}$ and the space dimension d. Similarly, when the functions are piecewise H^1 only, we will use the inequality

$$\|v\|_{\omega_\mathbf{a}} \le C_{\mathrm{bPF},\omega_\mathbf{a}} h_{\omega_\mathbf{a}} \left(\|\nabla_h v\|_{\omega_\mathbf{a}}^2 + \sum_{e\in\mathcal{E}_h,\, \mathbf{a}\in e} h_e^{-1} \|\Pi_e^0 [\![v]\!]\|_e^2 \right)^{\frac{1}{2}} \tag{5.4.2b}$$

valid for all $v \in H^1(\mathcal{T}_h)$ such that $(v,1)_{\omega_\mathbf{a}} = 0$ when $\mathbf{a} \in \mathcal{V}_h^{\rm int}$, and where the constant $C_{\rm bPF,\omega_\mathbf{a}}$ depends only on $\kappa_\mathcal{T}$ and d. Inequality (5.4.2b) may be seen as a local version of (5.2.4) on the patch domain $\omega_\mathbf{a}$, with the mean value condition $(v,1)_{\omega_\mathbf{a}} = 0$ for $\mathbf{a} \in \mathcal{V}_h^{\rm int}$ or appearance of boundary faces $e \in \mathcal{E}_h \subset \partial\omega_\mathbf{a}$ for $\mathbf{a} \in \mathcal{V}_h^{\rm ext}$ replacing the boundary term $\langle v,1\rangle_{\partial\Omega}$ from (5.2.4). Define $C_{\rm cont,PF} := \max_{\mathbf{a}\in\mathcal{V}_h}\{1 + C_{\rm PF,\omega_\mathbf{a}} h_{\omega_\mathbf{a}} \|\nabla\psi_\mathbf{a}\|_{\infty,\omega_\mathbf{a}}\}$ and $C_{\rm cont,bPF} := \max_{\mathbf{a}\in\mathcal{V}_h}\{1 + C_{\rm bPF,\omega_\mathbf{a}} h_{\omega_\mathbf{a}} \|\nabla\psi_\mathbf{a}\|_{\infty,\omega_\mathbf{a}}\}$. The constants $C_{\rm cont,PF}$ and $C_{\rm cont,bPF}$ only depend on the mesh regularity parameter $\kappa_\mathcal{T}$ and the space dimension d and can be fully estimated from above, see the discussion in [105, proofs of Lemmas 3.12 and 3.13 and Section 4.3.2]. In particular, there holds, see Carstensen and Funken [61, Theorem 3.1] or Braess *et al.* [32, Section 3]

$$\|\nabla(\psi_\mathbf{a} v)\|_{\omega_\mathbf{a}} \le C_{\rm cont,PF} \|\nabla v\|_{\omega_\mathbf{a}} \qquad \forall v \in H^1_*(\omega_\mathbf{a}),\ \forall \mathbf{a} \in \mathcal{V}_h. \tag{5.4.3}$$

5.4.2 Stability of the equilibrated flux and eigenvector reconstructions

Recently, Costabel and McIntosh [79, Corollary 3.4], Demkowicz *et al.* [84, Theorem 7.1], and Demkowicz *et al.* [83, Theorem 6.1] have shown fundamental results on the right inverse of respectively the divergence, the normal trace, and the trace operators for polynomial data on a single (reference) tetrahedron. Therefrom, the two following key *stability* results for the constructions of Definitions 5.3.2 and 5.3.3 follow:

$$\|\psi_\mathbf{a}\nabla_\theta u_{ih} + \boldsymbol{\sigma}_{ih}^\mathbf{a}\|_{\omega_\mathbf{a}} \le C_{\rm st} \sup_{v\in H^1_*(\omega_\mathbf{a});\,\|\nabla v\|_{\omega_\mathbf{a}}=1} \langle {\rm Res}_\theta(u_{ih},\lambda_{ih}), \psi_\mathbf{a} v\rangle_{V',V}, \tag{5.4.4a}$$

$$\|\nabla_h(\psi_\mathbf{a} u_{ih} - s_{ih}^\mathbf{a})\|_{\omega_\mathbf{a}} \le C_{\rm st} \inf_{v\in H^1_0(\omega_\mathbf{a})} \|\nabla_h(\psi_\mathbf{a} u_{ih} - v)\|_{\omega_\mathbf{a}}, \tag{5.4.4b}$$

where the constant $C_{\rm st} > 0$ only depends on the mesh shape regularity parameter $\kappa_\mathcal{T}$ and the space dimension d. Indeed, (5.4.4a) has been shown in Braess *et al.* [32, Theorem 7] in two space dimensions and [106, Corollaries 3.3 and 3.6] in three space dimensions, whereas (5.4.4b) is proven in Ern and Vohralík [105, Corollary 3.16] in two space dimensions and [106, Corollary 3.1] in three space dimensions. In [106, Corollaries 3.3 and 3.6], we merely need to set $\boldsymbol{\tau}_p = \psi_\mathbf{a}\nabla_\theta u_{ih}$, $r_K = \psi_\mathbf{a}(\lambda_{ih}u_{ih} + \nabla\cdot(\nabla_\theta u_{ih}))|_K$ for any simplex K in the patch $\mathcal{T}_\mathbf{a}$, and $r_F = \psi_\mathbf{a}[\![\nabla_\theta u_{ih}]\!]\cdot\mathbf{n}_F$ for any face F in the patch $\mathcal{T}_\mathbf{a}$ to infer (5.4.4a) for interior vertices. Similarly, to see (5.4.4b), it is enough to take $\tau_p = \psi_\mathbf{a} u_{ih}$ and $r_F = \psi_\mathbf{a}[\![u_{ih}]\!]_F$ in the notation of [106, Corollary 3.1]. We also remark that computable upper bounds on $C_{\rm st}$ are discussed in [105, Lemma 3.23].

5.4.3 Dual norm of the residual and nonconformity bounds

Our a posteriori error estimates below will rely on the two following intermediate results:

Corollary 5.4.1 (Upper and lower bounds on the dual norm of the residual)**.** *Let $(u_{ih},\lambda_{ih}) \in \mathbb{P}_p(\mathcal{T}_h)\times\mathbb{R}^+$ satisfy Assumption 5.3.1 and let $\boldsymbol{\sigma}_{ih}, s_{ih}$ be respectively constructed following Definitions 5.3.2 and 5.3.3. Then*

$$\|{\rm Res}_\theta(s_{ih},\lambda_{ih})\|_{-1} \le \left(\frac{\lambda_{ih}}{\sqrt{\lambda_1}}\|u_{ih}-s_{ih}\| + \|\nabla s_{ih} + \boldsymbol{\sigma}_{ih}\|\right), \tag{5.4.5a}$$

$$\|\nabla_\theta u_{ih} + \boldsymbol{\sigma}_{ih}\| \le (d+1) C_{\rm st} C_{\rm cont,PF} \|{\rm Res}_\theta(u_{ih},\lambda_{ih})\|_{-1}. \tag{5.4.5b}$$

Proof. Let $v \in V$ with $\|\nabla v\| = 1$ be fixed. Using the definition of the residual (5.2.7a), the consistency of the definition of the discrete gradient (5.2.6), adding and subtracting $(\boldsymbol{\sigma}_{ih}, \nabla v)$,

and employing the Green theorem and the equilibrium property (5.3.1b),

$$\begin{aligned}\langle \mathrm{Res}_\theta(s_{ih},\lambda_{ih}),v\rangle_{V',V} &= \lambda_{ih}(s_{ih},v)-(\nabla s_{ih},\nabla v)\\ &= (\lambda_{ih}s_{ih}-\nabla\cdot\boldsymbol{\sigma}_{ih},v)-(\nabla s_{ih}+\boldsymbol{\sigma}_{ih},\nabla v)\\ &= \lambda_{ih}(s_{ih}-u_{ih},v)-(\nabla s_{ih}+\boldsymbol{\sigma}_{ih},\nabla v).\end{aligned}$$

Thus, the characterization (5.2.7b) of the dual norm of the residual, the Cauchy–Schwarz inequality, and the Poincaré–Friedrichs inequality (5.2.2) yield (5.4.5a). To show (5.4.5b), let us first note that

$$\begin{aligned}&\sup_{v\in H^1_*(\omega_{\mathbf{a}});\,\|\nabla v\|_{\omega_{\mathbf{a}}}=1}\langle \mathrm{Res}_\theta(u_{ih},\lambda_{ih}),\psi_{\mathbf{a}}v\rangle_{V',V}\\ &\le \|\mathrm{Res}_\theta(u_{ih},\lambda_{ih})\|_{-1,\omega_{\mathbf{a}}}\sup_{v\in H^1_*(\omega_{\mathbf{a}});\,\|\nabla v\|_{\omega_{\mathbf{a}}}=1}\|\nabla(\psi_{\mathbf{a}}v)\|_{\omega_{\mathbf{a}}}\\ &\le \|\mathrm{Res}_\theta(u_{ih},\lambda_{ih})\|_{-1,\omega_{\mathbf{a}}}C_{\mathrm{cont,PF}},\end{aligned}\tag{5.4.6}$$

where $\|\mathrm{Res}_\theta(u_{ih},\lambda_{ih})\|_{-1,\omega_{\mathbf{a}}}:=\sup_{v\in H^1_0(\omega_{\mathbf{a}});\,\|\nabla v\|_{\omega_{\mathbf{a}}}=1}\langle \mathrm{Res}_\theta(u_{ih},\lambda_{ih}),v\rangle_{V',V}$, using that for any $v\in H^1_*(\omega_{\mathbf{a}})$, $\psi_{\mathbf{a}}v\in H^1_0(\omega_{\mathbf{a}})$ and (5.4.3). Since $(\nabla_\theta u_{ih}+\boldsymbol{\sigma}_{ih})|_K=\sum_{\mathbf{a}\in\mathcal{V}_K}(\psi_{\mathbf{a}}\nabla_\theta u_{ih}+\boldsymbol{\sigma}_{ih}^{\mathbf{a}})|_K$ for any $K\in\mathcal{T}_h$, where $\mathcal{V}_K$ stands for the set of the vertices of the element K, and since any simplex has $d+1$ vertices,

$$\begin{aligned}\|\nabla_\theta u_{ih}+\boldsymbol{\sigma}_{ih}\|^2 &= \sum_{K\in\mathcal{T}_h}\Big\|\sum_{\mathbf{a}\in\mathcal{V}_K}(\psi_{\mathbf{a}}\nabla_\theta u_{ih}+\boldsymbol{\sigma}_{ih}^{\mathbf{a}})\Big\|_K^2\\ &\le (d+1)\sum_{K\in\mathcal{T}_h}\sum_{\mathbf{a}\in\mathcal{V}_K}\|\psi_{\mathbf{a}}\nabla_\theta u_{ih}+\boldsymbol{\sigma}_{ih}^{\mathbf{a}}\|_K^2\\ &= (d+1)\sum_{\mathbf{a}\in\mathcal{V}_h}\|\psi_{\mathbf{a}}\nabla_\theta u_{ih}+\boldsymbol{\sigma}_{ih}^{\mathbf{a}}\|_{\omega_{\mathbf{a}}}^2.\end{aligned}$$

Now relying on (5.4.4a) and (5.4.6), we infer

$$\|\nabla_\theta u_{ih}+\boldsymbol{\sigma}_{ih}\|^2\le (d+1)C_{\mathrm{st}}^2C_{\mathrm{cont,PF}}^2\sum_{\mathbf{a}\in\mathcal{V}_h}\|\mathrm{Res}_\theta(u_{ih},\lambda_{ih})\|_{-1,\omega_{\mathbf{a}}}^2.$$

Finally, an estimate for combination of negative norms on recovering subdomains, see, for example, [77, Theorem 5.1] and the references therein, implies (5.4.5b). □ □

Corollary 5.4.2 (Nonconformity lower bound)**.** *For $(u_{ih},\lambda_{ih})\in\mathbb{P}_p(\mathcal{T}_h)\times\mathbb{R}^+$, let s_{ih} be constructed following Definition 5.3.3. Then*

$$\begin{aligned}\|\nabla_h(u_{ih}-s_{ih})\|\le\Bigg(&2(d+1)^2C_{\mathrm{st}}^2C_{\mathrm{cont,bPF}}^2\|\nabla_h(u_i-u_{ih})\|^2\\ &+2d(d+1)C_{\mathrm{st}}^2C_{\mathrm{cont,bPF}}^2\sum_{e\in\mathcal{E}_h}h_e^{-1}\|\Pi_e^0[\![u_{ih}]\!]\|_e^2\Bigg)^{\frac12}.\end{aligned}\tag{5.4.7}$$

Proof. This result can be shown as in [105, Lemma 3.22 and Section 4.3.2], relying on (5.4.2b) and crucially on (5.4.4b). □ □

5.5 Elliptic regularity bounds on the Riesz representation of the residual

An important ingredient for our estimates is a bound on $\|\mathscr{r}_{s_{ih}}\|$ of the Riesz representation $\mathscr{r}_{s_{ih}} \in V$ of the residual $\mathrm{Res}_\theta(s_{ih}, \lambda_{ih})$ given by (5.2.8a). We now derive a sharp estimate on $\|\mathscr{r}_{s_{ih}}\|$ under an elliptic regularity assumption.

Let $\zeta_{(\mathscr{r}_{s_{ih}})}$ be the weak solution of the Laplace *source problem* $-\Delta\zeta_{(\mathscr{r}_{s_{ih}})} = \mathscr{r}_{s_{ih}}$ in Ω and $\zeta_{(\mathscr{r}_{s_{ih}})} = 0$ on $\partial\Omega$, i.e., $\zeta_{(\mathscr{r}_{s_{ih}})} \in V$ is such that

$$(\nabla\zeta_{(\mathscr{r}_{s_{ih}})}, \nabla v) = (\mathscr{r}_{s_{ih}}, v) \qquad \forall v \in V. \tag{5.5.1}$$

We will use an Aubin–Nitsche duality argument, see the references in [52, Section 5.1] for the conforming setting and, e.g., Antonietti *et al.* [7, 8] and the references therein for the nonconforming setting. Recalling the lowest-order $H^1_0(\Omega)$- and $\mathbf{H}(\mathrm{div},\Omega)$-conforming finite element spaces V_h^1 and $\mathbf{V}_h^0$ from Section 5.2.2, and denoting Π_0 the $L^2(\Omega)$-orthogonal projection onto piecewise constants, let:

Assumption 5.5.1 (Elliptic regularity)**.** *The solution $\zeta_{(\mathscr{r}_{s_{ih}})}$ of problem* (5.5.1) *belongs to the space $H^{1+\delta}(\Omega)$, $0 < \delta \le 1$, so that the approximation and stability estimates*

$$\min_{v_h \in V_h^1} \|\nabla(\zeta_{(\mathscr{r}_{s_{ih}})} - v_h)\| \le C_{\mathrm{I}} h^\delta |\zeta_{(\mathscr{r}_{s_{ih}})}|_{H^{1+\delta}(\Omega)}, \tag{5.5.2a}$$

$$|\zeta_{(\mathscr{r}_{s_{ih}})}|_{H^{1+\delta}(\Omega)} \le C_{\mathrm{S}} \|\mathscr{r}_{s_{ih}}\| \tag{5.5.2b}$$

are satisfied. Let moreover, for a suitable $\mathbf{v}_h \in \mathbf{V}_h^0$ such that $\nabla\cdot\mathbf{v}_h = \Pi_0(\mathscr{r}_{s_{ih}})$, the approximation and stability estimates

$$\|\nabla\zeta_{(\mathscr{r}_{s_{ih}})} + \mathbf{v}_h\| \le \bar{C}_{\mathrm{I}} h^\delta |\zeta_{(\mathscr{r}_{s_{ih}})}|_{H^{1+\delta}(\Omega)}, \tag{5.5.3a}$$

$$\|\mathbf{v}_h\| \le \bar{C}_{\mathrm{S}} \|\nabla\zeta_{(\mathscr{r}_{s_{ih}})}\| \tag{5.5.3b}$$

hold. Let finally the inverse inequality

$$\|\mathbf{v}_h\cdot\mathbf{n}_e\|_e \le C_{\mathrm{inv}} h_e^{-\frac{1}{2}} \|\mathbf{v}_h\|_K \qquad \forall K \in \mathcal{T}_h, \forall e \in \mathcal{E}_K \tag{5.5.4}$$

hold for all $\mathbf{v}_h \in \mathbf{V}_h^0$, where $\mathcal{E}_K$ stands for the faces of the simplex K.

Remark 5.5.2 (Constants C_{I} and C_{S})**.** *Let Ω be a convex polygon in two space dimensions. Then it is classical that $\zeta_{(\mathscr{r}_{s_{ih}})} \in H^2(\Omega)$ and $|\zeta_{(\mathscr{r}_{s_{ih}})}|_{H^2(\Omega)} = \|\Delta\zeta_{(\mathscr{r}_{s_{ih}})}\| = \|\mathscr{r}_{s_{ih}}\|$, so that $\delta = 1$ and $C_{\mathrm{S}} = 1$, see Grisvard [122, Theorem 4.3.1.4]. In this situation, calculable bounds on C_{I} can be found in Liu and Kikuchi [177] and Carstensen* et al. *[65], see also the references therein. In particular, according to [177, equation (46)] with the notation therefrom, $C_{\mathrm{I}} = 0.493 \max_{K\in\mathcal{T}_h} \frac{1+|\cos(\theta_K)|}{\sin(\theta_K)} \sqrt{\frac{\nu_+(\alpha_K,\theta_K)}{2}} \frac{h_K^{[177]}}{h_K}$ for unstructured triangular meshes and $C_{\mathrm{I}} \le \frac{0.493}{\sqrt{2}}$ for a mesh formed by isosceles right-angled triangles.*

Remark 5.5.3 (Constants $\bar{C}_{\mathrm{I}}$, $\bar{C}_{\mathrm{S}}$, and C_{inv})**.** *As above, the ideal case is $\zeta_{(\mathscr{r}_{s_{ih}})} \in H^2(\Omega)$, which happens in particular when Ω is a convex polygon in two space dimensions. Then $\delta = 1$ and calculable bounds on $\bar{C}_{\mathrm{I}}$ can be found in Mao and Shi [186] and Carstensen* et al. *[65] for the choice $\mathbf{v}_h$ as the Raviart–Thomas–Nédélec interpolate of $-\nabla\zeta_{(\mathscr{r}_{s_{ih}})}$. In particular, following [65],*

$$\bar{C}_{\mathrm{I}} = \max_{K\in\mathcal{T}_h} \max_{\alpha \text{ angle of } K} \sqrt{\frac{1/4 + 2/j_{1,1}^2}{1 - |\cos(\alpha)|}}, \tag{5.5.5}$$

where $j_{1,1} \approx 3.8317059702$ *is the first positive root of the Bessel function* J_1*. This in particular gives* $\bar{C}_{\mathrm{I}} = \sqrt{1/4 + 2/j_{1,1}^2} \approx 0.6215$ *for a structured mesh with isosceles right-angled triangles. For this interpolate,* (5.5.3b) *holds, without any regularity assumption beyond* $-\nabla\zeta_{(\boldsymbol{\imath} s_{ih})} \in \mathbf{L}^q(\Omega)$*,* $q > 2$*. Finally,* (5.5.4) *holds for any* $\mathbf{v}_h \in \mathbf{V}_h^0$ *and* C_{inv} *only depends on the shape regularity of the mesh and on the space dimension* d*, as* $\mathbf{v}_h$ *is from the lowest-order space.*

Theorem 5.5.4 (Elliptic regularity bound on $\|\boldsymbol{\imath} s_{ih}\|$)**.** *Let* $(u_{ih}, \lambda_{ih}) \in H^1(\mathcal{T}_h) \times \mathbb{R}^+$ *and let Assumptions 5.3.1 and 5.5.1 hold. Then*

$$\begin{aligned}\|\boldsymbol{\imath} s_{ih}\| \le \frac{\lambda_{ih}}{\lambda_1}\|u_{ih}-s_{ih}\| + C_{\mathrm{I}}C_{\mathrm{S}}h^{\delta}\|\mathrm{Res}_\theta(u_{ih},\lambda_{ih})\|_{-1} + \bar{C}_{\mathrm{I}}C_{\mathrm{S}}h^{\delta}\|\nabla_\theta(u_{ih}-s_{ih})\| \\ + \|\Pi_0(u_{ih}-s_{ih})\| + |\theta-1|(d+1)\frac{C_{\mathrm{inv}}\bar{C}_{\mathrm{S}}}{\sqrt{\lambda_1}}\left\{\sum_{e\in\mathcal{E}_h} h_e^{-1}\|\Pi_e^0[\![u_{ih}]\!]\|_e^2\right\}^{\frac{1}{2}}. \end{aligned} \tag{5.5.6}$$

Proof. By the definition (5.5.1) of $\zeta_{(\boldsymbol{\imath} s_{ih})}$, the definition (5.2.8a) of $\boldsymbol{\imath} s_{ih}$, the definition (5.2.7a) of $\mathrm{Res}_\theta(s_{ih}, \lambda_{ih})$, and the orthogonality Assumption 5.3.1,

$$\begin{aligned}\|\boldsymbol{\imath} s_{ih}\|^2 &= (\nabla\zeta_{(\boldsymbol{\imath} s_{ih})}, \nabla\,\boldsymbol{\imath} s_{ih}) = \lambda_{ih}(s_{ih}, \zeta_{(\boldsymbol{\imath} s_{ih})}) - (\nabla s_{ih}, \nabla\zeta_{(\boldsymbol{\imath} s_{ih})}) \\ &= \lambda_{ih}(s_{ih}-u_{ih}, \zeta_{(\boldsymbol{\imath} s_{ih})}) + \lambda_{ih}(u_{ih}, \zeta_{(\boldsymbol{\imath} s_{ih})}) \\ &\quad - (\nabla_\theta u_{ih}, \nabla\zeta_{(\boldsymbol{\imath} s_{ih})}) - (\nabla_\theta(s_{ih}-u_{ih}), \nabla\zeta_{(\boldsymbol{\imath} s_{ih})}) \\ &= \lambda_{ih}(s_{ih}-u_{ih}, \zeta_{(\boldsymbol{\imath} s_{ih})}) + \lambda_{ih}(u_{ih}, \zeta_{(\boldsymbol{\imath} s_{ih})}-\zeta_h) \\ &\quad - (\nabla_\theta u_{ih}, \nabla(\zeta_{(\boldsymbol{\imath} s_{ih})}-\zeta_h)) - (\nabla_\theta(s_{ih}-u_{ih}), \nabla\zeta_{(\boldsymbol{\imath} s_{ih})}),\end{aligned}$$

where $\zeta_h \in V_h^1$ is arbitrary. One more application of (5.2.7a), (5.2.8a) then leads to

$$\begin{aligned}&\|\boldsymbol{\imath} s_{ih}\|^2 \\ &= \lambda_{ih}(s_{ih}-u_{ih}, \zeta_{(\boldsymbol{\imath} s_{ih})}) + (\nabla\,\boldsymbol{\imath} u_{ih}, \nabla(\zeta_{(\boldsymbol{\imath} s_{ih})}-\zeta_h)) - (\nabla_\theta(s_{ih}-u_{ih}), \nabla\zeta_{(\boldsymbol{\imath} s_{ih})}) \\ &\le \lambda_{ih}\|s_{ih}-u_{ih}\|\|\zeta_{(\boldsymbol{\imath} s_{ih})}\| + \|\nabla\,\boldsymbol{\imath} u_{ih}\|\|\nabla(\zeta_{(\boldsymbol{\imath} s_{ih})}-\zeta_h)\| - (\nabla_\theta(s_{ih}-u_{ih}), \nabla\zeta_{(\boldsymbol{\imath} s_{ih})}),\end{aligned}$$

where we have also employed the Cauchy–Schwarz inequality. Now the Poincaré–Friedrichs inequality (5.2.2) gives $\|\zeta_{(\boldsymbol{\imath} s_{ih})}\| \le \|\nabla\zeta_{(\boldsymbol{\imath} s_{ih})}\|/\sqrt{\lambda_1}$ and we have from (5.5.1) that $\|\nabla\zeta_{(\boldsymbol{\imath} s_{ih})}\| \le \|\boldsymbol{\imath} s_{ih}\|/\sqrt{\lambda_1}$. For the second term above, we need to take the best ζ_h and employ the estimates (5.5.2) to arrive at

$$\begin{aligned}\|\boldsymbol{\imath} s_{ih}\|^2 \le \left(\frac{\lambda_{ih}}{\lambda_1}\|u_{ih}-s_{ih}\| + C_{\mathrm{I}}C_{\mathrm{S}}h^{\delta}\|\nabla\,\boldsymbol{\imath} u_{ih}\|\right)\|\boldsymbol{\imath} s_{ih}\| \\ - (\nabla\zeta_{(\boldsymbol{\imath} s_{ih})}, \nabla_\theta(s_{ih}-u_{ih})).\end{aligned}$$

Let now $\mathbf{v}_h \in \mathbf{V}_h^0$ be such that $\nabla\cdot\mathbf{v}_h = \Pi_0(\boldsymbol{\imath} s_{ih})$. Definition (5.2.5) of the discrete gradient and the fact that $\mathbf{v}_h \in \mathbf{V}_h^0 \subset \mathbf{V}^0(\mathcal{T}_h)$ give

$$\begin{aligned}-(\mathbf{v}_h, \nabla_\theta u_{ih}) &= -(\mathbf{v}_h, \nabla_h u_{ih}) + \theta\sum_{e\in\mathcal{E}_h}(\mathbf{v}_h, \mathfrak{l}_e([\![u_{ih}]\!])) \\ &= -(\mathbf{v}_h, \nabla_h u_{ih}) + \theta\sum_{e\in\mathcal{E}_h}\langle\{\!\{\mathbf{v}_h\}\!\}\cdot\mathbf{n}_e, [\![u_{ih}]\!]\rangle_e.\end{aligned}$$

Thus, using that $\mathbf{v}_h \in \mathbf{H}(\mathrm{div}, \Omega)$ (so that $\{\!\{\mathbf{v}_h\}\!\}\cdot\mathbf{n}_e = \mathbf{v}_h\cdot\mathbf{n}_e$), $s_{ih} \in V$, and elementwise the Green theorem gives

$$(\mathbf{v}_h, \nabla_\theta(s_{ih}-u_{ih})) = -\sum_{K\in\mathcal{T}_h}(\nabla\cdot\mathbf{v}_h, s_{ih}-u_{ih})_K + (\theta-1)\sum_{e\in\mathcal{E}_h}\langle\mathbf{v}_h\cdot\mathbf{n}_e, [\![u_{ih}]\!]\rangle_e.$$

The last term above actually disappears when the jumps of u_{ih} are of mean value 0, i.e., $\langle [\![u_{ih}]\!], 1\rangle_e = 0$ for all $e \in \mathcal{E}_h$, or when $\theta = 1$. As $\mathbf{v}_h{\cdot}\mathbf{n}_e \in \mathbb{P}_0(e)$, we can, in general, at least replace $[\![u_{ih}]\!]$ by $\Pi_e^0[\![u_{ih}]\!]$ and estimate this term using the inverse inequality (5.5.4) and Cauchy–Schwarz one, as each simplex has $d+1$ faces

$$\left|\sum_{e\in\mathcal{E}_h}\langle \mathbf{v}_h{\cdot}\mathbf{n}_e, \Pi_e^0[\![u_{ih}]\!]\rangle_e\right| \le \sum_{e\in\mathcal{E}_h}\left\{\|\mathbf{v}_h\|_{K\in\mathcal{T}_h;\, e\in\mathcal{E}_K} C_{\mathrm{inv}} h_e^{-\frac{1}{2}} \|\Pi_e^0[\![u_{ih}]\!]\|_e\right\}$$
$$\le (d+1) C_{\mathrm{inv}} \|\mathbf{v}_h\| \left\{\sum_{e\in\mathcal{E}_h} h_e^{-1}\|\Pi_e^0[\![u_{ih}]\!]\|_e^2\right\}^{\frac{1}{2}}.$$

Thus, for $\mathbf{v}_h$ satisfying (5.5.3) and under the stability assumption (5.5.2b), we infer

$$\begin{aligned}
&-(\nabla\zeta_{(\boldsymbol{\imath}_{s_{ih}})}, \nabla_\theta(s_{ih}-u_{ih}))\\
=\;& -(\nabla\zeta_{(\boldsymbol{\imath}_{s_{ih}})} + \mathbf{v}_h, \nabla_\theta(s_{ih}-u_{ih})) + (\Pi_0(\boldsymbol{\imath}_{s_{ih}}), u_{ih}-s_{ih})\\
&+(\theta-1)\sum_{e\in\mathcal{E}_h}\langle \mathbf{v}_h{\cdot}\mathbf{n}_e, \Pi_e^0[\![u_{ih}]\!]\rangle_e\\
=\;& -(\nabla\zeta_{(\boldsymbol{\imath}_{s_{ih}})} + \mathbf{v}_h, \nabla_\theta(s_{ih}-u_{ih})) + (\boldsymbol{\imath}_{s_{ih}}, \Pi_0(u_{ih}-s_{ih}))\\
&+(\theta-1)\sum_{e\in\mathcal{E}_h}\langle \mathbf{v}_h{\cdot}\mathbf{n}_e, \Pi_e^0[\![u_{ih}]\!]\rangle_e\\
\le\;& \bar{C}_{\mathrm{I}} C_{\mathrm{S}} h^\delta \|\boldsymbol{\imath}_{s_{ih}}\| \|\nabla_\theta(u_{ih}-s_{ih})\| + \|\boldsymbol{\imath}_{s_{ih}}\| \|\Pi_0(u_{ih}-s_{ih})\|\\
&+|\theta-1|(d+1)\frac{C_{\mathrm{inv}}\bar{C}_{\mathrm{S}}}{\sqrt{\lambda_1}}\|\boldsymbol{\imath}_{s_{ih}}\| \left\{\sum_{e\in\mathcal{E}_h} h_e^{-1}\|\Pi_e^0[\![u_{ih}]\!]\|_e^2\right\}^{\frac{1}{2}}.
\end{aligned}$$

Combining the above estimates with the characterization (5.2.8b) of $\|\nabla \boldsymbol{\imath}_{u_{ih}}\|$ finishes the proof. □ □

5.6 Guaranteed and computable upper and lower bounds in a unified framework

We combine here the results of Sections 5.4 and 5.5 together with the key generic equivalences of [52, Section 3] to derive the actual guaranteed and computable eigenvalue and eigenvector bounds in a unified framework.

5.6.1 Eigenvalues

We first tackle the question of upper and lower bounds for the i-th eigenvalue λ_i. We refer to [52, Remark 5.4 and 5.5] for the discussion on obtaining the auxiliary eigenvalue bounds $\underline{\lambda}_1$, $\overline{\lambda}_{i-1}$, $\overline{\lambda}_i$, and $\underline{\lambda}_{i+1}$.

Theorem 5.6.1 (Guaranteed lower bounds for the i-th eigenvalue)**.** *Let the i-th exact eigenvalue λ_i, $i \ge 1$, be simple and suppose the auxiliary bounds $\underline{\lambda}_1 \le \lambda_1$, $\lambda_i \le \overline{\lambda}_i$, $\underline{\lambda}_{i+1} \le \lambda_{i+1}$, as well as $\lambda_{i-1} \le \overline{\lambda}_{i-1}$ when $i > 1$, for $\underline{\lambda}_1, \overline{\lambda}_i, \underline{\lambda}_{i+1}, \overline{\lambda}_{i-1} > 0$. Let the approximate eigenvector–eigenvalue pair $(u_{ih}, \lambda_{ih}) \in \mathbb{P}_p(\mathcal{T}_h) \times \mathbb{R}^+$ verify Assumption 5.3.1, as well as the inclusion*

$$\overline{\lambda}_{i-1} < \lambda_{ih} \text{ when } i > 1, \qquad \lambda_{ih} < \underline{\lambda}_{i+1}. \tag{5.6.1}$$

Let the equilibrated flux reconstruction $\boldsymbol{\sigma}_{ih}$ be given by Definition 5.3.2 and the eigenvector reconstruction s_{ih} by Definition 5.3.3, with $s_{ih} \neq 0$. Denote the principal estimator

$$\eta_{i,\mathrm{res}} := \frac{1}{\|s_{ih}\|}\left(\frac{\lambda_{ih}}{\sqrt{\underline{\lambda}_1}}\|u_{ih}-s_{ih}\| + \|\nabla s_{ih} + \boldsymbol{\sigma}_{ih}\|\right) \tag{5.6.2}$$

together with the discrete relative gaps

$$c_{ih} := \max\left\{\left(\frac{\lambda_{ih}}{\overline{\lambda}_{i-1}}-1\right)^{-1}, \left(1-\frac{\lambda_{ih}}{\underline{\lambda}_{i+1}}\right)^{-1}\right\}, \tag{5.6.3a}$$

$$\tilde{c}_{ih} := \max\left\{\overline{\lambda}_{i-1}^{-\frac{1}{2}}\left(\frac{\lambda_{ih}}{\overline{\lambda}_{i-1}}-1\right)^{-1}, \underline{\lambda}_{i+1}^{-\frac{1}{2}}\left(1-\frac{\lambda_{ih}}{\underline{\lambda}_{i+1}}\right)^{-1}\right\} \tag{5.6.3b}$$

and the scaled eigenvector reconstruction $\tilde{s}_{ih} := \frac{s_{ih}}{\|s_{ih}\|}$. Then, the i-th eigenvalue lower bound is

$$\|\nabla \tilde{s}_{ih}\|^2 - \eta_i^2 \leq \lambda_i, \tag{5.6.4}$$

where the complete estimator η_i takes different forms in the following three cases:
Case A (No smallness assumption) *If the sign characterization $(u_i, \tilde{s}_{ih}) \geq 0$ is known to hold, the lower i-th eigenvalue estimate* (5.6.4) *is valid with*

$$\eta_i^2 := (1 + (\lambda_{ih} + \overline{\lambda}_i)2\tilde{c}_{ih}^2)\eta_{i,\mathrm{res}}^2. \tag{5.6.5}$$

If only $(\tilde{s}_{ih}, \chi_i) > 0$ holds for the sign characterization function χ_i of Section 5.2.1, the factor 2 in (5.6.5) *needs to be replaced by $2(1 - \|\tilde{s}_{ih} - \Pi_i \tilde{s}_{ih}\|)^{-1}$, where $\Pi_i \tilde{s}_{ih}$ stands for the $L^2(\Omega)$-orthogonal projection of $\tilde{s}_{ih}$ on the span of χ_i.*
Case B (Improved estimates under a smallness assumption) *Assume the sign characterization $(\tilde{s}_{ih}, \chi_i) > 0$ and define*

$$\overline{\alpha}_{ih} := \sqrt{2}\tilde{c}_{ih}\eta_{i,\mathrm{res}}, \tag{5.6.6a}$$

where $\overline{\alpha}_{ih}$ is a computable bound on the $L^2(\Omega)$ error $\|u_i - \tilde{s}_{ih}\|$. Let the smallness assumption

$$\overline{\alpha}_{ih} \leq \min\left\{\left(\frac{2\underline{\lambda}_1}{\overline{\lambda}_i}\right)^{\frac{1}{2}}, \|\chi_i\|^{-1}(\tilde{s}_{ih}, \chi_i)\right\} \tag{5.6.6b}$$

hold, so that in particular $\frac{\overline{\lambda}_i}{\underline{\lambda}_1}\frac{\overline{\alpha}_{ih}^2}{4}$ is bounded by $\frac{1}{2}$ and tends to zero; when $i = 1$, taking $\underline{\lambda}_1 = \overline{\lambda}_i = \lambda_i$ is possible and makes the fraction $\frac{\underline{\lambda}_1}{\overline{\lambda}_i}$ vanish. Then the lower i-th eigenvalue estimate (5.6.4) *holds with*

$$\eta_i^2 := c_{ih}^2\left(1 - \frac{\overline{\lambda}_i}{\underline{\lambda}_1}\frac{\overline{\alpha}_{ih}^2}{4}\right)^{-1}\eta_{i,\mathrm{res}}^2. \tag{5.6.7}$$

Case C (Optimal estimates under elliptic regularity assumption) *Assume the elliptic regularity of Assumption 5.5.1 together with the sign characterization $(\tilde{s}_{ih}, \chi_i) > 0$. Define the $L^2(\Omega)$ estimators $\overline{\alpha}_{ih}$ of $\|u_i - \tilde{s}_{ih}\|$ by*

$$\begin{aligned}\overline{\alpha}_{ih} := \frac{\sqrt{2}c_{ih}}{\|s_{ih}\|}\Bigg(&\frac{\lambda_{ih}}{\underline{\lambda}_1}\|u_{ih}-s_{ih}\| + C_{\mathrm{I}}C_{\mathrm{S}}h^{\delta}\|\nabla_\theta u_{ih} + \boldsymbol{\sigma}_{ih}\| \\ &+ \bar{C}_{\mathrm{I}}C_{\mathrm{S}}h^{\delta}\|\nabla_\theta(u_{ih}-s_{ih})\| + \|\Pi_0(u_{ih}-s_{ih})\| \\ &+ |\theta-1|(d+1)\frac{C_{\mathrm{inv}}\bar{C}_{\mathrm{S}}}{\sqrt{\underline{\lambda}_1}}\left\{\sum_{e\in\mathcal{E}_h} h_e^{-1}\|\Pi_e^0[\![u_{ih}]\!]\|_e^2\right\}^{\frac{1}{2}}\Bigg).\end{aligned} \tag{5.6.8a}$$

Then, under the smallness assumption

$$\overline{\alpha}_{ih} \leq \|\chi_i\|^{-1}(\tilde{s}_{ih}, \chi_i), \tag{5.6.8b}$$

the lower i-th eigenvalue estimate (5.6.4) *holds with* η_i *given by*

$$\eta_i^2 := \eta_{i,\mathrm{res}}^2 + (\lambda_{ih} + \overline{\lambda}_i)\overline{\alpha}_{ih}^2. \tag{5.6.9}$$

Remark 5.6.2 (Form of the complete estimator η_i)**.** *In Cases A and B above, we immediately see*

$$\eta_i = m_{ih}\eta_{i,\mathrm{res}}, \quad m_{ih} := (1 + (\lambda_{ih} + \overline{\lambda}_i)2\tilde{c}_{ih}^2)^{\frac{1}{2}} \text{ and } m_{ih} := c_{ih}\left(1 - \frac{\overline{\lambda}_i}{\underline{\lambda}_1}\frac{\overline{\alpha}_{ih}^2}{4}\right)^{-\frac{1}{2}}$$

(up to the possible replacement of the factor 2 *in Case A) (*m_{ih} *is bounded by* $c_{ih}\sqrt{2}$ *and tends to* c_{ih} *in Case B). Thus the complete estimator* η_i *indeed takes the form* (5.1.3b) *announced in the introduction, where in particular the key role of* $\eta_{i,\mathrm{res}}$ *given by* (5.6.2) *and the unfavorable multiplication by the discrete relative gaps* c_{ih} *or* $\tilde{c}_{ih}$ *of* (5.6.3) *are obvious. In Case C,* η_i *rather takes an additive form, with the key estimator* $\eta_{i,\mathrm{res}}$ *supplemented by a term containing a multiplication by the discrete relative gap* c_{ih}*, which is, though, typically of higher order.*

Theorem 5.6.3 (Guaranteed upper bound for the i-th eigenvalue)**.** *For a given* $i \geq 1$*, let the approximate eigenvectors* $u_{kh} \in \mathbb{P}_p(\mathcal{T}_h)$*,* $1 \leq k \leq i$*, be arbitrary, and let* s_{kh}*,* $1 \leq k \leq i$*, be their eigenvector reconstructions by Definition 5.3.3. Suppose that* s_{kh}*,* $1 \leq k \leq i$*, are linearly independent. Then*

$$\lambda_i \leq \max_{\boldsymbol{\xi}\in\mathbb{R}^i,\, \|\boldsymbol{\xi}\|=1} \frac{\|\nabla \sum_{k=1}^i \boldsymbol{\xi}_k s_{kh}\|^2}{\|\sum_{k=1}^i \boldsymbol{\xi}_k s_{kh}\|^2}, \tag{5.6.10}$$

where $\|\boldsymbol{\xi}\|^2 = \sum_{k=1}^i \boldsymbol{\xi}_k^2$*. In particular*

$$\lambda_1 \leq \frac{\|\nabla s_{1h}\|^2}{\|s_{1h}\|^2} = \|\nabla \tilde{s}_{1h}\|^2. \tag{5.6.11}$$

Proof of Theorem 5.6.3. The statement follows immediately from the min–max principle

$$\lambda_i = \min_{V_i \subset V,\, \dim V_i = i} \max_{v \in V_i} \frac{\|\nabla v\|^2}{\|v\|^2}. \qquad \square$$

□

In what concerns Theorem 5.6.1, we prove the three cases separately:

Proof of Theorem 5.6.1, **Case A**. It is immediate from estimate (5.4.5a) of Corollary 5.4.1 and from definition (5.6.2) of the principal estimator $\eta_{i,\mathrm{res}}$ together with the scaling $\tilde{s}_{ih} = s_{ih}/\|s_{ih}\|$ that the dual norm of the residual of $(\tilde{s}_{ih}, \lambda_{ih})$ can be estimated as

$$\|\mathrm{Res}_\theta(\tilde{s}_{ih}, \lambda_{ih})\|_{-1} \leq \eta_{i,\mathrm{res}}. \tag{5.6.12}$$

If $(u_i, \tilde{s}_{ih}) \geq 0$ is known to hold, define $\overline{\alpha}_{ih}$ by (5.6.6a). From [52, Lemma 3.2] in combination with (5.2.8b), we then infer the $L^2(\Omega)$ bound

$$\|u_i - \tilde{s}_{ih}\| \leq \sqrt{2}\tilde{c}_{ih}\|\mathrm{Res}_\theta(\tilde{s}_{ih}, \lambda_{ih})\|_{-1} \leq \overline{\alpha}_{ih}. \tag{5.6.13}$$

Now the upper bound in [52, Theorem 3.4], in combination with the first bound of [52, Theorem 3.5], gives

$$\|\nabla \tilde{s}_{ih}\|^2 - \lambda_i \leq \|\nabla(u_i - \tilde{s}_{ih})\|^2 \leq \|\mathrm{Res}_\theta(\tilde{s}_{ih}, \lambda_{ih})\|_{-1}^2 + (\lambda_{ih} + \overline{\lambda}_i)\overline{\alpha}_{ih}^2, \tag{5.6.14}$$

and one more use of (5.6.12) proves (5.6.4) with η_i given by (5.6.5). If only $(\tilde{s}_{ih}, \chi_i) > 0$ holds, we take

$$\overline{\alpha}_{ih} := \sqrt{2}(1 - \|\tilde{s}_{ih} - \Pi_i \tilde{s}_{ih}\|)^{-\frac{1}{2}} \tilde{c}_{ih} \eta_{i,\mathrm{res}}$$

and proceed as in [52, proof of Theorem 5.1, Case A] to find

$$\|u_i - \tilde{s}_{ih}\| \leq \sqrt{2}(1 - \|\tilde{s}_{ih} - \Pi_i \tilde{s}_{ih}\|)^{-\frac{1}{2}} \tilde{c}_{ih} \|\mathrm{Res}_\theta(\tilde{s}_{ih}, \lambda_{ih})\|_{-1} \leq \overline{\alpha}_{ih}$$

instead of (5.6.13), and we conclude as above. □ □

Proof of Theorem 5.6.1, **Case B**. The second condition in (5.6.6b) implies that assumptions of [52, Lemma 3.3] are verified for $\tilde{s}_{ih}$. Thus the $L^2(\Omega)$ bound (5.6.13) is valid for $\overline{\alpha}_{ih}$ given by (5.6.6a). The first condition in (5.6.6b) then allows us to use the improved estimate in [52, Theorem 3.5]. In combination with the upper bound in [52, Theorem 3.4], this gives

$$\|\nabla \tilde{s}_{ih}\|^2 - \lambda_i \leq \|\nabla(u_i - \tilde{s}_{ih})\|^2 \leq c_{ih}^2 \left(1 - \frac{\overline{\lambda}_i}{\underline{\lambda}_1} \frac{\overline{\alpha}_{ih}^2}{4}\right)^{-1} \|\mathrm{Res}_\theta(\tilde{s}_{ih}, \lambda_{ih})\|_{-1}^2, \tag{5.6.15}$$

and we conclude by (5.6.12). □ □

Proof of Theorem 5.6.1, **Case C**. Theorem 5.5.4 together with the bound $\|\mathrm{Res}_\theta(u_{ih}, \lambda_{ih})\|_{-1} \leq \|\nabla_\theta u_{ih} + \boldsymbol{\sigma}_{ih}\|$ and the scaling $\tilde{s}_{ih} = s_{ih}/\|s_{ih}\|$ imply

$$\|\mathscr{E}_{\tilde{s}_{ih}}\| \leq \frac{\overline{\alpha}_{ih}}{\sqrt{2} c_{ih}},$$

where $\overline{\alpha}_{ih}$ is given by (5.6.8a). Next, condition (5.6.8b) implies that assumptions of [52, Lemma 3.3] are verified for $\tilde{s}_{ih}$ and consequently $(u_i, \tilde{s}_{ih}) \geq 0$. Thus [52, Lemma 3.1] again gives the computable $L^2(\Omega)$ bound

$$\|u_i - \tilde{s}_{ih}\| \leq \overline{\alpha}_{ih}.$$

We then conclude as in Case A via (5.6.14). □ □

5.6.2 Eigenvectors

We now turn to the estimates on the error in the approximation of the i-th exact eigenvector u_i by u_{ih} and their efficiency and robustness with respect to the polynomial degree of u_{ih}.

Theorem 5.6.4 (Guaranteed and robust bounds for the i-th eigenvector error)**.** *Let the assumptions of Theorem 5.6.1 be verified. Then the energy eigenvector error can be bounded via*

$$\|\nabla_\theta(u_i - u_{ih})\| \leq \eta_i + \|\nabla_\theta(u_{ih} - \tilde{s}_{ih})\|, \tag{5.6.16}$$

where η_i is defined in the three cases A, B, and C respectively by (5.6.5), (5.6.7), *and* (5.6.9). *Moreover, this estimate is efficient as*

$$\begin{aligned} \eta_i + \|\nabla_\theta(u_{ih} - \tilde{s}_{ih})\| \leq C_i \Bigg(& \|\nabla_\theta(u_i - u_{ih})\| + \left\{ \sum_{e \in \mathcal{E}_h} h_e^{-1} \|\Pi_e^0 [\![u_{ih}]\!]\|_e^2 \right\}^{\frac{1}{2}} \\ & + |1 - \|u_{ih}\|| + \left|\lambda_{ih} - \|\nabla_\theta u_{ih}\|^2\right| \Bigg), \end{aligned} \tag{5.6.17}$$

where the generic constants C_i can be determined from the detailed estimates

- *efficiency of* $\|\nabla_\theta(u_{ih}-\tilde{s}_{ih})\|$*:*

$$\|\nabla_\theta(u_{ih}-\tilde{s}_{ih})\| \le \|\nabla_\theta(u_{ih}-s_{ih})\| + |1-\|s_{ih}\||\frac{\|\nabla s_{ih}\|}{\|s_{ih}\|}, \tag{5.6.18a}$$

$$\|\nabla_\theta(u_{ih}-s_{ih})\| \le \|\nabla_h(u_{ih}-s_{ih})\| + |\theta|\sqrt{d+1}C_{\mathrm{inv}}\left\{\sum_{e\in\mathcal{E}_h} h_e^{-1}\|\Pi_e^0[\![u_{ih}]\!]\|_e^2\right\}^{\frac{1}{2}}, \tag{5.6.18b}$$

together with (5.4.7)*, inequalities* (5.6.21) *and* (5.6.19) *below, and*

$$\|\nabla_h(u_i-u_{ih})\| \le \|\nabla_\theta(u_i-u_{ih})\| + |\theta|\sqrt{d+1}C_{\mathrm{inv}}\left\{\sum_{e\in\mathcal{E}_h} h_e^{-1}\|\Pi_e^0[\![u_{ih}]\!]\|_e^2\right\}^{\frac{1}{2}}; \tag{5.6.18c}$$

- *efficiency of* $\|u_{ih}-s_{ih}\|$ *(first part of* $\eta_{i,\mathrm{res}}$*):*

$$\|u_{ih}-s_{ih}\| \le C_{\mathrm{bF}}\left(\|\nabla_h(u_{ih}-s_{ih})\|^2 + \sum_{e\in\mathcal{E}_h^{\mathrm{int}}} h_e^{-1}\|\Pi_e^0[\![u_{ih}]\!]\|_e^2 + \langle u_{ih},1\rangle_{\partial\Omega}^2\right)^{\frac{1}{2}}, \tag{5.6.19a}$$

$$|\langle u_{ih},1\rangle_{\partial\Omega}| \le h^{\frac{1}{2}}|\partial\Omega|^{\frac{1}{2}}\left\{\sum_{e\in\mathcal{E}_h^{\mathrm{ext}}} h_e^{-1}\|\Pi_e^0[\![u_{ih}]\!]\|_e^2\right\}^{\frac{1}{2}}, \tag{5.6.19b}$$

together with (5.4.7) *and* (5.6.18c)*;*

- *efficiency of* $\|\nabla s_{ih}+\boldsymbol{\sigma}_{ih}\|$ *(second part of* $\eta_{i,\mathrm{res}}$*):*

$$\|\nabla s_{ih}+\boldsymbol{\sigma}_{ih}\| \le \|\nabla_\theta u_{ih}+\boldsymbol{\sigma}_{ih}\| + \|\nabla_\theta(u_{ih}-s_{ih})\|, \tag{5.6.20a}$$

$$\|\nabla_\theta u_{ih}+\boldsymbol{\sigma}_{ih}\| \le (d+1)C_{\mathrm{st}}C_{\mathrm{cont,PF}}\left(\frac{\lambda_{ih}}{\sqrt{\lambda_1}}\|u_{ih}-s_{ih}\| + \|\nabla_\theta(u_{ih}-s_{ih})\| + \|\mathrm{Res}_\theta(s_{ih},\lambda_{ih})\|_{-1}\right), \tag{5.6.20b}$$

$$\|\mathrm{Res}_\theta(s_{ih},\lambda_{ih})\|_{-1} \le \frac{\|s_{ih}\|}{\sqrt{\lambda_i}}|\lambda_{ih}-\lambda_i| + \overline{C}_{ih}^{\frac{1}{2}}\|s_{ih}\|\|\nabla(u_i-\tilde{s}_{ih})\|, \tag{5.6.20c}$$

$$\overline{C}_{ih} := 1 \text{ if } i=1,$$

$$\overline{C}_{ih} := \max\left\{\left(\frac{\lambda_{ih}}{\lambda_1}-1\right)^2, 1\right\} \text{ if } i>1,$$

$$|\lambda_{ih}-\lambda_i| \le \|\nabla_\theta(u_i-u_{ih})\|^2 + 2\|\nabla_\theta(u_i-u_{ih})\|\|\nabla_\theta u_{ih}\| + \left|\lambda_{ih}-\|\nabla_\theta u_{ih}\|^2\right|, \tag{5.6.20d}$$

together with $\|\nabla(u_i-\tilde{s}_{ih})\| \le \|\nabla_\theta(u_i-u_{ih})\| + \|\nabla_\theta(u_{ih}-\tilde{s}_{ih})\|$*,* (5.6.18)*,* (5.6.19)*, and* (5.6.21) *below;*

- *inequalities for the scaling terms:*

$$|1-\|s_{ih}\|| \le |1-\|u_{ih}\|| + \|u_{ih}-s_{ih}\|, \tag{5.6.21a}$$

$$\|\nabla s_{ih}\| \le \|\nabla_\theta u_{ih}\| + \|\nabla_\theta(u_{ih}-s_{ih})\|, \tag{5.6.21b}$$

$$\|u_{ih}\| - \|u_{ih}-s_{ih}\| \le \|s_{ih}\| \le \|u_{ih}\| + \|u_{ih}-s_{ih}\|; \tag{5.6.21c}$$

- *note that $\overline{\alpha}_{ih}$ given by* (5.6.8a) *only contains terms treated above (possibly with multiplicative factors).*

Proof. The reliability (5.6.16) is a combination of the triangle inequality

$$\|\nabla_\theta(u_i - u_{ih})\| \le \|\nabla_\theta(u_i - \tilde{s}_{ih})\| + \|\nabla_\theta(u_{ih} - \tilde{s}_{ih})\|$$

together with the bounds $\|\nabla_\theta(u_i - \tilde{s}_{ih})\| = \|\nabla(u_i - \tilde{s}_{ih})\| \le \eta_i$ shown in (5.6.14) and (5.6.15) in the proof of Theorem 5.6.1. The rest of the proof is dedicated to showing the efficiency (5.6.17).

We first examine the second term on the right-hand side of (5.6.16). The definition $\tilde{s}_{ih} := \frac{s_{ih}}{\|s_{ih}\|}$ and the triangle inequality give (5.6.18a), since

$$\|\nabla(s_{ih} - \tilde{s}_{ih})\| = |1 - \|s_{ih}\||\frac{\|\nabla s_{ih}\|}{\|s_{ih}\|}.$$

As for the first term therein,

$$\begin{aligned}
\|\nabla_\theta(u_{ih} - s_{ih})\| &= \left\| \nabla_h(u_{ih} - s_{ih}) - \theta \sum_{e\in\mathcal{E}_h} \mathfrak{l}_e([\![u_{ih}]\!]) \right\| \\
&\le \|\nabla_h(u_{ih} - s_{ih})\| + |\theta| \left\{ \sum_{K\in\mathcal{T}_h} \left\| \sum_{e\in\mathcal{E}_K} \mathfrak{l}_e([\![u_{ih}]\!]) \right\|_K^2 \right\}^{\frac12} \\
&\le \|\nabla_h(u_{ih} - s_{ih})\| + |\theta| \left\{ \sum_{K\in\mathcal{T}_h} (d+1) \sum_{e\in\mathcal{E}_K} \|\mathfrak{l}_e([\![u_{ih}]\!])\|_K^2 \right\}^{\frac12} \\
&= \|\nabla_h(u_{ih} - s_{ih})\| + |\theta| \left\{ (d+1) \sum_{e\in\mathcal{E}_h} \|\mathfrak{l}_e([\![u_{ih}]\!])\|_{\mathcal{T}_e}^2 \right\}^{\frac12},
\end{aligned}$$

a direct consequence of the definition of the discrete gradient (5.2.5), of the triangle and Cauchy–Schwarz inequalities, and of the fact that $\mathfrak{l}_e([\![u_{ih}]\!])$ is only supported on the (1 or 2) elements in $\mathcal{T}_e$ containing the face e. Next, the definition of the face lifting $\mathfrak{l}_e$ from Section 5.2.3, the fact that $\mathbf{v}_h{\cdot}\mathbf{n}_e$ are constants for $\mathbf{v}_h \in \mathbf{V}^0(\mathcal{T}_e)$, and the inverse inequality (5.5.4) give

$$\begin{aligned}
\|\mathfrak{l}_e([\![u_{ih}]\!])\|_{\mathcal{T}_e} &= \sup_{\mathbf{v}_h\in\mathbf{V}^0(\mathcal{T}_e);\,\|\mathbf{v}_h\|_{\mathcal{T}_e}=1} (\mathfrak{l}_e([\![u_{ih}]\!]), \mathbf{v}_h)_{\mathcal{T}_e} \\
&= \sup_{\mathbf{v}_h\in\mathbf{V}^0(\mathcal{T}_e);\,\|\mathbf{v}_h\|_{\mathcal{T}_e}=1} \langle \{\!\{\mathbf{v}_h\}\!\}{\cdot}\mathbf{n}_e, \Pi_e^0[\![u_{ih}]\!]\rangle_e \\
&\le C_{\mathrm{inv}} h_e^{-\frac12} \|\Pi_e^0[\![u_{ih}]\!]\|_e.
\end{aligned}$$

Combining the two above bounds gives (5.6.18b). Finally, (5.6.18c) follows by, using again (5.2.5),

$$\|\nabla_h(u_i - u_{ih})\| = \left\| \nabla_\theta(u_i - u_{ih}) - \theta \sum_{e\in\mathcal{E}_h} \mathfrak{l}_e([\![u_{ih}]\!]) \right\|$$

and proceeding as above for the liftings. Concerning the second term in (5.6.18a), the multiplicative factor $\|\nabla s_{ih}\|$ approaches $\|\nabla_\theta u_{ih}\| \approx \sqrt{\lambda}_{ih}$ as manifested in (5.6.21b), the multiplicative factor $\|s_{ih}\|$ is of order 1 when $\|u_{ih}\| \approx 1$ as shown in (5.6.21c), and $|1 - \|s_{ih}\||$ is bounded in (5.6.21a) by the consistency term $|1 - \|u_{ih}\||$ and the estimator $\|u_{ih} - s_{ih}\|$ efficient via (5.6.19). Thus the efficiency for the term $\|\nabla_\theta(u_{ih} - \tilde{s}_{ih})\|$ as announced in (5.6.17) follows.

We now turn to the $L^2(\Omega)$-term $\|u_{ih} - s_{ih}\|$, the first part of the estimator $\eta_{i,\mathrm{res}}$ given by (5.6.2). Note that $\eta_{i,\mathrm{res}}$ forms the principal part of η_i in all three cases A, B, and C, and that the scaling factor $1/\|s_{ih}\|$ is of order 1, see (5.6.21c). First, (5.6.19a) is a consequence of the broken Poincaré–Friedrichs inequality (5.2.4) with $v = u_{ih} - s_{ih}$ and of the fact that the jumps of s_{ih} are zero. The last term in (5.6.19a) can then still be bounded by

$$\langle u_{ih}, 1\rangle_{\partial\Omega}^2 = \left\{ \sum_{e\in\mathcal{E}_h^{\mathrm{ext}}} \langle \Pi_e^0[\![u_{ih}]\!], 1\rangle_e \right\}^2 \le h \left\{ \sum_{e\in\mathcal{E}_h^{\mathrm{ext}}} h_e^{-\frac{1}{2}} \|\Pi_e^0[\![u_{ih}]\!]\|_e |e|^{\frac{1}{2}} \right\}^2,$$

so that (5.6.19b) follows by another Cauchy–Schwarz inequality. The efficiency of $\|u_{ih} - s_{ih}\|$ is then completed by (5.4.7) and (5.6.18c). Numerically, though, we have observed that $\|u_{ih} - s_{ih}\|$ converges still one order faster than what (5.6.19) suggests, so that it becomes negligible in practice.

We now turn to the second part of the estimator $\eta_{i,\mathrm{res}}$ of (5.6.2), $\|\nabla s_{ih} + \boldsymbol{\sigma}_{ih}\|$. To begin with, (5.6.20a) follows by the triangle inequality; the second term therein has been treated in (5.6.18). For $\|\nabla_\theta u_{ih} + \boldsymbol{\sigma}_{ih}\|$, we have the crucial bound (5.4.5b). As, however, $u_{ih} \notin H_0^1(\Omega)$, we need to get back from $\|\mathrm{Res}_\theta(u_{ih}, \lambda_{ih})\|_{-1}$ to $\|\mathrm{Res}_\theta(s_{ih}, \lambda_{ih})\|_{-1}$ to prove the efficiency. For this purpose, let $v \in V$ with $\|\nabla v\| = 1$ be fixed. Using the residual definition (5.2.7a), the Cauchy–Schwarz and Poincaré–Friedrichs (5.2.2) inequalities, and the dual norm definition (5.2.7b),

$$\begin{aligned}
&\langle \mathrm{Res}_\theta(u_{ih}, \lambda_{ih}), v\rangle_{V',V} \\
&= \lambda_{ih}(u_{ih}, v) - (\nabla_\theta u_{ih}, \nabla v) \\
&= \lambda_{ih}(u_{ih} - s_{ih}, v) - (\nabla_\theta(u_{ih} - s_{ih}), \nabla v) + \langle \mathrm{Res}_\theta(s_{ih}, \lambda_{ih}), v\rangle_{V',V} \\
&\le \frac{\lambda_{ih}}{\sqrt{\lambda_1}} \|u_{ih} - s_{ih}\| + \|\nabla_\theta(u_{ih} - s_{ih})\| + \|\mathrm{Res}_\theta(s_{ih}, \lambda_{ih})\|_{-1},
\end{aligned}$$

so that (5.6.20b) follows. We know from (5.6.18) and (5.6.19) that the first two terms herein are efficient, so that we pursue with the last one only. To start with, note that $\|\mathrm{Res}_\theta(s_{ih}, \lambda_{ih})\|_{-1} = \|s_{ih}\| \|\mathrm{Res}_\theta(\tilde{s}_{ih}, \lambda_{ih})\|_{-1}$; then (5.6.20c) follows from the proof of the second bound in [52, Theorem 3.5]. To finish, develop

$$\begin{aligned}
\lambda_{ih} - \lambda_i &= \lambda_{ih} - \|\nabla_\theta(u_i - u_{ih} + u_{ih})\|^2 \\
&= \lambda_{ih} - \|\nabla_\theta(u_i - u_{ih})\|^2 - 2(\nabla_\theta(u_i - u_{ih}), \nabla_\theta u_{ih}) - \|\nabla_\theta u_{ih}\|^2,
\end{aligned}$$

which proves (5.6.20d) and together with (5.6.18), (5.6.19), and (5.6.21) gives the requested efficiency. □ □

5.6.3 Comments

We now give some comments on the results of Theorems 5.6.1, 5.6.3, and 5.6.4; a discussion for the conforming setting can be found in [52, Section 5.3].

Remark 5.6.5 (Vanishing consistency terms)**.** *Nonconforming finite elements (with $\theta = 0$ in the discrete gradient* (5.2.5)*) are a particular example of a numerical method where both consistency terms $|1 - \|u_{ih}\||$ and $\left|\lambda_{ih} - \|\nabla_\theta u_{ih}\|^2\right|$ are zero and thus vanish in* (5.6.17)*, see Section 5.7.1 below.*

Remark 5.6.6 (Jumps of mean value zero)**.** *A particular situation arises when $\langle [\![u_{ih}]\!], 1\rangle_e = 0$ for all $e \in \mathcal{E}_h$, i.e., when the jumps over the mesh faces in the eigenvector approximation vanish in mean value. Then the discrete and broken gradient coincide, i.e., $\nabla_\theta = \nabla_h$ (see*

Section 5.2.3) and all the mean value jump terms of the form $h_e^{-\frac{1}{2}}\|\Pi_e^0[\![u_{ih}]\!]\|_e$ *of the present paper vanish, in particular in* (5.6.8a) *and in* (5.6.17). *Moreover,* (5.4.7) *and* (5.6.19a) *simplify respectively to*

$$\|\nabla_h(u_{ih}-s_{ih})\| \le (d+1)C_{\rm st}C_{\rm cont,bPF}\|\nabla_h(u_i-u_{ih})\|, \tag{5.6.22a}$$

$$\|u_{ih}-s_{ih}\| \le (d+1)C_{\rm st}C_{\rm cont,bPF}C_{\rm bF}\|\nabla_h(u_i-u_{ih})\|, \tag{5.6.22b}$$

see [105, Lemma 3.22 and Section 4.3.2]. This very favorable context arises namely for nonconforming and mixed finite elements, as we will see below in Sections 5.7.1 and 5.7.3.

Remark 5.6.7 (Jump-free estimators for the symmetric discontinuous Galerkin method)**.** *The jump terms in the estimator* $\overline{\alpha}_{ih}$ *given by* (5.6.8a) *also vanish when* $\theta = 1$*, which is typically the situation for the symmetric discontinuous Galerkin method of Section 5.7.2 below.*

Remark 5.6.8 (Alternative eigenvector reconstruction and vanishing jumps for the symmetric discontinuous Galerkin method)**.** *An alternative eigenvector reconstruction to that of Definition 5.3.3 is possible in two space dimensions following [105, Remark 3.11] when* $\left(\nabla_\theta u_{ih}, \begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\nabla\psi_{\mathbf{a}}\right)_{\omega_{\mathbf{a}}} = 0$ *for all* $\mathbf{a}\in\mathcal{V}_h$*. This is in particular satisfied for the symmetric variant of the discontinuous Galerkin method of Section 5.7.2 below. This alternative reconstruction remarkably yields*

$$\|\nabla_\theta(u_{ih}-s_{ih})\| \le (d+1)C_{\rm st}C_{\rm cont,P}\|\nabla_\theta(u_i-u_{ih})\|$$

in place of (5.6.18b), (5.4.7)*, and* (5.6.18c).
Here the constant $C_{\rm cont,P} := \max_{\mathbf{a}\in\mathcal{V}_h}\{1+C_{{\rm P},\omega_{\mathbf{a}}}h_{\omega_{\mathbf{a}}}\|\nabla\psi_{\mathbf{a}}\|_{\infty,\omega_{\mathbf{a}}}\}$ *with* $C_{{\rm P},\omega_{\mathbf{a}}}$ *given by* (5.4.2a) *where zero mean value condition* $(v,1)_{\omega_{\mathbf{a}}} = 0$ *is also imposed for boundary vertices. This is an equivalent of the bound* (5.6.22a)*, again without any jump terms. Then, all the principal estimators are efficient without any jump term, fully mimicking the situation of Remark 5.6.6.*

Remark 5.6.9 (Optimal efficiency and polynomial-degree robustness)**.** *Theorem 5.6.4 shows that both estimators* η_i *and* $\|\nabla_\theta(u_{ih}-\tilde{s}_{ih})\|$ *are equivalent to the eigenvector energy error* $\|\nabla_\theta(u_i-u_{ih})\|$ *for nonconforming finite elements, see Remarks 5.6.5 and 5.6.6. A similar case can arise for the symmetric discontinuous Galerkin method, see Remark 5.6.8. Taking into account that the size of our confidence interval for the i-th eigenvalue of Theorem 5.6.1 is* η_i^2*, this gives a fully optimal theory with in particular* polynomial-degree-robustness*. In the general case, the jumps in mean values of* u_{ih}*,* $\{\sum_{e\in\mathcal{E}_h} h_e^{-1}\|\Pi_e^0[\![u_{ih}]\!]\|_e^2\}^{\frac{1}{2}}$*, may be added to the error in the form* $\{\sum_{e\in\mathcal{E}_h} h_e^{-1}\|\Pi_e^0[\![u_i-u_{ih}]\!]\|_e^2\}^{\frac{1}{2}}$*, as typically done in discontinuous Galerkin methods, and similarly for the consistency terms, so as to still have equivalence between the eigenvector energy error and its estimate. Note, however, that the efficiency constant* C_i *in* (5.6.17) *contains the discrete relative gap* c_{ih} *or* $\tilde{c}_{ih}$ *of* (5.6.3) *in the form these latter are included as multiplicative factors in the complete estimator* η_i*, see Remark 5.6.2; only in Case C and when* $\|u_{ih}-s_{ih}\|$ *and consequently* $\|\Pi_0(u_{ih}-s_{ih})\|$ *decay as* $h^\delta\|\nabla_\theta u_{ih}+\boldsymbol{\sigma}_{ih}\|$*, the influence of these discrete relative gaps vanishes in the limit. Moreover, the factor* $\overline{C}_{ih}$ *from* (5.6.20c) *deteriorates the efficiency for increasing eigenvalues, except in our mixed finite elements setting of Section 5.7.3 where it does not appear.*

Remark 5.6.10 (Cheaper flux and potential reconstructions)**.** *The bound* (5.6.4) *for eigenvalues and the upper bound* (5.6.16) *for eigenvectors stay valid for cheaper (by one polynomial degree) flux and potential reconstructions, where* $\mathbf{V}_h^p(\omega_{\mathbf{a}})\times Q_h^p(\omega_{\mathbf{a}})$ *and* $\mathbb{P}_p(\mathcal{T}_{\mathbf{a}})\cap H_0^1(\omega_{\mathbf{a}})$ *are used in Section 5.3.2, instead of* $\mathbf{V}_h^{p+1}(\omega_{\mathbf{a}})\times Q_h^{p+1}(\omega_{\mathbf{a}})$ *and* $\mathbb{P}_{p+1}(\mathcal{T}_{\mathbf{a}})\cap H_0^1(\omega_{\mathbf{a}})$*. This is often completely sufficient in practice, albeit the theoretical polynomial degree robustness* (5.6.17) *may be lost.*

5.7 Application to common nonconforming numerical methods

We verify in this section the conditions of the application of our results to three common nonconforming numerical discretizations of the Laplace eigenvalue problem (5.2.1).

5.7.1 Nonconforming finite elements

Let V_h be spanned by functions v_h from $\mathbb{P}_p(\mathcal{T}_h)$, $p \geq 1$, such that $\langle [\![v_h]\!], q_h \rangle_e = 0$ for all $q_h \in \mathbb{P}_{p-1}(e)$ and all $e \in \mathcal{E}_h$. The nonconforming finite element method for problem (5.2.1) reads, cf. [82, 140, 141, 64, 253]: find $(u_{ih}, \lambda_{ih}) \in V_h \times \mathbb{R}^+$ with $(u_{ih}, u_{jh}) = \delta_{ij}$, $1 \leq i, j \leq \dim V_h$, such that

$$(\nabla_h u_{ih}, \nabla_h v_h) = \lambda_{ih}(u_{ih}, v_h) \qquad \forall v_h \in V_h; \tag{5.7.1}$$

the sign of the eigenvector u_{ih} is fixed by $(u_{ih}, \chi_i) > 0$. As the jump mean values in the space V_h are zero, $\nabla_\theta = \nabla_h$ follows from (5.2.5) (we can, e.g., take $\theta = 0$). Then definition (5.7.1) directly implies Assumption 5.3.1 (take $v_h = \psi_{\mathbf{a}} \in V_h$ in (5.7.1)). Thus, upon the verification/satisfaction of condition (5.6.6b) (in Case B) and of (5.6.8b) (in Case C), all the results of Theorems 5.6.1, 5.6.3, and 5.6.4 hold. We actually have clear eigenvector efficiency without jumps and consistency terms ($\lambda_{ih} = \|\nabla_h u_{ih}\|^2$ follows by taking $v_h = u_{ih}$ in (5.7.1)), see Remarks 5.6.5 and 5.6.6, and optimally convergent eigenvalue and eigenvector bounds, see Remark 5.6.9.

5.7.2 Discontinuous Galerkin finite elements

Set $V_h := \mathbb{P}_p(\mathcal{T}_h)$, $p \geq 1$, without any continuity requirement. The discontinuous Galerkin finite element method for problem (5.2.1), cf. [7, 113] and the references therein, reads: find $(u_{ih}, \lambda_{ih}) \in V_h \times \mathbb{R}^+$ with $(u_{ih}, u_{jh}) = \delta_{ij}$, $1 \leq i, j \leq \dim V_h$, such that

$$\begin{aligned} &\sum_{K \in \mathcal{T}_h} (\nabla_h u_{ih}, \nabla_h v_h)_K - \sum_{e \in \mathcal{E}_h} \{ \langle \{\!\{ \nabla_h u_{ih} \}\!\} \cdot \mathbf{n}_e, [\![v_h]\!] \rangle_e + \theta \langle \{\!\{ \nabla_h v_h \}\!\} \cdot \mathbf{n}_e, [\![u_{ih}]\!] \rangle_e \} \\ &+ \sum_{e \in \mathcal{E}_h} \langle \nu h_e^{-1} [\![u_{ih}]\!], [\![v_h]\!] \rangle_e = \lambda_{ih}(u_{ih}, v_h) \qquad \forall v_h \in V_h; \end{aligned} \tag{5.7.2}$$

the sign of u_{ih} is fixed by $(u_{ih}, \chi_i) > 0$. Here ν is a positive stabilization parameter and the parameter $\theta \in \{-1, 0, 1\}$ defines the discrete gradient (5.2.5) in Section 5.2.3 and corresponds respectively to the nonsymmetric, incomplete, and symmetric variants. The system matrix corresponding to (5.7.2) is only symmetric for $\theta = 1$. In the other cases, we tacitly assume that the i-th eigenvalue λ_{ih} that one computes is real. This typically happens for the first eigenvalue and more generally for all simple eigenvalues, cf. the numerical experiments in [7, Section 7.1.2].

With the concept of the discrete gradient (5.2.5), the orthogonality of Assumption 5.3.1 is immediately satisfied. Indeed, it is enough to take $v_h = \psi_{\mathbf{a}} \in V_h$ in (5.7.2) and take into account the facts that $\psi_{\mathbf{a}}$ has no jumps as well as that $\nabla \psi_{\mathbf{a}} \in [\mathbb{P}_0(\mathcal{T}_h)]^d \subset \mathbf{V}^0(\mathcal{T}_h)$. Thus all the results of Theorems 5.6.1, 5.6.3, and 5.6.4 hold upon the satisfaction of their assumptions. Recall also that 1) for $\theta = 0$, the broken ∇_h and discrete ∇_θ gradients coincide; 2) the jumps are here generally not of mean value zero, $\langle [\![u_{ih}]\!], 1 \rangle_e \neq 0$ for $e \in \mathcal{E}_h$, so that Remark 5.6.6 does not apply here; 3) the choice $\theta = 1$ leads to a remarkable situation where the jumps vanish from $\overline{\alpha_{ih}}$ given by (5.6.8a) and consequently from all three considered estimators η_i in Theorem 5.6.1, see Remark 5.6.7; 4) the choice $\theta = 1$ and the alternative eigenvector reconstruction of Remark 5.6.8 make the jumps vanish also from all the important parts in the efficiency bounds of Theorem 5.6.4.

5.7.3 Mixed finite elements

Let $\overline{\mathbf{V}}_h \times \overline{Q}_h$ be any pair of of the usual mixed finite element spaces, see [36, 214] and also Section 5.2.2 for the Raviart–Thomas–Nédélec case. The mixed finite element method for problem (5.2.1) looks for the triple $\overline{\boldsymbol{\sigma}}_{ih} \in \overline{\mathbf{V}}_h$, $\bar{u}_{ih} \in \overline{Q}_h$, and $\lambda_{ih} \in \mathbb{R}^+$ such that $(u_{ih}, u_{jh}) = \delta_{ij}$, $1 \le i, j \le \dim \overline{Q}_h$, and, cf. [190, 94, 29, 143] and the references therein,

$$(\overline{\boldsymbol{\sigma}}_{ih}, \mathbf{v}_h) - (\bar{u}_{ih}, \nabla{\cdot}\mathbf{v}_h) = 0 \qquad \forall \mathbf{v}_h \in \overline{\mathbf{V}}_h, \tag{5.7.3a}$$

$$(\nabla{\cdot}\overline{\boldsymbol{\sigma}}_{ih}, q_h) = \lambda_{ih}(\bar{u}_{ih}, q_h) \qquad \forall q_h \in \overline{Q}_h, \tag{5.7.3b}$$

where the sign of the i-th eigenvector that we are interested in is fixed by $(\bar{u}_{ih}, \chi_i) > 0$.

The low-order eigenvector approximation $\bar{u}_{ih}$ is typically elementwise postprocessed in mixed finite element methods. In particular, following Arnold and Brezzi [11], Arbogast and Chen [9], and Vohralík [242], there exists for each pair $\overline{\mathbf{V}}_h \times \overline{Q}_h$ a piecewise polynomial space M_h such that $u_{ih} \in M_h$ can be prescribed by

$$\Pi_{\overline{Q}_h(K)}(u_{ih}|_K) = \bar{u}_{ih}|_K \qquad \forall K \in \mathcal{T}_h, \tag{5.7.4a}$$

$$\Pi_{\overline{\mathbf{V}}_h(K)}((-\nabla_h u_{ih})|_K) = \overline{\boldsymbol{\sigma}}_{ih}|_K \qquad \forall K \in \mathcal{T}_h, \tag{5.7.4b}$$

where $\Pi_{\overline{Q}_h(K)}$ is the $L^2(K)$-orthogonal projection onto $\overline{Q}_h(K)$ and $\Pi_{\overline{\mathbf{V}}_h(K)}$ is the $[L^2(K)]^d$-orthogonal projection onto $\overline{\mathbf{V}}_h(K)$. Let p denote the polynomial degree of the approximation u_{ih} resulting from (5.7.4), i.e., $u_{ih} \in \mathbb{P}_p(\mathcal{T}_h)$ as throughout the paper. A remarkable fact is that (5.7.4) and (5.7.3a) imply

$$\langle [\![u_{ih}]\!], v_h \rangle_e = 0 \qquad \forall v_h \in \overline{\mathbf{V}}_h{\cdot}\mathbf{n}_e(e),\ \forall e \in \mathcal{E}_h,$$

so that in particular the zero mean-value condition, cf. Remark 5.6.6, is satisfied. Consequently, $\nabla_\theta = \nabla_h$, see (5.2.5) (and we can, e.g., take $\theta = 0$). The computed flux $\overline{\boldsymbol{\sigma}}_{ih}$ can typically serve directly as an equilibrated flux reconstruction in mixed finite elements, see [105, Section 4.4] and the references therein. However, in the present eigenvalue case, it only follows from (5.7.3b) that $\nabla{\cdot}\overline{\boldsymbol{\sigma}}_{ih} = \lambda_{ih}\bar{u}_{ih}$, and not $\nabla{\cdot}\overline{\boldsymbol{\sigma}}_{ih} = \lambda_{ih}u_{ih}$ as requested in the equilibrium property (5.3.1b) and necessary in the proof of the upper bound (5.4.5a). We can, however, postprocess elementwise the flux $\overline{\boldsymbol{\sigma}}_{ih}$ as well: choose a mixed space $\mathbf{V}_h^q$ with a sufficient polynomial degree q such that $M_h \subset \nabla{\cdot}\mathbf{V}_h^q$. Denoting by $\mathbf{n}_K$ the outward unit normal to K, define

$$\boldsymbol{\sigma}_{ih}|_K := \arg \min_{\substack{\mathbf{v}_h \in \mathbf{V}_h^q(K),\ \mathbf{v}_h{\cdot}\mathbf{n}_K = \overline{\boldsymbol{\sigma}}_{ih}{\cdot}\mathbf{n}_K \text{ on } \partial K \\ \nabla{\cdot}\mathbf{v}_h = \lambda_{ih}u_{ih}}} \|\overline{\boldsymbol{\sigma}}_{ih} - \mathbf{v}_h\|_K \qquad \forall K \in \mathcal{T}_h \tag{5.7.5}$$

instead of (5.3.3a) of Definition 5.3.2. The well-posedness of (5.7.5) follows from (5.7.3b) and (5.7.4a). Note that $\boldsymbol{\sigma}_{ih}$ is only a slight local elementwise modification of $\overline{\boldsymbol{\sigma}}_{ih}$, preserving the normal component while improving the divergence.

With the just described setting, all the eigenvalue results of Section 5.6 hold true in the following sense:

Corollary 5.7.1 (Eigenvalue bounds for mixed finite elements)**.** *Let $\underline{\lambda}_1, \overline{\lambda}_i, \underline{\lambda}_{i+1}, \overline{\lambda}_{i-1} > 0$ be the usual auxiliary bounds. Let (u_{ih}, λ_{ih}) be given by* (5.7.3)–(5.7.4)*. Construct s_{ih} from u_{ih} following Definition 5.3.3 and $\boldsymbol{\sigma}_{ih}$ by* (5.7.5)*. Then, the bounds* (5.6.4) *and* (5.6.10) *of respectively Theorems 5.6.1 and 5.6.3 hold true (in the three cases A, B, and C).*

Concerning the eigenvectors, the guaranteed upper bounds (5.4.5a) and consequently (5.6.16) do hold even if u_{ih} does not satisfy Assumption 5.3.1; the key is that $\nabla{\cdot}\boldsymbol{\sigma}_{ih} = \lambda_{ih}u_{ih}$ that we have arranged in (5.7.5). For the efficiency, recall first that $\nabla_\theta = \nabla_h$, so that (5.6.18b)

and (5.6.18c) are redundant here. Next, the bounds (5.4.7) and (5.6.19), or more precisely their improvements (5.6.22), only exploit the construction of s_{ih} from u_{ih} via Definition 5.3.3 and are thus also valid. Unfortunately, (5.4.5b) and consequently (5.6.20) do not hold in general, as $\boldsymbol{\sigma}_{ih}$ is not constructed from u_{ih} by Definition 5.3.2 but via (5.7.5). In order to restore fully optimal (guaranteed, efficient, and polynomial-degree robust) eigenvector error control, we proceed as in [105, Section 4.4], see also the references therein. Invoking the triangle inequality $\|\nabla u_i + \boldsymbol{\sigma}_{ih}\| \le \|\nabla_h(u_i - u_{ih})\| + \|\nabla_h u_{ih} + \boldsymbol{\sigma}_{ih}\|$, we have the following optimal simultaneous error control in $\nabla_h u_{ih}$ and $-\boldsymbol{\sigma}_{ih}$:

Corollary 5.7.2 (Eigenvector bounds for mixed finite elements)**.** *Let the assumptions of Corollary 5.7.1 be satisfied. Then, in the three cases A, B, and C of Theorem 5.6.1,*

$$\|\nabla_h(u_i - u_{ih})\| + \|\nabla u_i + \boldsymbol{\sigma}_{ih}\| \le 2(\eta_i + \|\nabla_h(u_{ih} - \tilde{s}_{ih})\|) + \|\nabla_h u_{ih} + \boldsymbol{\sigma}_{ih}\|.$$

This bound is efficient as (5.6.22) *holds together with*

$$\|\nabla_h u_{ih} + \boldsymbol{\sigma}_{ih}\| \le \|\nabla_h(u_i - u_{ih})\| + \|\nabla u_i + \boldsymbol{\sigma}_{ih}\|.$$

5.8 Numerical experiments

We now numerically illustrate our estimates on two test cases in $\mathbb{R}^2$, for the nonconforming finite element method of order $p = 1$ and the symmetric ($\theta = 1$) discontinuous Galerkin finite element method of order $p = 1$. We actually only study the first eigenpair; results for the higher eigenpairs are similar as in [52]. We use the cheaper Raviart–Thomas–Nédélec space of degree $p = 1$ for the flux equilibration instead of $p + 1$, as discussed in Remark 5.6.10. This still gives guaranteed bounds and we do not observe any asymptotic loss of efficiency. The implementation was done in the FreeFem++ code [132]. If the additional elliptic regularity for the corresponding source problem of Assumption 5.5.1 holds, so that Case C of Theorems 5.6.1 and 5.6.4 can be used, we observe that the last term of (5.6.8a) vanishes in the two considered numerical methods. We consider full $H^2(\Omega)$ regularity and use the constants $C_{\mathrm{S}} = 1$ and $\delta = 1$ given in Remark 5.5.2 and set C_{I} and $\bar{C}_{\mathrm{I}}$ following respectively Remarks 5.5.2 and 5.5.3.

5.8.1 Nonconforming finite element method

We test here the performance of the lowest-order ($p = 1$) nonconforming finite element method as described in Section 5.7.1.

Unit square

We start by testing the framework on a geometry where everything is known explicitly: the unit square $\Omega = (0, 1)^2$. The eigenvalues on a square of size H being $\pi^2(k^2 + l^2)/H^2$, $k, l \in \mathbb{N}^\star$, the first and second eigenvalues are $\lambda_1 = 2\pi^2$ and $\lambda_2 = 5\pi^2$, respectively. The first eigenfunction is given by $u_1(x, y) = 2\sin(\pi x)\sin(\pi y)$. We can here apply the refined elliptic regularity of Case C, since $d = 2$ and the domain is convex. The conditions (5.6.1) and (5.6.6b), (5.6.8b) respectively, are satisfied on all the meshes considered here, using $\underline{\lambda}_1 := 1.5\pi^2$, $\underline{\lambda}_2 := 4.5\pi^2$, and $\overline{\lambda}_1 := \|\nabla \tilde{s}_{1h}\|^2$ following (5.6.11) as the auxiliary bounds in Theorems 5.6.1 and 5.6.4.

Figure 5.1 illustrates the convergence of the energy error in the eigenfunction $\|\nabla_\theta(u_1 - u_{1h})\|$ and its upper bound $\eta_1 + \|\nabla_\theta(u_{1h} - \tilde{s}_{1h})\|$ for a sequence of uniform and structured meshes (left) and a sequence of unstructured quasi-uniform meshes (right). This test confirms that the convergence rate for the upper bound is the same as the one of the error in the approximation of the eigenvector.

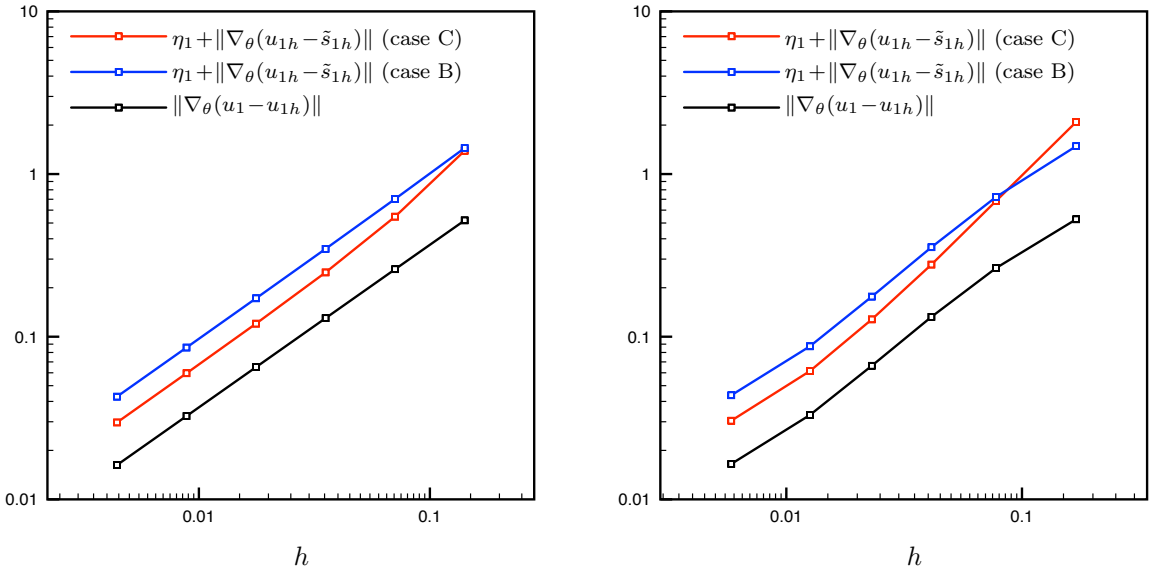

Figure 5.1 – [Unit square, nonconforming method] Error in the eigenvector approximation and its upper bound for the choice $\underline{\lambda}_1 = 1.5\pi^2$, $\underline{\lambda}_2 = 4.5\pi^2$; sequence of structured (left) and unstructured but quasi-uniform (right) meshes

N	h	ndof	λ_1	λ_{1h}	$\|\nabla\tilde{s}_{1h}\|^2-\eta_1^2$	$\|\nabla\tilde{s}_{1h}\|^2$	$E_{\lambda,\mathtt{rel}}$	$I^{\mathtt{ub}}_{u,\mathtt{eff}}$
10	0.1414	320	19.7392	19.6850	18.8966	19.8262	4.80e-02	2.68
20	0.0707	1240	19.7392	19.7257	19.6495	19.7616	5.69e-03	2.11
40	0.0354	4880	19.7392	19.7358	19.7246	19.7448	1.02e-03	1.91
80	0.0177	19360	19.7392	19.7384	19.7361	19.7406	2.29e-04	1.85
160	0.0088	77120	19.7392	19.7390	19.7385	19.7396	5.53e-05	1.83
320	0.0044	307840	19.7392	19.7392	19.7390	19.7393	1.37e-05	1.83

Table 5.1 – [Structured mesh, unit square, nonconforming method, case C] Lower and upper bounds of the exact eigenvalue λ_1, the relative eigenvalue error, and the eigenvector effectivity index; case $\underline{\lambda}_1 = 1.5\pi^2$, $\underline{\lambda}_2 = 4.5\pi^2$

We present in Tables 5.1 and 5.2 precise values of the lower and upper bounds $\|\nabla\tilde{s}_{1h}\|^2 - \eta_1^2 \le \lambda_1 \le \|\nabla\tilde{s}_{1h}\|^2$ on the exact eigenvalue λ_1, the relative error and the effectivity index of the upper bound $\|\nabla_\theta(u_1 - u_{1h})\| \le \eta_1 + \|\nabla_\theta(u_{1h} - \tilde{s}_{1h})\|$, given respectively by

$$E_{\lambda,\mathtt{rel}} := 2\frac{\|\nabla\tilde{s}_{1h}\|^2 - (\|\nabla\tilde{s}_{1h}\|^2 - \eta_1^2)}{\|\nabla\tilde{s}_{1h}\|^2 + (\|\nabla\tilde{s}_{1h}\|^2 - \eta_1^2)} = \frac{2\,\eta_1^2}{2\,\|\nabla\tilde{s}_{1h}\|^2 - \eta_1^2}, \tag{5.8.1a}$$

$$I^{\mathtt{ub}}_{u,\mathtt{eff}} := \frac{\eta_1 + \|\nabla_\theta(u_{1h} - \tilde{s}_{1h})\|}{\|\nabla_\theta(u_1 - u_{1h})\|}. \tag{5.8.1b}$$

We observe rather convincing results.

L-shaped domain

We next consider the L-shaped domain $\Omega := (-1,1)^2 \setminus [0,1]\times[-1,0]$, where $\lambda_1 \approx 9.6397238440$ is known to high accuracy [234]. Including Ω into the square $\Omega^+ := (-1,1)^2$, cf. [52, Remark 5.4], we take $\underline{\lambda}_1 := \lambda_1(\Omega^+) = \pi^2/2$ and $\underline{\lambda}_2 := 15.1753$ from Table 1 of [176] in Theorems 5.6.1 and 5.6.4.

We first consider a sequence of unstructured quasi-uniform meshes, with N elements partitioning the edges of Ω of length 2 and $N/2$ elements the edges of length 1. Figure 5.2

N	h	ndof	λ_1	λ_{1h}	$\|\nabla\tilde{s}_{1h}\|^2-\eta_1^2$	$\|\nabla\tilde{s}_{1h}\|^2$	$E_{\lambda,\mathtt{rel}}$	$I_{u,\mathtt{eff}}^{\mathtt{ub}}$
10	0.1698	386	19.7392	19.6556	17.1037	19.8250	1.47e-01	3.97
20	0.0776	1486	19.7392	19.7157	19.5482	19.7604	1.08e-02	2.58
40	0.0413	5762	19.7392	19.7335	19.7167	19.7448	1.42e-03	2.10
80	0.0230	22789	19.7392	19.7377	19.7353	19.7406	2.66e-04	1.93
160	0.0126	91355	19.7392	19.7389	19.7384	19.7396	5.89e-05	1.86
320	0.0058	366520	19.7392	19.7391	19.7390	19.7393	1.41e-05	1.84

Table 5.2 – [Untructured mesh, unit square, nonconforming method, case C] Lower and upper bounds of the exact eigenvalue λ_1, the relative eigenvalue error, and the eigenvector effectivity index; case $\underline{\lambda}_1 = 1.5\pi^2$, $\underline{\lambda}_2 = 4.5\pi^2$

N	h	ndof	λ_1	λ_{1h}	$\|\nabla\tilde{s}_{1h}\|^2-\eta_1^2$	$\|\nabla\tilde{s}_{1h}\|^2$	$E_{\lambda,\mathtt{rel}}$	$I_{u,\mathtt{eff}}^{\mathtt{ub}}$
10	0.3041	266	9.6397	9.2966	-4.1909	9.7861	–	6.02
20	0.1670	1069	9.6397	9.5155	7.8895	9.6926	2.05e-01	4.19
40	0.0839	4148	9.6397	9.5933	9.0782	9.6578	6.19e-02	4.12
80	0.0459	16699	9.6397	9.6227	9.4514	9.6459	2.04e-02	4.09
160	0.0234	64991	9.6397	9.6331	9.5703	9.6420	7.46e-03	4.08
320	0.0125	259147	9.6397	9.6372	9.6138	9.6406	2.78e-03	4.07

Table 5.3 – [Untructured mesh, L-shaped domain, nonconforming method] Lower and upper bounds of the exact eigenvalue λ_1, the relative eigenvalue error, and the eigenvector effectivity index; case $\underline{\lambda}_1 = \pi^2/2$ and $\underline{\lambda}_2 = 15.1753$

(left) illustrates the convergence of the energy error $\|\nabla_\theta(u_1 - u_{1h})\|$ and its upper bound $\eta_1 + \|\nabla_\theta(u_{1h} - \tilde{s}_{1h})\|$. Details and eigenvalue convergence results are presented in Table 5.3. All the theoretical results are nicely confirmed.

We finally test adaptive refinement using the local character of the eigenvector estimator for each $K \in \mathcal{T}_h$ given by

$$\left(1 + 2(\lambda_{1h} + \|\nabla\tilde{s}_{1h}\|^2)\underline{\lambda}_2^{-1}\left(1 - \frac{\lambda_{1h}}{\underline{\lambda}_2}\right)^{-2}\right)\frac{1}{\|s_{1h}\|^2}$$
$$\left(\frac{\lambda_{1h}^2}{\underline{\lambda}_1}\|u_{1h} - s_{1h}\|_K^2 + \|\nabla s_{1h} + \boldsymbol{\sigma}_{1h}\|_K^2\right) + \|\nabla_\theta(u_{1h} - \tilde{s}_{1h})\|_K^2,$$

in case A and

$$\left(1 - \frac{\lambda_{1h}}{\underline{\lambda}_2}\right)^{-2}\left(1 - \frac{\overline{\alpha}_{1h}^2}{4}\right)^{-1}\frac{1}{\|s_{1h}\|^2}\left(\frac{\lambda_{1h}^2}{\underline{\lambda}_1}\|u_{1h} - s_{1h}\|_K^2 + \|\nabla s_{1h} + \boldsymbol{\sigma}_{1h}\|_K^2\right)$$
$$+ \|\nabla_\theta(u_{1h} - \tilde{s}_{1h})\|_K^2,$$

in case B of Theorems 5.6.1 and 5.6.4. We employ the Dörfler marking [92] with $\theta = 0.6$ and the newest vertex bisection mesh refinement. The same lower bounds $\underline{\lambda}_1 = \pi^2/2$ and $\underline{\lambda}_2 = 15.1753$ as for the uniform refinement have been used for the auxiliary bounds. Figure 5.2 (right) illustrates the error in the eigenvector and its bound using (5.6.16). The optimal convergence rate is indicated by dashed lines. The initial mesh is structured with 47 degrees of freedom and the conditions (5.6.1) and (5.6.6b) are both satisfied starting from 296 degrees of freedom. The transition from Case A to Case B in Theorem 5.6.1 is marked by a dotted line. Table 5.4 then presents more details of the adaptive procedure, which in particular leads to quite good effectivity indices.

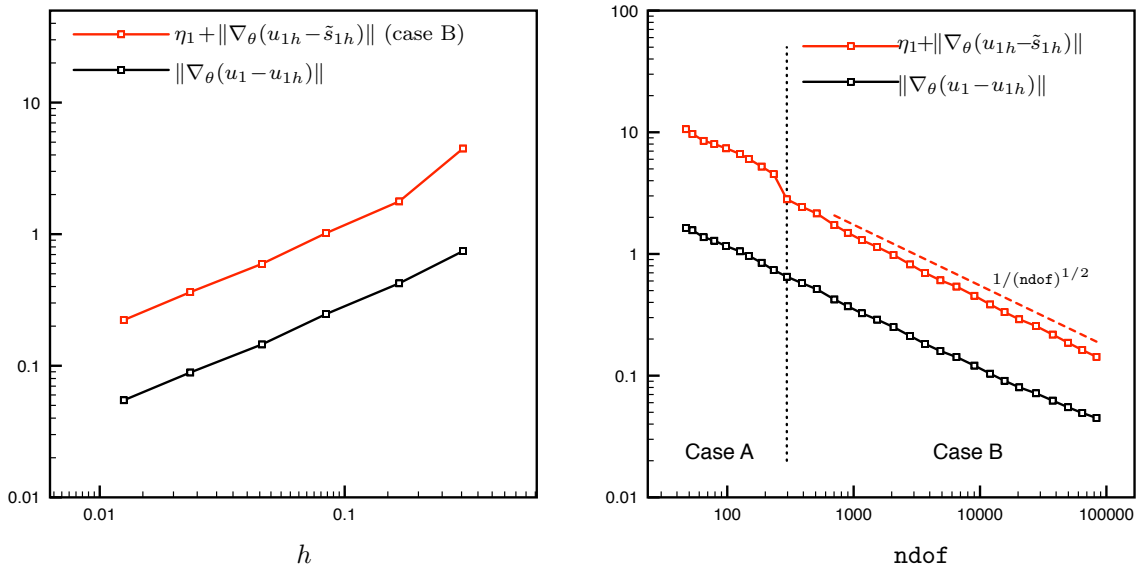

Figure 5.2 – [Unstructured and adaptive mesh refinement, L-shaped domain, nonconforming method] Error in the eigenvector and its upper bound for a quasi-uniform refinement (left) and adaptive refinement (right).

Level	ndof	λ_1	λ_{1h}	$\|\nabla\tilde{s}_{1h}\|^2-\eta_1^2$	$\|\nabla\tilde{s}_{1h}\|^2$	$E_{\lambda,\mathtt{rel}}$	$I^{\mathtt{ub}}_{u,\mathtt{eff}}$
5	98	9.6397	8.9699	-29.6187	9.9072	–	6.36
10	296	9.6397	9.4403	4.8193	9.7445	6.76e-01	4.32
15	1161	9.6397	9.5868	8.6628	9.6646	1.09e-01	3.99
20	4860	9.6397	9.6275	9.4310	9.6457	2.25e-02	3.81
25	20429	9.6397	9.6369	9.5925	9.6411	5.06e-03	3.62
30	83472	9.6397	9.6390	9.6284	9.6401	1.21e-03	3.18

Table 5.4 – [Adaptive mesh refinement, L-shaped domain, nonconforming method] Lower and upper bounds of the exact eigenvalue λ_1, the relative eigenvalue error, and the eigenvector effectivity index; case $\underline{\lambda}_1 = \pi^2/2$ and $\underline{\lambda}_2 = 15.1753$

5.8.2 Discontinuous Galerkin finite element method

In order to test the framework on another method, we have taken the symmetric version ($\theta = 1$) of the discontinuous Galerkin finite element method as presented in Section 5.7.2, using piecewise affine basis functions ($p = 1$) and the penalty parameter $\nu = 10$.

Unit square

We consider again first the case of the unit square $\Omega = (0,1)^2$. The test case and the constants used are the same as presented in Section 5.8.1.

Figure 5.3 illustrates the convergence of the energy error in the eigenfunction $\|\nabla_\theta(u_1-u_{1h})\|$ and its guaranteed and computable a posteriori estimate $\eta_1 + \|\nabla_\theta(u_{1h} - \tilde{s}_{1h})\|$ for a sequence of uniform and structured meshes (left) and a sequence of unstructured quasi-uniform meshes (right). As the auxiliary eigenvalue lower bounds, we have taken again $\underline{\lambda}_1 = 1.5\pi^2$ and $\underline{\lambda}_2 = 4.5\pi^2$. This test in particular confirms that the convergence rate of our estimate is the same as the one of the error. Tables 5.5 and 5.6 reveal again more details on our estimates.

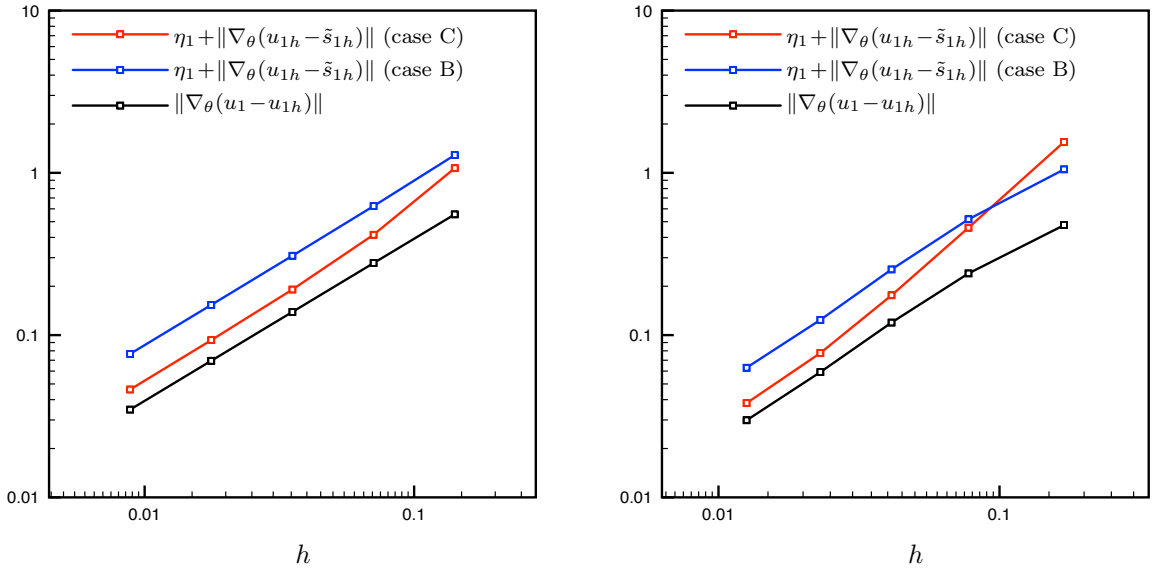

Figure 5.3 – [Unit square, discontinuous Galerkin method] Error in the eigenvector approximation and its upper bound for the choice $\underline{\lambda}_1 = 1.5\pi^2$, $\underline{\lambda}_2 = 4.5\pi^2$; sequence of structured (left) and unstructured but quasi-uniform (right) meshes

N	h	ndof	λ_1	λ_{1h}	$\|\nabla\tilde{s}_{1h}\|^2-\eta_1^2$	$\|\nabla\tilde{s}_{1h}\|^2$	$E_{\lambda,\mathtt{rel}}$	$I^{\mathtt{ub}}_{u,\mathtt{eff}}$
10	0.1414	600	19.7392	20.0333	19.1803	20.0101	4.23e-02	1.93
20	0.0707	2400	19.7392	19.8169	19.6907	19.8099	6.03e-03	1.50
40	0.0354	9600	19.7392	19.7591	19.7324	19.7572	1.26e-03	1.37
80	0.0177	38400	19.7392	19.7442	19.7378	19.7438	2.99e-04	1.34
160	0.0088	153600	19.7392	19.7405	19.7389	19.7403	7.09e-05	1.33

Table 5.5 – [Structured mesh, unit square, discontinuous Galerkin method, case C] Lower and upper bounds of the exact eigenvalue λ_1, the relative eigenvalue error, and the eigenvector effectivity index; case $\underline{\lambda}_1 = 1.5\pi^2$, $\underline{\lambda}_2 = 4.5\pi^2$

L-shaped domain

We consider again as for the nonconforming method the L-shaped domain $\Omega := (-1,1)^2 \setminus [0,1]\times[-1,0]$ as a second test problem. As motivated in Section 5.8.1, we take $\underline{\lambda}_1 := \pi^2/2$ and $\underline{\lambda}_2 = 15.1753$ in Theorems 5.6.1 and 5.6.4.

Figure 5.4 (left) illustrates the convergence of the energy error $\|\nabla_\theta(u_1-u_{1h})\|$ and its upper bound $\eta_1+\|\nabla_\theta(u_{1h}-\tilde{s}_{1h})\|$. Details are presented in Table 5.7. Again, all the theoretical results are nicely confirmed, with in particular excellent effectivity indices.

We finally test adaptive refinement as outlined in Section 5.8.1. Figure 5.4 (right) illustrates the error in the eigenvector and its bound using (5.6.16). The optimal convergence rate is indicated by dashed lines. The initial mesh is structured with 47 degrees of freedom and the conditions (5.6.1) and (5.6.6b) are all satisfied starting from 591 degrees of freedom. The transition from Case A to Case B in Theorem 5.6.1 is marked by a dotted line. Table 5.8 then presents more details of the adaptive procedure.

N	h	ndof	λ_1	λ_{1h}	$\|\nabla\tilde{s}_{1h}\|^2-\eta_1^2$	$\|\nabla\tilde{s}_{1h}\|^2$	$E_{\lambda,\mathrm{rel}}$	$I_{u,\mathrm{eff}}^{\mathrm{ub}}$
10	0.1698	732	19.7392	19.9432	17.8788	19.9501	1.10e-01	3.26
20	0.0776	2892	19.7392	19.7928	19.6264	19.7939	8.50e-03	1.91
40	0.0413	11364	19.7392	19.7526	19.7295	19.7529	1.18e-03	1.47
80	0.0230	45258	19.7392	19.7425	19.7381	19.7426	2.28e-04	1.31
160	0.0126	182070	19.7392	19.7400	19.7390	19.7401	5.35e-05	1.28

Table 5.6 – [Untructured mesh, unit square, discontinuous Galerkin method, case C] Lower and upper bounds of the exact eigenvalue λ_1, the relative eigenvalue error, and the eigenvector effectivity index; case $\underline{\lambda}_1 = 1.5\pi^2$, $\underline{\lambda}_2 = 4.5\pi^2$

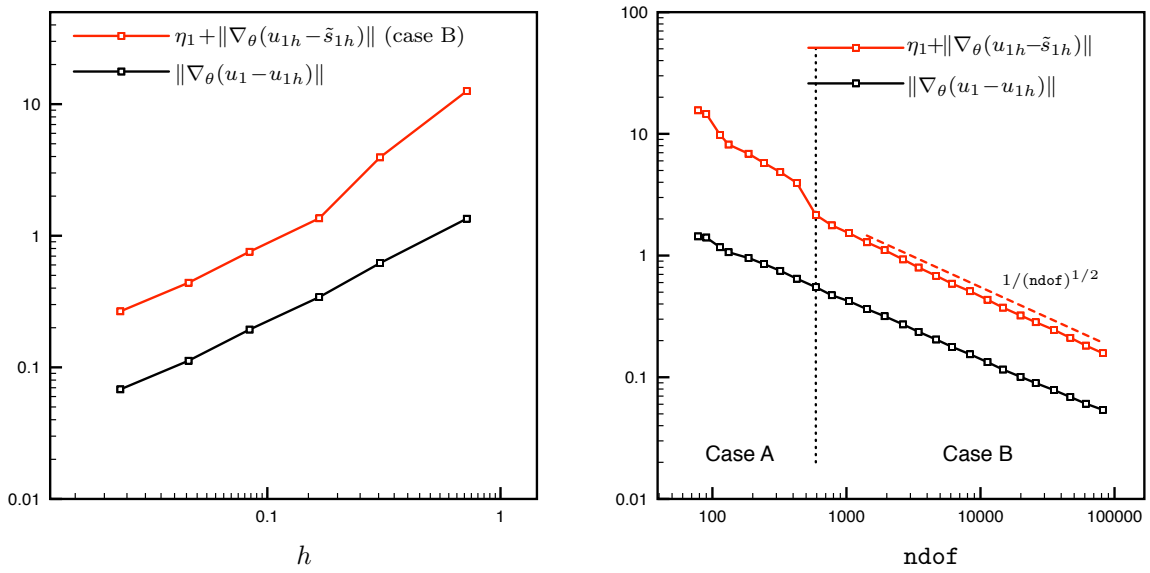

Figure 5.4 – [Unstructured and adaptive mesh refinement, L-shaped domain, discontinuous Galerkin method] Error in the eigenvector and its upper bound for a quasi-uniform refinement (left) and adaptive refinement (right)

5.9 Concluding remarks

The motivation of the present paper was to develop a general theory of eigenvalue and eigenvector a posteriori error estimates, enabling to take into account basically any numerical method. This in particular means that we need to admit the violation of the constraints $u_{ih} \in H_0^1(\Omega)$, $\|u_{ih}\| = 1$, $\|\nabla_\theta u_{ih}\|^2 = \lambda_{ih}$, and $\lambda_{ih} \geq \lambda_i$. Our bounds from Section 5.6 achieve this and we have seen in Section 5.7 that three common nonconforming numerical methods fit perfectly the framework. Moreover, typically, not all the above constraints are violated. Then parts of the results of Section 5.6 simplify importantly.

We have focused here for simplicity on the treatment of the case where the underlying algebraic eigenvalue solvers are exact, so that the present Assumption 5.3.1 can be satisfied. The framework is, however, built rich enough to take into account inexact solvers, following [197, 104] and the references therein, as we have demonstrated it in [52]. The resulting estimates are then valid on an arbitrary eigenvalue iterative solver step, enable to distinguish the different error components, and yield (local) adaptive stopping criteria. A preliminary example in the context of the Gross–Pitaevskii nonlinear eigenvalue problem is given in [49].

The approximation polynomial degree p was considered fixed here and we have only treated the case of matching simplicial meshes. Extension to variable polynomial degree and nonmatch-

N	h	ndof	λ_1	λ_{1h}	$\|\nabla\tilde{s}_{1h}\|^2-\eta_1^2$	$\|\nabla\tilde{s}_{1h}\|^2$	$E_{\lambda,\mathtt{rel}}$	$I_{u,\mathtt{eff}}^{\mathtt{ub}}$
5	0.7165	90	9.6397	10.7897	-128.5909	11.0700	–	9.32
10	0.3041	492	9.6397	9.9085	-3.4330	9.9928	–	6.36
20	0.1670	2058	9.6397	9.7044	8.3596	9.7448	1.53e-01	3.97
40	0.0839	8136	9.6397	9.6576	9.2512	9.6729	4.46e-02	3.90
80	0.0459	33078	9.6397	9.6447	9.5110	9.6506	1.46e-02	3.92
160	0.0234	129342	9.6397	9.6413	9.5929	9.6436	5.27e-03	3.92

Table 5.7 – [Untructured mesh, L-shaped domain, discontinuous Galerkin method] Lower and upper bounds of the exact eigenvalue λ_1, the relative eigenvalue error, and the eigenvector effectivity index; case $\underline{\lambda}_1=\pi^2/2$ and $\underline{\lambda}_2=15.1753$

Level	ndof	λ_1	λ_{1h}	$\|\nabla\tilde{s}_{1h}\|^2-\eta_1^2$	$\|\nabla\tilde{s}_{1h}\|^2$	$E_{\lambda,\mathtt{rel}}$	$I_{u,\mathtt{eff}}^{\mathtt{ub}}$
5	186	9.6397	10.2136	-30.6026	10.3629	–	7.19
10	777	9.6397	9.8154	7.2388	9.8388	3.04e-01	3.75
15	3453	9.6397	9.6865	9.1572	9.6902	5.66e-02	3.38
20	14706	9.6397	9.6509	9.5335	9.6517	1.23e-02	3.23
25	61137	9.6397	9.6425	9.6144	9.6426	2.93e-03	3.00

Table 5.8 – [Adaptive mesh refinement, L-shaped domain, discontinuous Galerkin method] Lower and upper bounds of the exact eigenvalue λ_1, the relative eigenvalue error, and the eigenvector effectivity index; case $\underline{\lambda}_1=\pi^2/2$ and $\underline{\lambda}_2=15.1753$

ing simplicial and quadrilateral meshes is straightforward following [91], where also corresponding *hp* (mesh and polynomial degree) adaptive refinement strategies are developed. It should be rather easy to generalize them to the present eigenvalue setting.

5.10 Appendix

The current analysis was presented for the Laplace operator of (5.1.1). The generic equivalences can, however, be extended to a larger class of operators that we show in part 5.10.1 of this appendix, for a conforming approximation. We next complement in part 5.10.2 the estimate of Theorem 5.6.3 by a further possible improvement of the first eigenvalue upper bound.

5.10.1 Extension to a generic operator

We formulate here the results of [52, Theorems 3.4 and 3.5] for conforming approximations and any bounded-below self-adjoint operator with compact resolvent, see, e.g., Helffer [134]. This comprises for example the operator $A:=-\Delta+w$ with domain $D(A):=\{v\in H_0^1(\Omega);\Delta v\in L^2(\Omega)\}$, which is self-adjoint on $L^2(\Omega)$ whenever $w\in L^\infty(\Omega)$. It appears that only the operator considered $(-\Delta)$ and the norms ($\|\cdot\|$, $\|\nabla\cdot\|$, and $\|\cdot\|_{-1}$) need to be changed.

Let $\mathcal{H}$ be a separable Hilbert space endowed with a scalar product denoted by $(\cdot,\cdot)_{\mathcal{H}}$. Now let A be a bounded-below self-adjoint operator on $\mathcal{H}$ with domain $D(A)$ and compact resolvent. There exists a non-decreasing sequence of real numbers $(\lambda_k)_{k\ge1}$ such that $\lambda_k\to\infty$ and an orthonormal basis $(u_k)_{k\ge1}$ of $\mathcal{H}$ consisting of vectors of $D(A)$ such that

$$\forall k\ge1,\quad A\,u_k=\lambda_k u_k.$$

Making the additional assumption that the k-th eigenvalue of A is simple, that is $\lambda_{k-1}<\lambda_k<\lambda_{k+1}$, the k-th eigenvector is unique up to the sign. Up to shifting the operator A by a constant $c\in\mathbb{R}^+$ such that $c+A$ is a positive definite operator, we can suppose that A is a positive definite operator, in which case $(\lambda_k)_{k\ge1}$ is a sequence of positive numbers. This

enables to define an operator $A^{\frac{1}{2}}$ analogous to the operator $|\nabla|$ in the previous case (recall that $\||\nabla v|\| = \|\nabla v\|$ for $v \in H^1(\Omega)$) by its domain

$$D(A^{\frac{1}{2}}) := \left\{ v \in \mathcal{H}; \quad \sum_{k\geq 1} \lambda_k |(v, u_k)_{\mathcal{H}}|^2 < +\infty \right\}$$

and its expression

$$A^{\frac{1}{2}} : v \in D(A^{\frac{1}{2}}) \mapsto \sum_{k\geq 1} \sqrt{\lambda_k}(v, u_k)_{\mathcal{H}} u_k.$$

Replace now $-\Delta$ by A; for the norms, the scalar product $(\cdot,\cdot)_{\mathcal{H}}$ of the Hilbert space $\mathcal{H}$ substitutes the L^2 scalar product $(\cdot,\cdot)$, and naturally the norm of $\|\cdot\|_{\mathcal{H}}$ replaces the L^2-norm $\|\cdot\|$. The energy norm $\|\nabla\cdot\|$ is changed into $\|A^{\frac{1}{2}}\cdot\|_{\mathcal{H}}$, and the duality pairing $\langle\cdot,\cdot\rangle_{V',V}$ becomes $\langle\cdot,\cdot\rangle_{D(A^{\frac{1}{2}})',D(A^{\frac{1}{2}})}$.

Let $(w_i, \lambda_{ih}) \in D(A^{\frac{1}{2}}) \times \mathbb{R}^+$ with $\|w_i\|_{\mathcal{H}} = 1$ and $(w_i, \chi_i)_{\mathcal{H}} > 0$ be given, for $\chi_i \in \mathcal{H}$, $i \geq 1$ fixed. Its residual $\mathrm{Res}_\theta(w_i, \lambda_{ih}) \in D(A^{\frac{1}{2}})'$ is now defined by

$$\langle \mathrm{Res}_\theta(w_i, \lambda_{ih}), v\rangle_{D(A^{\frac{1}{2}})',D(A^{\frac{1}{2}})} := \lambda_{ih}(w_i, v)_{\mathcal{H}} - (A^{\frac{1}{2}} w_i, A^{\frac{1}{2}} v)_{\mathcal{H}} \qquad \forall v \in D(A^{\frac{1}{2}}),$$

with the dual norm

$$\|\mathrm{Res}_\theta(w_i, \lambda_{ih})\|_{D(A^{\frac{1}{2}})'} := \sup_{v\in D(A^{\frac{1}{2}}),\, \|A^{\frac{1}{2}} v\|_{\mathcal{H}}=1} \langle \mathrm{Res}_\theta(w_i, \lambda_{ih}), v\rangle_{D(A^{\frac{1}{2}})',D(A^{\frac{1}{2}})}.$$

The Riesz representation of the residual $\boldsymbol{\imath}_{w_i} \in D(A^{\frac{1}{2}})$ is given by

$$(A^{\frac{1}{2}} \boldsymbol{\imath}_{w_i}, A^{\frac{1}{2}} v)_{\mathcal{H}} = \langle \mathrm{Res}_\theta(w_i, \lambda_{ih}), v\rangle_{D(A^{\frac{1}{2}})',D(A^{\frac{1}{2}})} \qquad \forall v \in D(A^{\frac{1}{2}}).$$

Let

$$\lambda_{i-1} < \lambda_{ih} \quad \text{when } i > 1, \quad \lambda_{ih} < \lambda_{i+1}, \tag{5.10.1}$$

and

$$\alpha_{ih} := \sqrt{2} C_{ih}^{-\frac{1}{2}} \|\boldsymbol{\imath}_{w_i}\|_{\mathcal{H}} \leq \|\chi_i\|_{\mathcal{H}}^{-1} (w_i, \chi_i)_{\mathcal{H}}, \tag{5.10.2}$$

where

$$C_{ih} := \min\left\{ \left(1 - \frac{\lambda_{ih}}{\lambda_{i-1}}\right)^2, \left(1 - \frac{\lambda_{ih}}{\lambda_{i+1}}\right)^2 \right\}.$$

The generalizations of [52, Theorems 3.4 and 3.5] then read:

Theorem 5.10.1 (Eigenvalue bounds)**.** *Let $(w_i, \lambda_{ih}) \in D(A^{\frac{1}{2}}) \times \mathbb{R}^+$ with $\|w_i\|_{\mathcal{H}} = 1$ and $(w_i, \chi_i)_{\mathcal{H}} > 0$, $i \geq 1$. Let assumptions* (5.10.1) *and* (5.10.2) *be satisfied. Then*

$$\|A^{\frac{1}{2}}(u_i - w_i)\|_{\mathcal{H}}^2 - \lambda_i \alpha_{ih}^2 \leq \|A^{\frac{1}{2}} w_i\|_{\mathcal{H}}^2 - \lambda_i \leq \|A^{\frac{1}{2}}(u_i - w_i)\|_{\mathcal{H}}^2. \tag{5.10.3a}$$

If, moreover $\alpha_{1h} \leq \sqrt{2}$, then, for $i = 1$,

$$\frac{1}{2}\left(1 - \frac{\lambda_1}{\lambda_2}\right)\left(1 - \frac{\alpha_{1h}^2}{4}\right) \|A^{\frac{1}{2}}(u_1 - w_1)\|_{\mathcal{H}}^2 \leq \|A^{\frac{1}{2}} w_1\|_{\mathcal{H}}^2 - \lambda_1. \tag{5.10.3b}$$

Let

$$\overline{C}_{ih} := 1 \text{ if } i = 1, \quad \overline{C}_{ih} := \max\left\{\left(\frac{\lambda_{ih}}{\lambda_1} - 1\right)^2, 1\right\} \text{ if } i > 1$$

and

$$\gamma_{ih} := \begin{cases} \|A^{\frac{1}{2}}(u_i - w_i)\|_{\mathcal{H}}^2 & \text{if } \lambda_i \le \|A^{\frac{1}{2}}(w_i)\|_{\mathcal{H}}^2 \text{ is known to hold,} \\ \max\{\|A^{\frac{1}{2}}(u_i - w_i)\|_{\mathcal{H}}^2, \lambda_i \alpha_{ih}^2\} & \text{otherwise.} \end{cases} \tag{5.10.4}$$

Then we also have:

Theorem 5.10.2 (Eigenvector bounds)**.** *Let the assumptions of Theorem 5.10.1 be satisfied. Then*

$$\|A^{\frac{1}{2}}(u_i - w_i)\|_{\mathcal{H}}^2 \le \|\mathrm{Res}_\theta(w_i, \lambda_{ih})\|_{D(A^{\frac{1}{2}})'}^2 + (\lambda_{ih} + \lambda_i)\alpha_{ih}^2, \tag{5.10.5a}$$

$$\|\mathrm{Res}_\theta(w_i, \lambda_{ih})\|_{D(A^{\frac{1}{2}})'}^2 \le \frac{\left(\left|\lambda_{ih} - \|A^{\frac{1}{2}} w_i\|_{\mathcal{H}}^2\right| + \gamma_{ih}\right)^2}{\lambda_i} + \overline{C}_{ih}\|A^{\frac{1}{2}}(u_i - w_i)\|_{\mathcal{H}}^2. \tag{5.10.5b}$$

If, moreover $\alpha_{ih}^2 \le 2\frac{\lambda_1}{\lambda_i}$*, then*

$$\|A^{\frac{1}{2}}(u_i - w_i)\|_{\mathcal{H}}^2 \le C_{ih}^{-1}\left(1 - \frac{\lambda_i}{\lambda_1}\frac{\alpha_{ih}^2}{4}\right)^{-1} \|\mathrm{Res}_\theta(w_i, \lambda_{ih})\|_{D(A^{\frac{1}{2}})'}^2.$$

5.10.2 Further improvement of the first eigenvalue upper bound

In [52, Theorem 5.2], a further improvement of the eigenvalue upper bounds of Theorem 5.6.3 was possible. We now extend it to the present setting, for the first eigenvalue.

We first need to generalize the conforming local residual lifting from [52, Section 4.3] to the present setting. Let for each vertex $\mathbf{a} \in \mathcal{V}_h$, $X_h^{\mathbf{a}}$ be an arbitrary finite-dimensional subspace of the space $H_*^1(\omega_{\mathbf{a}})$ from (5.4.1). Typically, $X_h^{\mathbf{a}} := \mathbb{P}_{p+1}(\mathcal{T}_{\mathbf{a}}) \cap H_*^1(\omega_{\mathbf{a}})$, similarly as in Section 5.3.2. We will now solve *homogeneous local Neumann* (Neumann–Dirichlet close to the boundary) *problems* on the patches $\omega_{\mathbf{a}}$ via conforming primal counterparts of problems (5.3.3a):

Definition 5.10.3 (Conforming local Neumann problems)**.** *For each* $\mathbf{a} \in \mathcal{V}_h$*, define* $r_{1h}^{\mathbf{a}} \in X_h^{\mathbf{a}}$ *by*

$$(\nabla r_{1h}^{\mathbf{a}}, \nabla v_h)_{\omega_{\mathbf{a}}} = \langle \mathrm{Res}_\theta(s_{1h}, \lambda_{1h}), \psi_{\mathbf{a}} v_h\rangle_{V',V} \qquad \forall v_h \in X_h^{\mathbf{a}}. \tag{5.10.6}$$

Then set

$$r_{1h} := \sum_{\mathbf{a}\in\mathcal{V}_h} \psi_{\mathbf{a}} r_{1h}^{\mathbf{a}} \in V.$$

The functions $r_{1h}^{\mathbf{a}}$ are discrete Riesz representations of the local residual of the pair (s_{1h}, λ_{1h}) with hat-weighted test functions. Note that the right-hand side in (5.10.6) does not necessarily satisfy the usually required Neumann compatibility condition $(\psi_{\mathbf{a}}\lambda_{1h}s_{1h} - \nabla s_{1h}\cdot\nabla\psi_{\mathbf{a}}, 1)_{\omega_{\mathbf{a}}} = 0$ for $\mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}$, so that (5.10.6) cannot hold for a constant function $v_h = 1$ on $\omega_{\mathbf{a}}$. Assumption 5.3.1 is in particular not required for s_{1h}; this does not influence the existence and uniqueness of $r_{1h}^{\mathbf{a}}$ (the system matrix in (5.10.6) is regular). Note also that $r_{1h}^{\mathbf{a}} \notin V$ (when extended by zero outside of $\omega_{\mathbf{a}}$) but $\psi_{\mathbf{a}} r_{1h}^{\mathbf{a}} \in H_0^1(\omega_{\mathbf{a}})$, whence the sum r_{1h} belongs to V. For this construction, we have:

Lemma 5.10.4 (Lower dual residual bound)**.** *Let* $(u_{1h}, \lambda_{1h}) \in \mathbb{P}_p(\mathcal{T}_h) \times \mathbb{R}^+$ *be arbitrary. Construct* s_{1h} *by Definition 5.3.3 and* r_{1h} *by Definition 5.10.3. Then*

$$\frac{\langle \mathrm{Res}_\theta(s_{1h}, \lambda_{1h}), r_{1h}\rangle_{V',V}}{\|\nabla r_{1h}\|} \le \|\mathrm{Res}_\theta(s_{1h}, \lambda_{1h})\|_{-1}.$$

Proof. The proof is trivial from (5.2.7b) and from the fact that $r_{1h} \in V$ for Definition 5.10.3. Importantly, this bound is positive, see [197, proof of Theorem 2]. □ □

Equipped with these tools, we can now hopefully improve the upper bound (5.6.11) in Theorem 5.6.3 (we actually only mimic the Case B of Theorem 5.6.1, the other cases can be treated similarly).

Theorem 5.10.5 (Possible improvement of the first eigenvalue upper bound)**.** *Let $\underline{\lambda}_1, \underline{\lambda}_2$ be as in Theorem 5.6.1. Let $(u_{1h}, \lambda_{1h}) \in \mathbb{P}_p(\mathcal{T}_h) \times \mathbb{R}^+$, $p \geq 1$, be arbitrary. Let s_{1h} be constructed following Definition 5.3.3 and r_{1h} following Definition 5.10.3. Let $(s_{1h}, \chi_1) > 0$ and*

$$\begin{aligned}\overline{\alpha}_{1h} &:= \sqrt{2}\left(1 - \frac{\lambda_{1h}}{\underline{\lambda}_2}\right)^{-1} \underline{\lambda}_2^{-\frac{1}{2}} \frac{1}{\|s_{1h}\|}\left(\frac{\lambda_{1h}}{\sqrt{\underline{\lambda}_1}}\|u_{1h} - s_{1h}\| + \|\nabla s_{1h} + \boldsymbol{\sigma}_{1h}\|\right) \\ &\leq \min\left\{\sqrt{2}, \|\chi_1\|^{-1}(\tilde{s}_{1h}, \chi_1)\right\},\end{aligned}$$

with $\tilde{s}_{1h} := \frac{s_{1h}}{\|s_{1h}\|}$. Then

$$\lambda_1 \leq \|\nabla \tilde{s}_{1h}\|^2 - \tilde{\eta}_1,$$

where

$$\begin{aligned}\tilde{\eta}_1 &:= \max\left\{\frac{1}{4}\left(1 - \frac{\|\nabla \tilde{s}_{1h}\|^2}{\underline{\lambda}_2}\right)\left(1 - \frac{\overline{\alpha}_{1h}^2}{4}\right)\left(\sqrt{d_h} - \left(\underline{\lambda}_1 + 2\left|\lambda_{1h} - \|\nabla \tilde{s}_{1h}\|^2\right|\right)\right), 0\right\}, \\ d_h &:= \underline{\lambda}_1^2 + 4\underline{\lambda}_1 \frac{\langle \mathrm{Res}_\theta(\tilde{s}_{1h}, \lambda_{1h}), r_{1h}\rangle_{V',V}^2}{\|\nabla r_{1h}\|^2} + 4\underline{\lambda}_1\left|\lambda_{1h} - \|\nabla \tilde{s}_{1h}\|^2\right|.\end{aligned}$$

Proof. Note first that all the assumptions of [52, Theorems 3.4 and 3.5] are satisfied. We start by the second bound in [52, Theorem 3.4] which immediately implies, using $\underline{\lambda}_1 \leq \lambda_1$, $\underline{\lambda}_2 \leq \lambda_2$, and $\lambda_1 \leq \|\nabla \tilde{s}_{1h}\|$,

$$\lambda_1 \leq \|\nabla \tilde{s}_{1h}\|^2 - \frac{1}{2}\left(1 - \frac{\|\nabla \tilde{s}_{1h}\|^2}{\underline{\lambda}_2}\right)\left(1 - \frac{\overline{\alpha}_{1h}^2}{4}\right)\|\nabla(u_1 - \tilde{s}_{1h})\|^2.$$

Similarly, the second bound in [52, Theorem 3.5] now takes the form

$$\|\mathrm{Res}_\theta(\tilde{s}_{1h}, \lambda_{1h})\|_{-1}^2 \leq \frac{\left(\left|\lambda_{1h} - \|\nabla \tilde{s}_{1h}\|^2\right| + \|\nabla\left(u_1 - \tilde{s}_{1h}\right)\|^2\right)^2}{\lambda_1} + \|\nabla(u_1 - \tilde{s}_{1h})\|^2.$$

Denote $l_h := \left|\lambda_{1h} - \|\nabla \tilde{s}_{1h}\|^2\right|$, $R_h := \langle \mathrm{Res}_\theta(\tilde{s}_{1h}, \lambda_{1h}), r_{1h}\rangle_{V',V}^2 / \|\nabla r_{1h}\|^2$, as well as $e_h := \|\nabla\left(u_1 - \tilde{s}_{1h}\right)\|^2$. Combined with Lemma 5.10.4 and $0 < \underline{\lambda}_1 \leq \lambda_1$, this last inequality implies

$$e_h^2 + e_h\left(\underline{\lambda}_1 + 2l_h\right) - \left(\underline{\lambda}_1 R_h - l_h^2\right) \geq 0.$$

Note that the discriminant of this quadratic inequality is the term d_h and that it is non-negative. Thus

$$e_h \geq \frac{-\left(\underline{\lambda}_1 + 2l_h\right) + \sqrt{d_h}}{2}$$

and the desired bound follows. Note finally that for this estimate to actually improve on (5.6.11), $\tilde{\eta}_1$ needs to be positive, which follows when $\underline{\lambda}_1 R_h > l_h^2$ and $\|\nabla \tilde{s}_{1h}\|^2 < \underline{\lambda}_2$. □ □

Acknowledgements

This work was supported by the ANR project MANIF "Mathematical and numerical issues in first-principle molecular simulation". The last author has also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 647134 GATIPOR).

Part IV

A posteriori error estimation for a nonlinear eigenvalue problem

Chapter 6

A posteriori analysis of a nonlinear Gross–Pitaevskii-type eigenvalue problem

We expose in this chapter the results of [98]. This work was done in collaboration with Yvon Maday.

Abstract

In this paper, we provide a first full *a posteriori* error analysis for variational approximations of the ground state eigenvector of a nonlinear elliptic problem of the Gross-Pitaevskii type, more precisely of the form $-\Delta u + Vu + u^3 = \lambda u$, $\|u\|_{L^2} = 1$, with periodic boundary conditions in one dimension. Denoting by (u_N, λ_N) the variational approximation of the ground state eigenpair (u, λ) based on a Fourier spectral approximation and (u_N^k, λ_N^k) the approximate solution at the k^{th} iteration of an algorithm used to solve the nonlinear problem, we first provide a precised *a priori* analysis of the convergence rates of $\|u - u_N\|_{H^1}$, $\|u - u_N\|_{L^2}$, $|\lambda - \lambda_N|$ and then present original *a posteriori* estimates in the convergence rates of $\|u - u_N^k\|_{H^1}$ when N and k go to infinity. We introduce a residual standing for the global error $R_N^k = -\Delta u_N^k + V u_N^k + (u_N^k)^3 - \lambda_N^k u_N^k$ and we divide it into two residuals characterizing respectively the error due to the discretization of the space and the finite number of iterations when solving the problem numerically. Finally, in a series of numerical tests, we illustrate numerically the performances of this *a posteriori* analysis.

6.1 Introduction

Nonlinear eigenvalue problems are involved in many application fields such as nonlinear mechanics, theoretical physics and electronic structure calculations. The numerical simulation of these problems demands a lot of computational resources both due to the accuracy that is generally required in the applications which implies the use of a large number of degrees of freedom and also due to the nonlinear nature of the models that leads to iterative solution techniques with a large number of steps. The tuning of the two above ingredients involved in the approximation methods (number of degrees of freedom and number of iterations) is, most of the times and in the best cases guided both by empirical reasons and by the available volume of computing resources. From the mathematical point of view, these questions are related to the numerical analysis of the discretization approaches that allow to establish in a rigorous way the link between the discretization and solution parameters and the error between the approximate solution(s) and the exact one(s). In the frame of the numerical analysis, and related to error bounds, we can distinguish between two types of contributions: the *a priori* analysis and the *a posteriori* analysis. The *a priori* version allows to qualify the convergence of the approximation methods when the number of degrees of freedom and/or the amount of work necessary for the computation of the discrete solution increase. This is generally done by upper bounding the error by a constant times the best approximation given by the projection of the exact solution onto the discrete space. The above constant that appears in the *a priori* analysis is generally not fully known, nor actually the distance between the solution and its projection. Most of the times, the latter is evaluated from the regularity property of the solution that is at best only roughly estimated. On the contrary, the *a posteriori* analysis provides a (more or less) precise upper bound of the actual error after a computation has been performed. This bound involves only quantities that are or can easily be evaluated at a lower cost than the computation of the discrete solution. Note that *a posteriori* analysis, thanks to the notion of indicators, may tell you what to do in order to improve the accuracy, but will not tell you what to do to diminish the current error by e.g. a factor 2. On the contrary, *a posteriori* analysis provides a stopping criteria when the desired accuracy is reached, an *a priori* estimator fails to do so.

The (*a priori*) numerical analysis of such nonlinear eigenvalue problems is quite recent and relies on the papers [255], [43], [73], [161], [44], [72] and the references therein. These papers only consider the discretization error due to the use of a given number of degrees of freedom in order to approximate the problem of interest. For an analysis of the convergence of the iterative algorithms to solve the nonlinear eigenvalue problem (or the associated nonlinear minimization problem), the papers issued from [54], [41], [154] provide *a priori* convergence results and allow to understand the basics for the failure of some classical approaches and how to remedy.

As is standard, all these *a priori* approaches allow to state that, provided that you put enough computing resources, the approximation will be good. Such results are classically insufficient because the amount of required computing resources for large problems is very often out of the possibility that you can afford. This is the reason why *a posteriori* approaches (estimators and indicators) have been designed. As far as we know, the first paper in the direction of *a posteriori* estimates is [184], where the analysis of the Hartree–Fock problem was performed and error bounds (i.e. upper and lower bounds) for the ground state energy was proposed. We refer also to the more recent contributions [81], [73], [80].

The present paper is the first of a series that aims at providing precise information on the accuracy of the approximation as a function of the number of degrees of freedom that are used and the number of iterations at which we stop the numerical process. For the sake of clarity in the tools that we use, the analysis is explained on a nonlinear equation that enters in the class

of Gross-Pitaevskii equations ([204]) and we focus on a one dimensional example to present both the theory and the numerical simulation that illustrate it. This allows us to propose *a posteriori* estimates and indicators based on residual techniques that discriminate the effect of the discretization parameter (the number of degrees of freedom) from the parameters attached to the solution procedure (i.e. the number of iterations). The interest of this indicator is to balance the complexity of the computation by tuning properly the number of iterations to the achievable best fit that the discretization allows. Note that the indicator is not used to refine *locally* the discretization mesh — this one being by construction uniform — however the indicator informs on the level of accuracy that is a direct consequence of the number of degrees of freedom used in the discretization and indicates also if it should be (uniformly) increased. The generalization of these tools for the more difficult problem of the Kohn–Sham problem involves a series of technical difficulties and is on its way (see [49]). We refer to the [45] for a general presentation of the mathematical models and approaches for their simulations in computational quantum chemistry.

In this paper, we focus on the following nonlinear eigenvalue problems arising in the study of variational models of the form

$$I = \inf \left\{ E(v),\ v \in X,\ \int_\Omega v^2 = 1 \right\}, \tag{6.1.1}$$

where Ω is here simply the unit cell $(0,1)$ of a periodic lattice $\mathcal{R}$ of $\mathbb{R}$, the energy functional E is of the form

$$E(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 + \frac{1}{2}\int_\Omega V v^2 + \frac{1}{4}\int_\Omega v^4, \tag{6.1.2}$$

and $X = H^1_\#(\Omega)$ is the Sobolev space, defined in the more general settings for any $s \in \mathbb{R}$,

$$H^s_\#(\Omega) = \{ v|_\Omega,\ v \in H^s_{\rm loc}(\mathbb{R}) \mid v \text{ is 1-periodic} \},$$

provided with the norm denoted as $\|.\|_{H^s}$ and for any $k \in \mathbb{N}$,

$$C^k_\#(\Omega) = \left\{ v|_\Omega,\ v \in C^k(\mathbb{R}) \mid v \text{ is 1-periodic} \right\}.$$

Problem (6.1.1) has a unique solution (up to the sign, see e.g. [43]).

Remark 6.1.1. *For simplicity, the periodic cell is considered here of size 1. But it is easy to rescale the equation if the periodic cell is different from $\Omega = (0,1)$. The only change would be in the coefficient in front of the nonlinearity. The analysis would not be changed and is actually naturally scaled since we do take care of the various constants in our estimates.*

Let us remind the following Gagliardo Nirenberg inequality[1]

[1]For any $v \in H^1(\Omega)$ we can indeed write

$$\forall x, y \in \Omega, \quad v^2(x) \le v^2(y) + 2\sqrt{\int_\Omega v^2}\sqrt{\int_\Omega v'^2},$$

from which we deduce, after integration in the y variable that

$$\forall x \in \Omega, \quad v^2(x) \le \int_\Omega v^2 + 2\sqrt{\int_\Omega v^2}\sqrt{\int_\Omega v'^2} \le \sqrt{5}\sqrt{\int_\Omega v^2}\sqrt{\int_\Omega v^2 + \int_\Omega v'^2},$$

thus the Gagliardo Nirenberg inequality; we also derive

$$\forall x \in \Omega, \quad v^2(x) \le \int_\Omega v^2 + 2\sqrt{\int_\Omega v^2}\sqrt{\int_\Omega v'^2} \le (\frac{1+\sqrt{5}}{2})(\int_\Omega v^2 + \int_\Omega v'^2),$$

and the Sobolev embedding constant follows from $\int_\Omega v^p \le \|v\|^{p-2}_{L^\infty}\int_\Omega v^2$

$$\forall v \in H^1_{\#}(\Omega), \quad \|v\|^2_{L^\infty(\Omega)} \leq \sqrt{5}\|v\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)}. \tag{6.1.3}$$

In addition, for $p \in [1, \infty]$, let us denote by C_p the Sobolev constant such that

$$\forall v \in X, \quad \|v\|_{L^p} \leq C_p \|v\|_{H^1}, \tag{6.1.4}$$

which holds with $C_p = (\frac{1+\sqrt{5}}{2})^{\frac{1}{2}-\frac{1}{p}}$.

In what follows, we shall assume that $V \in L^p(\Omega)$ for some $p \geq 2$, in addition, we assume that V is lower bounded[2]. It was shown in [43] that (6.1.1) has exactly two solutions: u and $-u$ in X with $u > 0$ in Ω. From the embedding of $H^1(\Omega)$ into $C^0(\overline{\Omega})$ (valid because we are in one dimension), $u \in C^0(\overline{\Omega})$. Moreover, E is Gâteaux differentiable on X and for any $v \in X$, $E'(v) = A_v v$ where

$$A_v = -\Delta + V + v^2. \tag{6.1.5}$$

Under the previous assumptions E is twice differentiable at any $v \in X$ and, by denoting $E''(v)$ the second derivative of E at v, we have for any $v, w, z \in X$,

$$\langle E''(v)w, z\rangle_{X',X} = \langle A_v w, z\rangle_{X',X} + 2\int_\Omega v^2 wz = \int_\Omega \nabla w.\nabla z + \int_\Omega Vwz + 3\int_\Omega v^2 wz. \tag{6.1.6}$$

Note that A_v defines a self-adjoint operator on $L^2(\Omega)$, with form domain X (see *e.g.* [212]). The function u therefore is solution to the Euler equation

$$\forall v \in X, \quad \langle A_u u - \lambda u, v\rangle_{X',X} = 0, \tag{6.1.7}$$

for some $\lambda \in \mathbb{R}$ (the Lagrange multiplier associated with the constraint $\|u\|^2_{L^2} = 1$) and equation (6.1.7), complemented with the constraint $\|u\|_{L^2} = 1$, takes the form of the following nonlinear eigenvalue problem

$$\left\{ \begin{array}{l} A_u u = \lambda u \\ \|u\|_{L^2} = 1, \end{array} \right. \text{ or again } \left\{ \begin{array}{l} -\Delta u + Vu + u^3 = \lambda u \\ \|u\|_{L^2} = 1, \end{array} \right. \tag{6.1.8}$$

which can be rewritten in a weak form as

$$\left\{ \begin{array}{l} \forall v \in X, \displaystyle\int_\Omega \nabla u \cdot \nabla v + \int_\Omega Vuv + \int_\Omega u^3 v = \lambda \int_\Omega uv \\ \displaystyle\int_\Omega u^2 = 1. \end{array} \right. \tag{6.1.9}$$

Let us remark that for any $v \in X$,

$$\langle E''(u)u - \lambda u, v\rangle_{X',X} = 2\int_\Omega u^3 v. \tag{6.1.10}$$

It should be noted in addition, that λ is the ground state eigenvalue of the linear operator A_u. An important result is that λ is a *simple* eigenvalue of A_u (see *e.g.* the Appendix of [43]).

A natural discretization in the periodic settings consists in using a Fourier basis. We denote by $(X_N)_{N>0}$ the family of finite-dimensional subspaces of X defined by

$$X_N = \text{Span}\left\{ e_k : x \mapsto e^{2ik\pi x}, |k| \leq N, k \in \mathbb{Z} \right\}.$$

[2]This assumption that the negative part of V: denoted as V_- is in $L^\infty(\Omega)$ is not crucial, it mainly allows us to simplify some estimates in what follows, in particular in section 6.3.1

Remind now that, for any $v \in L^2(\Omega)$,

$$v(x) = \sum_{\boldsymbol{k}\in\mathbb{Z}} \widehat{v}_k e_k(x),$$

where $\widehat{v}_k$ is the k^{th} Fourier coefficient of v:

$$\widehat{v}_k := \int_\Omega v(x)\,\overline{e_k(x)}\,dx = \int_\Omega v(x)\,e^{-2ik\pi x}\,dx.$$

For any real number s, we now endow the Sobolev space $H^s_\#(\Omega)$ with the equivalent norm expressed in Fourier modes as follows

$$\|v\|_{H^s} = \left(\sum_{k\in\mathbb{Z}} \left(1+|k|^2\right)^s |\widehat{v}_k|^2\right)^{1/2}, \tag{6.1.11}$$

in what follows we shall use only this definition of the H^s–norm (6.1.11). We obtain that for any $r \in \mathbb{R}$, and all $v \in H^r_\#(\Omega)$, the best approximation of v in $H^s_\#(\Omega)$ for any $s \le r$ is

$$\Pi_N v = \sum_{\boldsymbol{k}\in\mathbb{Z},|\boldsymbol{k}|\le N} \widehat{v}_k e_k. \tag{6.1.12}$$

The more regular v (the regularity being measured in terms of the Sobolev norms H^r), the faster the convergence of this truncated series to v: for any real numbers r and s with $s \le r$, we have (see e.g. [60])

$$\forall v \in H^r_\#(\Omega), \quad \|v - \Pi_N v\|_{H^s} \le \frac{1}{N^{r-s}}\|v\|_{H^r}, \tag{6.1.13}$$

and in particular for the solution u :

$$\min\{\|u - v_N\|_{H^1},\ v_N \in X_N\} \underset{N\to+\infty}{\longrightarrow} 0. \tag{6.1.14}$$

Let us now consider the variational approximation of (6.1.1) consisting in solving

$$I_N = \inf\left\{E(v_N),\ v_N \in X_N,\ \int_\Omega v_N^2 = 1\right\}. \tag{6.1.15}$$

Problem (6.1.15) has at least one minimizer u_N, which satisfies for some $\lambda_N \in \mathbb{R}$

$$\forall v_N \in X_N, \quad \langle A_{u_N} u_N - \lambda_N u_N, v_N\rangle_{X',X} = 0. \tag{6.1.16}$$

that is

$$\begin{cases} \forall v_N \in X_N, \int_\Omega \nabla u_N \cdot \nabla v_N + \int_\Omega V u_N v_N + \int_\Omega u_N^3 v_N = \lambda_N \int_\Omega u_N v_N, \\ \displaystyle\int_\Omega u_N^2 = 1. \end{cases} \tag{6.1.17}$$

A possible algorithm used to solve the equation numerically in the space X_N is the following: starting from a given pair (u_N^0, λ_N^0), we solve at each step the linear equation

$$\Pi_N\left(-\Delta u_N^{k*} + V u_N^{k*} + (u_N^{k-1})^2 u_N^{k*}\right) = \lambda_N^{k-1} u_N^{k-1}. \tag{6.1.18}$$

The discrete solution u_N^{k*} is completely determined by the knowledge of $(\lambda_N^{k-1}, u_N^{k-1})$. Since u_N^{k*} is *a priori* a non-normalized vector, we normalize it and define u_N^k by

$$u_N^k = \frac{u_N^{k*}}{\|u_N^{k*}\|_{L^2}}. \tag{6.1.19}$$

Finally, we define the approximation of the eigenvalue λ_N^k as a Rayleigh quotient being

$$\lambda_N^k = \frac{\displaystyle\int_\Omega (\nabla u_N^{k*})^2 + \int_\Omega V(u_N^{k*})^2 + \int_\Omega (u_N^{k*})^4}{\displaystyle\int_\Omega (u_N^{k*})^2} = \int_\Omega (\nabla u_N^k)^2 + \int_\Omega V(u_N^k)^2 + \int_\Omega (u_N^k)^4. \tag{6.1.20}$$

Remark 6.1.2. *In this paper, we do not consider error relatively to numerical integration since the nonlinearity $\int_\Omega (u_N^k)^4$ is only quartic hence easy to integrate. It is moreover possible to precompute the $4N+1$ coefficients of the Fourier projection of the potential $\Pi_{2N}V$, and since the problem relative to it is linear, exact integation can be performed.*

It should be noticed that the above algorithm corresponds to an extension of the inverse power method to this nonlinear eigenvalue problem. We can check numerically that such an algorithm converges (at least in all the simulations we have performed, possibly with a relaxation parameter — see the numerical results below). Moreover we can derive that the limit (λ_N, u_N) is a good approximation of the solution to problem (6.1.7). More precisely we can prove the following lemma.

Lemma 6.1.1. *Let us assume that there exists $u_N^* \in X_N$ with $\int_\Omega (u_N^*)^2 = 1$, such that the sequence $(u_N^{k*})_{k\geq 1}$ converges to u_N^* in H^1-norm when k goes to infinity, then*

- *the sequence $(\lambda_N^k)_{k\geq 1}$ converges to $\lambda_N^* = \displaystyle\int_\Omega (\nabla u_N^*)^2 + \int_\Omega V(u_N^*)^2 + \int_\Omega (u_N^*)^4$*
- *the sequence $(u_N^k)_{k\geq 1}$ converges to u_N^* in H^1-norm*
- *the limit (u_N^*, λ_N^*) verifies the nonlinear eigenvalue equation (6.1.17).*

In addition, if λ_N^* is the smallest eigenvalue of $A_{u_N^*}$, then u_N^* is solution in (6.1.15) and this solution is unique if N is large enough.

Remark 6.1.3. *The assumption "λ_N^* is the smallest eigenvalue of $A_{u_N^*}$" is needed because the inverse power method does not always converge to the lowest eigenvalue e.g. when, for a linear symmetric problem, the starting point is orthogonal to the targeted eigenvector. However, this hypothesis can be checked a posteriori by computing the first eigenvalue of the linear operator $A_{u_N^*}$ by another numerical method.*

Proof. The strong convergence of u_N^{k*} to u_N^* implies that the limit of the L^2-norm of u_N^{k*} is 1 and thus the sequence (u_N^k, λ_N^k) converges to (u_N^*, λ_N^*) in H^1-norm. Then, for any $\epsilon > 0$, there exists $k_0 \in \mathbb{N}$ such that for any $k \geq k_0$, $k \in \mathbb{N}$, the following assertions hold:

$$\begin{aligned} |\lambda_N^k - \lambda_N^*| &\leq \epsilon, \\ \|u_N^k - u_N^*\|_{H^1} &\leq \epsilon, \\ \|u_N^k - u_N^{k-1}\|_{H^1} &\leq \epsilon. \end{aligned} \tag{6.1.21}$$

For any $k > k_0$, using (6.1.3), we deduce that $\|u_N^k\|_{L^\infty} \leq \sqrt[4]{5}\sqrt{\epsilon + \|u_N^*\|_{H^1}}$. Let us now take $v_N \in X_N$, we have:

$$\begin{aligned}
&\int_\Omega \nabla u_N^* \nabla v_N + \int_\Omega V u_N^* v_N + \int_\Omega (u_N^*)^3 v_N - \lambda_N^* \int_\Omega u_N^* v_N \\
&\qquad = \int_\Omega \nabla u_N^k \nabla v_N + \int_\Omega V u_N^k v_N + \int_\Omega (u_N^{k-1})^2 u_N^k v_N - \lambda_N^{k-1} \int_\Omega u_N^{k-1} v_N \\
&\qquad + \int_\Omega \nabla (u_N^* - u_N^k) \nabla v_N + \int_\Omega V(u_N^* - u_N^k) v_N + \int_\Omega ((u_N^*)^3 - (u_N^k)^3) v_N \\
&\qquad + \int_\Omega ((u_N^k)^3 - (u_N^{k-1})^2 u_N^k) v_N + \lambda_N^{k-1} \int_\Omega u_N^{k-1} v_N - \lambda_N^* \int_\Omega u_N^* v_N
\end{aligned}$$

From (6.1.18) the second line is zero so we are left with

$$\begin{aligned}
&\int_\Omega \nabla u_N^* \nabla v_N + \int_\Omega V u_N^* v_N + \int_\Omega (u_N^*)^3 v_N - \lambda_N^* \int_\Omega u_N^* v_N \\
&= \int_\Omega \nabla (u_N^* - u_N^k) \nabla v_N + \int_\Omega V(u_N^* - u_N^k) v_N + \int_\Omega ((u_N^*)^2 + u_N^* u_N^k + (u_N^k)^2)(u_N^* - u_N^k) v_N \\
&\quad + \int_\Omega u_N^k (u_N^k + u_N^{k-1})(u_N^k - u_N^{k-1}) v_N + (\lambda_N^{k-1} - \lambda_N^*) \int_\Omega u_N^{k-1} v_N + \lambda_N^* \int_\Omega (u_N^{k-1} - u_N^*) v_N.
\end{aligned}$$

Hence

$$\begin{aligned}
&\left| \int_\Omega \nabla u_N^* \nabla v_N + \int_\Omega V u_N^* v_N + \int_\Omega (u_N^*)^3 v_N - \lambda_N^* \int_\Omega u_N^* v_N \right| \\
&\qquad \leq \|\nabla (u_N^* - u_N^k)\|_{L^2} \|\nabla v_N\|_{L^2} + \|V\|_{L^p} \|u_N^* - u_N^k\|_{L^\infty} \|v_N\|_{L^{p'}} \\
&\qquad + (\|u_N^*\|_{L^\infty}^2 + \|u_N^*\|_{L^\infty} \|u_N^k\|_{L^\infty} + \|u_N^k\|_{L^\infty}^2) \|u_N^* - u_N^k\|_{L^2} \|v_N\|_{L^2} \\
&\qquad + \|u_N^k\|_{L^\infty} (\|u_N^k\|_{L^\infty} + \|u_N^{k-1}\|_{L^\infty}) \|u_N^k - u_N^{k-1}\|_{L^2} \|v_N\|_{L^2} \\
&\qquad + |\lambda_N^{k-1} - \lambda_N^*| \|u_N^{k-1}\|_{L^2} \|v_N\|_{L^2} + |\lambda_N^*| \|u_N^{k-1} - u_N^*\|_{L^2} \|v_N\|_{L^2}.
\end{aligned}$$

with $\frac{1}{p} + \frac{1}{p'} = 1$ .Then from (6.1.21) and (6.1.3) we derive

$$\begin{aligned}
&\left| \int_\Omega \nabla u_N^* \nabla v_N + \int_\Omega V u_N^* v_N + \int_\Omega (u_N^*)^3 v_N - \lambda_N^* \int_\Omega u_N^* v_N \right| \\
&\qquad \leq \epsilon \|\nabla v_N\|_{L^2} + \epsilon \sqrt[4]{5}\, \|V\|_{L^p} \|v_N\|_{L^{p'}} + \epsilon \Big(\|u_N^*\|_{L^\infty}^2 + \sqrt[4]{5}\, \|u_N^*\|_{L^\infty} \sqrt{\epsilon + \|u_N^*\|_{H^1}} \\
&\qquad\qquad + \sqrt{5}(\epsilon + \|u_N^*\|_{H^1}) + 2\sqrt{5}(\epsilon + \|u_N^*\|_{H^1}) + 1 + |\lambda_N^*| \Big) \|v_N\|_{L^2}.
\end{aligned}$$

We then easily deduce that the limit (u_N^*, λ_N^*) verifies the non-linear eigenvalue equation (6.1.17).

Let us prove the last point of the lemma. We note that u_N defined in (6.1.16) and u_N^* the limit of the sequence above are both solutions to the same nonlinear (discrete) eigenvalue problem associated with λ_N (resp. λ_N^*) minimum eigenvalues of $\Pi_N A_{u_N}$ (resp. $\Pi_N A_{u_N^*}$). Besides, by definition, $\lambda_N \leq \lambda_N^*$.

We deduce

$$\langle (A_{u_N} - \lambda_N)(u_N - u_N^*), (u_N - u_N^*) \rangle = \lambda_N^* - \lambda_N + \int (u_N^2 - (u_N^*)^2)(u_N^*)^2$$

and

$$\left\langle (A_{u_N^*} - \lambda_N^*)(u_N - u_N^*), (u_N - u_N^*) \right\rangle = \lambda_N - \lambda_N^* + \int ((u_N^*)^2 - u_N^2) u_N^2$$

The two above lefthand sides are nonnegative. By adding up these two quantities, we deduce that

$$-\int \left((u_N^*)^2 - u_N^2\right)^2 \geq 0$$

which proves that $u_N = u_N^*$. We then deduce easily that $\lambda_N = \lambda_N^*$. □

It is an interesting result *per se*, to highlight the result that is established at the end of the previous proof, and that can be stated at the continuous level :

Lemma 6.1.2. *Let v be an element in X such that v is an eigenvector of A_v (as defined in (6.1.5)), associated with its smallest eigenvalue, then v is equal to u, the ground state of the energy E as defined in (6.1.2).*

The following property will be used several times in the analysis:

Lemma 6.1.3. *For any $v, w \in X$ such that $\int_\Omega v^2 = \int_\Omega w^2 = 1$,*

$$\int_\Omega v(v-w) = 1 - \int_\Omega vw = \frac{1}{2}\|v-w\|_{L^2}^2. \tag{6.1.22}$$

In the remainder of this paper, we denote by u the unique positive solution of (6.1.1) and by u_N a minimizer of the discretized problem (6.1.15) such that $(u_N, u)_{L^2} \geq 0$.

6.2 A priori analysis

The purpose of this section is to provide a precised *a priori* analysis for the approximation of problem (6.1.1) by (6.1.17) : we establish error bounds on $\|u_N - u\|_{H^1}$, $\|u_N - u\|_{L^2}$, $|\lambda_N - \lambda|$ and $E(u_N) - E(u)$. Actually, we follow and refine the analysis presented in the paper [43] where the *a priori* analysis was done in a more general framework.

We provide this *a priori* analysis of (6.1.1) for two reasons. Firstly the particular form of the energy functional and the fact that the problem in one-dimensional allows to simplify the proofs and understand better the basic ingredients that will be used in the next section. Secondly, and more importantly, we need to be as precise as possible in order to provide an accurate evaluation of the various constants that are involved in the error bounds of the *a posteriori* analysis whenever this is possible.

Lemma 6.2.1. (precised version of Lemma 1 of [43]) *If $V \in L^p(\Omega)$,with $\Omega = (0,1)$, there exist $\beta > 0$, $M_1, M_3 \in \mathbb{R}_+$ and $\gamma > 0$ such that for any $v \in X$ and any $N \in \mathbb{N}$,*

$$0 \leq \langle (A_u - \lambda)v, v\rangle_{X',X} \leq M_1 \|v\|_{H^1}^2 \tag{6.2.1}$$

$$\beta\|v\|_{H^1}^2 \leq \langle (E''(u) - \lambda)v, v\rangle_{X',X} \leq M_3\|v\|_{H^1}^2. \tag{6.2.2}$$

$$\gamma\|u_N - u\|_{H^1}^2 \leq \langle (A_u - \lambda)(u_N - u), (u_N - u)\rangle_{X',X}. \tag{6.2.3}$$

Moreover the constants are

$$M_m = 1 + \|V\|_{L^p} C_{2p'}^2 + m\|u\|_{L^\infty}^2 + |\lambda| \tag{6.2.4}$$

$$\beta = \frac{1}{2}\frac{\eta}{\eta+\chi}, \qquad \eta = \min(\lambda_2 - \lambda, 2), \qquad \chi = |\lambda| + 1 + \frac{5\|V\|_{L^p}^2}{2} \tag{6.2.5}$$

$$\gamma = \frac{1}{2}\frac{\eta}{\eta+2\chi} \tag{6.2.6}$$

where $\lambda_2 > \lambda$ is the the second smallest eigenvalue of A_u and $p' = (1-p^{-1})^{-1}$.

Proof. From (6.1.5), we have for every $v \in X$,

$$\begin{aligned} |\langle (A_u - \lambda)v, v\rangle_{X',X}| &\leq \|\nabla v\|_{L^2}^2 + \|V\|_{L^p}\|v\|_{L^{2p'}}^2 + \|u^2\|_{L^\infty}\|v\|_{L^2}^2 + |\lambda|\|v\|_{L^2}^2 \\ &\leq \left(1 + \|V\|_{L^p} C_{2p'}^2 + \|u^2\|_{L^\infty} + |\lambda|\right)\|v\|_{H^1}^2 = M_1\|v\|_{H^1}^2, \end{aligned}$$

which is (6.2.1) with $p' = (1-p^{-1})^{-1}$. Moreover, from (6.1.6)

$$\begin{aligned} |\langle (E''(u) - \lambda)v, v\rangle_{X',X}| &\leq |\langle (A_u - \lambda)v, v\rangle_{X',X}| + 2\|u^2\|_{L^\infty}\|v\|_{L^2}^2 \\ &\leq \left(1 + \|V\|_{L^p} C_{2p'}^2 + 3\|u^2\|_{L^\infty} + |\lambda|\right)\|v\|_{H^1}^2, \end{aligned}$$

hence the upper bound in (6.2.2) with constant M_3 defined in (6.2.4).

The fact that λ, the lowest eigenvalue of A_u, is simple (see the Appendix of [43]) provides the lower bound in (6.2.1). Indeed, the operator $A_u - \lambda$ is positive over the set $u^\perp$ defined as

$$u^\perp = \left\{ v \in X \mid \int_\Omega uv = 0 \right\}, \tag{6.2.7}$$

more precisely we have, for any $v \in X$,

$$\langle (A_u - \lambda)v, v\rangle_{X',X} \geq (\lambda_2 - \lambda)(\|v\|_{L^2}^2 - |(u,v)_{L^2}|^2) \geq \eta(\|v\|_{L^2}^2 - |(u,v)_{L^2}|^2) \geq 0, \tag{6.2.8}$$

with $\eta = \min(\lambda_2 - \lambda, 2)$. On the one hand for any $v \in X$,

$$\begin{aligned} \langle (E''(u) - \lambda)v, v\rangle_{X',X} &= \langle (A_u - \lambda)v, v\rangle_{X',X} + 2\int_\Omega u^2 v^2 \\ &\geq \eta(\|v\|_{L^2}^2 - |(u,v)_{L^2}|^2) + 2\int_\Omega u^2 v^2 \\ &\geq \eta\|v\|_{L^2}^2 + 2\int_\Omega u^2 v^2 - \eta\left(\int_\Omega uv\right)^2 \\ &\geq \eta\|v\|_{L^2}^2 + (2-\eta)\left(\int_\Omega uv\right)^2 \\ &\geq \eta\|v\|_{L^2}^2. \end{aligned} \tag{6.2.9}$$

On the other hand for any $v \in X$,

$$\begin{aligned} \langle (A_u - \lambda)v, v\rangle_{X',X} &\geq \|\nabla v\|_{L^2}^2 - \|V\|_{L^p}\|v\|_{L^\infty}^2 - |\lambda|\|v\|_{L^2}^2 \\ &\geq \|v\|_{H^1}^2 - \sqrt{5}\|V\|_{L^p}\|v\|_{L^2}\|v\|_{H^1} - (|\lambda|+1)\|v\|_{L^2}^2, \end{aligned}$$

by using the Gagliardo-Nirenberg inequality (6.1.3). Thanks to the inequality between arithmetic and geometric means applied to $\|v\|_{H^1}$ and $\|V\|_{L^p}\|v\|_{L^2}$, we deduce that

$$\begin{aligned} \langle (E''(u) - \lambda)v, v\rangle_{X',X} &\geq \langle (A_u - \lambda)v, v\rangle_{X',X} \\ &\geq \frac{1}{2}\|v\|_{H^1}^2 - \left(|\lambda| + 1 + \frac{5\|V\|_{L^p}^2}{2}\right)\|v\|_{L^2}^2. \end{aligned} \tag{6.2.10}$$

Combining (6.2.9) with (6.2.10) we get the lower bound in (6.2.2) with the constant β defined in (6.2.5).

To prove (6.2.3) we notice from (6.1.22) and the positivity of $(u, u_N)_{L^2}$ that

$$\|u_N\|_{L^2}^2 - |(u, u_N)_{L^2}|^2 \geq 1 - (u, u_N)_{L^2} = \frac{1}{2}\|u_N - u\|_{L^2}^2.$$

It therefore readily follows from (6.1.7) and (6.2.8) that

$$\langle (A_u - \lambda)(u_N - u), (u_N - u)\rangle_{X',X} = \langle (A_u - \lambda)(u_N), (u_N)\rangle_{X',X} \geq \frac{\eta}{2}\|u_N - u\|_{L^2}^2. \quad (6.2.11)$$

We also have from (6.2.10) that

$$\langle (A_u - \lambda)(u_N - u), (u_N - u)\rangle_{X',X} \geq \frac{1}{2}\|u_N - u\|_{H^1}^2 - \chi\|u_N - u\|_{L^2}^2 \quad (6.2.12)$$

with χ defined in (6.2.5). From (6.2.11) and (6.2.12) we can write

$$\langle (A_u - \lambda)(u_N - u), (u_N - u)\rangle_{X',X} \geq \frac{1}{2}\frac{\eta/2}{\eta/2 + \chi}\|u_N - u\|_{H^1}^2$$

Hence (6.2.3) with γ defined in (6.2.6). □

For $w \in X'$, we denote by ψ_w in $u^\perp$ defined in (6.2.7) the unique solution to the adjoint problem

$$\begin{cases} \text{find } \psi_w \in u^\perp \text{ such that} \\ \forall v \in u^\perp, \quad \langle (E''(u) - \lambda)\psi_w, v\rangle_{X',X} = \langle w, v\rangle_{X',X}. \end{cases} \quad (6.2.13)$$

The existence and uniqueness of the solution to (6.2.13) is a straightforward consequence of (6.2.2) and the Lax-Milgram lemma that also provides the estimate,

$$\forall w \in L^2(\Omega), \quad \|\psi_w\|_{H^1} \leq \beta^{-1}\|w\|_{X'} \leq \beta^{-1}\|w\|_{L^2}. \quad (6.2.14)$$

Besides this existence and stability result, the (very) simple elliptic regularity result, that we state without proof, follows

Lemma 6.2.2. *Assume $V \in L^p(\Omega)$, $p \geq 2$ then, there exists a constant $\tilde{C} = \frac{3\sqrt[4]{5}}{\beta}(\|V\|_{L^p} + 1) + \frac{\lambda}{\beta} + 1$ such that*

$$\|\psi_w\|_{H^2} \leq \tilde{C}\|w\|_{L^2}. \quad (6.2.15)$$

Let us now state the first *a priori* result of this section.

Theorem 6.2.1. *Under the previous assumptions,*

$$u_N \text{ converges strongly to } u \text{ in } H^1(\Omega) \text{ for } N \to +\infty. \quad (6.2.16)$$

In addition, there exists $C^E \in \mathbb{R}_+$ such that for any $N \in \mathbb{N}$,

$$\frac{\gamma}{2}\|u_N - u\|_{H^1}^2 \leq E(u_N) - E(u) \leq C^E\|u_N - u\|_{H^1}^2, \quad (6.2.17)$$

there exists $C^\lambda \in \mathbb{R}_+$ such that for any $N \in \mathbb{N}$,

$$|\lambda_N - \lambda| \leq C^\lambda\left(\|u_N - u\|_{H^1}^2 + \|u_N - u\|_{L^2}\right), \quad (6.2.18)$$

there exist $N_0 \in \mathbb{N}$ *and* $C^{H1} \in \mathbb{R}_+$ *such that for any* $N \geq N_0, N \in \mathbb{N}$,

$$\|u_N - u\|_{H^1} \leq C^{H1} \min_{v_N \in X_N} \|v_N - u\|_{H^1}, \tag{6.2.19}$$

and there exist $N_1 \in \mathbb{N}$ *and* $C^{L2} \in \mathbb{R}_+$ *such that for any* $N \geq N_1, N \in \mathbb{N}$,

$$\|u_N - u\|_{L^2}^2 \leq C^{L2} \|u_N - u\|_{H^1} \min_{\psi_N \in X_N} \|\psi_{u_N - u} - \psi_N\|_{H^1}. \tag{6.2.20}$$

Remark 6.2.1. *It should be noticed at this level that, even if the constants above* C^λ, C^{H1} *and* C^{L2} *can be estimated quite accurately, they involve* u *so as does* $\min_{v_N \in X_N} \|v_N - u\|_{H^1}$ *: it results that these estimates are not constructive.*

Proof. We have

$$\begin{aligned} E(u_N) - E(u) &= \frac{1}{2}\langle A_u u_N, u_N\rangle_{X',X} - \frac{1}{2}\langle A_u u, u\rangle_{X',X} + \frac{1}{2}\int_\Omega \left(\frac{u_N^4}{2} - \frac{u^4}{2} - u^2(u_N^2 - u^2)\right) \\ &= \frac{1}{2}\langle (A_u - \lambda)(u_N - u), (u_N - u)\rangle_{X',X} + \frac{1}{4}\int_\Omega (u_N^2 - u^2)^2. \end{aligned} \tag{6.2.21}$$

Using (6.2.3) and the fact that the second term on the right hand side is positive we get

$$E(u_N) - E(u) \geq \frac{\gamma}{2}\|u_N - u\|_{H^1}^2,$$

hence

$$\|u_N - u\|_{H^1}^2 \leq \frac{2}{\gamma}\left(E(u_N) - E(u)\right) \leq \frac{2}{\gamma} \inf_{v_N \in X_N, \|v_N\|_{L^2}=1} E(v_N) - E(u).$$

Let us now denote by

$$J_N = \min_{v_N \in X_N, \|v_N\|_{L^2}=1} \|v_N - u\|_{H^1}. \tag{6.2.22}$$

The functional E being strongly continuous on X, we obtain

$$\|u_N - u\|_{H^1}^2 \leq \frac{2}{\gamma}\varepsilon[J_N],$$

where ε tends to zero with its argument.

From the definition of $\Pi_N u$ the H^1-projector of u on X_N, $(\Pi_N u)_{N>0}$ converges to u in X when N goes to infinity. Denoting by $\widetilde{u}_N = \|\Pi_N u\|_{L^2}^{-1}\Pi_N u$ (which is well defined, since $u > 0$ means that it is not with zero average, hence $\Pi_N u$ is never null), we thus have

$$\begin{aligned} J_N &\leq \|\Pi_N u/\|\Pi_N u\|_{L^2} - u\|_{H^1} \\ &\leq \|\Pi_N u - u\|_{H^1} + \frac{\|\Pi_N u\|_{H^1}}{\|\Pi_N u\|_{L^2}}\,|1 - \|\Pi_N u\|_{L^2}| \\ &\leq \|\Pi_N u - u\|_{H^1} + \frac{\|\Pi_N u\|_{H^1}}{\|\Pi_N u\|_{L^2}}\|u - \Pi_N u\|_{L^2} \quad \text{(from the triangle inequality)} \\ &\leq \left(1 + \frac{\|\Pi_N u\|_{H^1}}{\|\Pi_N u\|_{L^2}}\right)\|\Pi_N u - u\|_{H^1} \\ &\leq 2(\|u\|_{H^1} + 1)\|\Pi_N u - u\|_{H^1} = 2(\|u\|_{H^1} + 1)\min_{v_N \in X_N}\|v_N - u\|_{H^1}, \end{aligned} \tag{6.2.23}$$

as soon as N is large enough, hence

$$J_N \underset{N \to +\infty}{\longrightarrow} 0,$$

and, consequently

$$\|u_N - u\|_{H^1} \underset{N\to+\infty}{\longrightarrow} 0,$$

that is (6.2.16). It follows that there exists $N_1 \in \mathbb{N}$ such that

$$\forall N > N_1, N \in \mathbb{N}, \quad \|u_N - u\|_{H^1} \leq \frac{1}{2} \quad \text{and} \quad \|u_N\|_{H^1} \leq 2\|u\|_{H^1}. \tag{6.2.24}$$

In addition,

$$\begin{aligned} E(u_N) - E(u) &= \frac{1}{2}\langle (A_u - \lambda)(u_N - u), (u_N - u)\rangle_{X',X} + \frac{1}{4}\int_\Omega (u_N^2 - u^2)^2 \\ &\leq \frac{M_1}{2}\|u_N - u\|_{H^1}^2 + \frac{1}{4}\|u_N - u\|_{L^\infty}^2 \|u_N + u\|_{L^2}^2 \quad \text{(from (6.2.1))} \\ &\leq \left(\frac{M_1}{2} + \sqrt{5}\right)\|u_N - u\|_{H^1}^2 \quad \text{(from (6.1.3) using } \|u\|_{L^2(\Omega)} = \|u_N\|_{L^2(\Omega)} = 1). \end{aligned}$$

Hence the upper bound in (6.2.17) with $C^E = \frac{M_1}{2} + \sqrt{5}$.

From (6.1.17) with $v_N = u_N$ and (6.1.9) with $v = u_N - u$, we remark that

$$\begin{aligned} \lambda_N - \lambda &= \lambda_N \int_\Omega u_N^2 - \lambda \int_\Omega u^2 \\ &= \int_\Omega (\nabla u_N)^2 + \int_\Omega V u_N^2 + \int_\Omega u_N^4 - \left(\int_\Omega (\nabla u)^2 + \int_\Omega V u^2 + \int_\Omega u^4\right) \\ &= \int_\Omega \nabla (u_N - u)^2 + \int_\Omega V (u_N - u)^2 + 2\int_\Omega \nabla u \cdot \nabla (u_N - u) \\ &\quad + 2\int_\Omega V u (u_N - u) + \int_\Omega u_N^4 - u^4 \\ &= \langle A_u (u_N - u), (u_N - u)\rangle_{X',X} - \int_\Omega u^2 (u_N - u)^2 + 2\lambda \int_\Omega u (u_N - u) \\ &\quad - 2\int_\Omega u^3 (u_N - u) + \int_\Omega u_N^4 - u^4 \\ &= \langle (A_u - \lambda)(u_N - u), (u_N - u)\rangle_{X',X} \\ &\quad + \int_\Omega u_N^2 (u_N + u)(u_N - u) \quad \text{(from (6.1.22))}, \end{aligned} \tag{6.2.25}$$

we also obtain

$$\begin{aligned} \left|\int_\Omega u_N^2 (u_N + u)(u_N - u)\right| &\leq \|u_N^2 (u_N + u)\|_{L^2} \|u_N - u\|_{L^2} \\ &\leq \|u_N\|_{L^\infty}^2 \|u_N + u\|_{L^2} \|u_N - u\|_{L^2} \\ &\leq \sqrt{5}\|u_N\|_{L^2}\|u_N\|_{H^1}\|u_N + u\|_{L^2}\|u_N - u\|_{L^2} \quad \text{(from (6.1.3))} \\ &\leq 4\sqrt{5}\|u\|_{H^1}\|u_N - u\|_{L^2} \quad \text{(from (6.2.24))}, \end{aligned}$$

and from (6.2.1)

$$\begin{aligned} |\lambda_N - \lambda| &\leq M_1 \|u_N - u\|_{H^1}^2 + 4\sqrt{5}\|u\|_{H^1}\|u_N - u\|_{L^2} \\ &\leq C^\lambda \left(\|u_N - u\|_{H^1}^2 + \|u_N - u\|_{L^2}\right), \end{aligned} \tag{6.2.26}$$

with $C^\lambda = \max\left(M_1, 4\sqrt{5}\|u\|_{H^1}\right)$.

In order to evaluate the H^1-norm of the error $u_N - u$, we first notice that

$$\forall v_N \in X_N, \quad \|u_N - u\|_{H^1} \leq \|u_N - v_N\|_{H^1} + \|v_N - u\|_{H^1}, \tag{6.2.27}$$

and from (6.2.2) that

$$\begin{aligned} \|u_N - v_N\|_{H^1}^2 &\leq \beta^{-1} \langle (E''(u) - \lambda)(u_N - v_N), (u_N - v_N)\rangle_{X',X} \\ &= \beta^{-1}\Big(\langle (E''(u) - \lambda)(u_N - u), (u_N - v_N)\rangle_{X',X} \\ &\qquad + \langle (E''(u) - \lambda)(u - v_N), (u_N - v_N)\rangle_{X',X} \Big). \end{aligned} \tag{6.2.28}$$

For any $w_N \in X_N$, using (6.1.6), (6.1.7) and (6.1.17)

$$\begin{aligned} \langle (E''(u) - \lambda)(u_N - u), w_N\rangle_{X',X} &= \langle (A_u - \lambda)(u_N - u), w_N\rangle_{X',X} + 2\int_\Omega u^2(u_N - u)w_N \\ &= \langle (A_u - \lambda)u_N, w_N\rangle_{X',X} + 2\int_\Omega u^2(u_N - u)w_N \\ &= \langle (A_u - A_{u_N})u_N, w_N\rangle_{X',X} \\ &\quad + (\lambda_N - \lambda)\int_\Omega u_N w_N + 2\int_\Omega u^2(u_N - u)w_N \\ &= (\lambda_N - \lambda)\int_\Omega u_N w_N + \int_\Omega \left(u^2 u_N - u_N^3 + 2u^2(u_N - u)\right) w_N \\ &= (\lambda_N - \lambda)\int_\Omega u_N w_N - \int_\Omega (u_N - u)^2(u_N + 2u)w_N. \end{aligned} \tag{6.2.29}$$

By using (6.1.22) with $v = u_N$ and $w = v_N$, (6.2.18) and (6.2.24), we obtain that for any $N \geq N_1, N \in \mathbb{N}$ and all $v_N \in X_N$ such that $\|v_N\|_{L^2} = 1$,

$$\begin{aligned} \left|\langle (E''(u) - \lambda)(u_N - u), (u_N - v_N)\rangle_{X',X}\right| &= \Big|(\lambda_N - \lambda)\langle u_N, u_N - v_N\rangle_{X',X} \\ &\quad - \int_\Omega (u_N - u)^2(u_N + 2u)(u_N - v_N)\Big| \\ &\leq \frac{1}{2}|\lambda_N - \lambda|\|u_N - v_N\|_{L^2}^2 \\ &\quad + \|u_N - v_N\|_{L^\infty}\|u_N - u\|_{L^2}\|u_N - u\|_{L^\infty}\|u_N + 2u\|_{L^2} \\ &\leq \frac{1}{2}C^\lambda\left(\|u_N - u\|_{H^1}^2 + \|u_N - u\|_{L^2}\right)\|u_N - v_N\|_{L^2}^2 \\ &\quad + 3\sqrt{5}\|u_N - v_N\|_{H^1}\|u_N - u\|_{L^2}\|u_N - u\|_{H^1}. \end{aligned} \tag{6.2.30}$$

It then follows from (6.2.2) that for any $N \geq N_1, N \in \mathbb{N}$ and all $v_N \in X_N$ such that $\|v_N\|_{L^2} = 1$,

$$\begin{aligned} \|u_N - v_N\|_{H^1}^2 &\leq \beta^{-1}\Big(\frac{1}{2}C^\lambda(\|u_N - u\|_{H^1}^2 + \|u_N - u\|_{L^2})\|u_N - v_N\|_{L^2}^2 \\ &\quad + 3\sqrt{5}\|u_N - v_N\|_{H^1}\|u_N - u\|_{L^2}\|u_N - u\|_{H^1} \\ &\quad + M_3\|u - v_N\|_{H^1}\|u_N - v_N\|_{H^1}\Big). \end{aligned}$$

So

$$\begin{aligned} \|u_N - v_N\|_{H^1} &\leq \beta^{-1}\Big(\frac{C^\lambda}{2}\left(\|u_N - u\|_{H^1}^2 + \|u_N - u\|_{L^2}\right)\|u_N - v_N\|_{H^1} \\ &\quad + 3\sqrt{5}\|u_N - u\|_{L^2}\|u_N - u\|_{H^1} + M_3\|u - v_N\|_{H^1}\Big), \end{aligned}$$

that is

$$\|u_N - v_N\|_{H^1}\left(1 - \beta^{-1}\frac{C^\lambda}{2}(\|u_N - u\|_{H^1}^2 + \|u_N - u\|_{L^2})\right) \leq \beta^{-1}\left(3\sqrt{5}\|u_N - u\|_{L^2}\|u_N - u\|_{H^1} + M_3\|u - v_N\|_{H^1}\right). \quad (6.2.31)$$

Since $\|u_N - u\|_{H^1} \underset{N\to+\infty}{\longrightarrow} 0$, there exists $N_2 \in \mathbb{N}$ such that $\forall N \geq N_2$,

$$\beta^{-1}\frac{C^\lambda}{2}(\|u_N - u\|_{H^1}^2 + \|u_N - u\|_{L^2}) \leq \frac{1}{2},$$

i.e.

$$\|u_N - v_N\|_{H^1} \leq \beta^{-1}\left(6\sqrt{5}\|u_N - u\|_{L^2}\|u_N - u\|_{H^1} + 2M_3\|u - v_N\|_{H^1}\right).$$

Then

$$\begin{aligned}\|u_N - u\|_{H^1} &\leq \|u_N - v_N\|_{H^1} + \|v_N - u\|_{H^1} \\ &\leq 6\sqrt{5}\beta^{-1}\|u_N - u\|_{L^2}\|u_N - u\|_{H^1} + (2M_3\beta^{-1} + 1)\|u - v_N\|_{H^1},\end{aligned}$$

hence

$$\|u_N - u\|_{H^1}\left(1 - 6\sqrt{5}\beta^{-1}\|u_N - u\|_{L^2}\right) \leq (2M_3\beta^{-1} + 1)\|u - v_N\|_{H^1}. \quad (6.2.32)$$

Besides, there exists $N_3 \in \mathbb{N}$ such that $\forall N \geq N_3$,

$$6\sqrt{5}\beta^{-1}\|u_N - u\|_{L^2} \leq \frac{1}{2}.$$

Then $\forall N \geq N_3$, $\|u_N - u\|_{H^1} \leq C^M\|u - v_N\|_{H^1}$ where $C^M \leq 2(2M_3\beta^{-1} + 1)$ (and $C^M \equiv C^M(N) \to 2M_3\beta^{-1} + 1$ as $N \to \infty$). Hence for any $N \geq N_3$, $\|u_N - u\|_{H^1} \leq C^M J_N$. From (6.2.23), we deduce

$$\|u_N - u\|_{H^1} \leq 2C^M(\|u\|_{H^1} + 1)\min_{v_N \in X_N}\|v_N - u\|_{H^1}$$

and (6.2.19) is proven with $C^{H1} \leq 4(2M_3\beta^{-1} + 1)(\|u\|_{H^1} + 1)$ (and $C^{H1} \equiv C^{H1}(N) \to (2M_3\beta^{-1} + 1)\|u\|_{H^1}$ as $N \to \infty$).

Let $\widetilde{\widetilde{u_N}}$ be the orthogonal projection, for the L^2 inner product, of u_N on the affine space $S = \left\{v \in L^2(\Omega) \,|\, \int_\Omega uv = 1\right\}$. One has

$$\widetilde{\widetilde{u_N}} \in X, \qquad \int_\Omega u\widetilde{\widetilde{u_N}} = 1, \qquad u_N - \widetilde{\widetilde{u_N}} \in S^\perp \quad \text{i.e.} \quad u_N - \widetilde{\widetilde{u_N}} \quad \text{is colinear to } u.$$

As $\int_\Omega(\widetilde{\widetilde{u_N}} - u)u = 0$, we have $\widetilde{\widetilde{u_N}} - u \in u^\perp$. Moreover

$$\int_\Omega(\widetilde{\widetilde{u_N}} - u_N)u = 1 - \int_\Omega u_N u = \frac{1}{2}\int_\Omega u_N^2 - \int_\Omega u_N u + \frac{1}{2}\int_\Omega u^2 = \frac{1}{2}\int_\Omega(u_N - u)^2,$$

hence

$$\widetilde{\widetilde{u_N}} - u_N = \frac{1}{2}\|u_N - u\|_{L^2}^2 u. \quad (6.2.33)$$

We can write the following

$$\begin{aligned}
\|u_N-u\|_{L^2}^2 &= \int_\Omega (u_N-u)(\widetilde{\widetilde{u_N}}-u)+\int_\Omega (u_N-u)(u_N-\widetilde{\widetilde{u_N}})\\
&= \int_\Omega (u_N-u)(\widetilde{\widetilde{u_N}}-u)-\frac{1}{2}\|u_N-u\|_{L^2}^2\int_\Omega (u_N-u)u \quad \text{(from (6.2.33))}\\
&= \int_\Omega (u_N-u)(\widetilde{\widetilde{u_N}}-u)+\frac{1}{2}\|u_N-u\|_{L^2}^2\left(1-\int_\Omega u_N u\right)\\
&= \int_\Omega (u_N-u)(\widetilde{\widetilde{u_N}}-u)+\frac{1}{4}\|u_N-u\|_{L^2}^4 \quad \text{(from (6.1.22))}\\
&= \langle (E''(u)-\lambda)\psi_{u_N-u},\widetilde{\widetilde{u_N}}-u\rangle_{X',X}+\frac{1}{4}\|u_N-u\|_{L^2}^4 \quad \text{(from (6.2.13))}\\
&= \langle (E''(u)-\lambda)\psi_{u_N-u},u_N-u\rangle_{X',X}\\
&\quad +\langle (E''(u)-\lambda)\psi_{u_N-u},\widetilde{\widetilde{u_N}}-u_N\rangle_{X',X}+\frac{1}{4}\|u_N-u\|_{L^2}^4\\
&= \langle (E''(u)-\lambda)(u_N-u),\psi_{u_N-u}\rangle_{X',X}\\
&\quad +\frac{1}{2}\|u_N-u\|_{L^2}^2\langle (E''(u)-\lambda)u,\psi_{u_N-u}\rangle_{X',X}+\frac{1}{4}\|u_N-u\|_{L^2}^4 \quad \text{(from (6.2.33))}\\
&= \langle (E''(u)-\lambda)(u_N-u),\psi_{u_N-u}\rangle_{X',X}\\
&\quad +\|u_N-u\|_{L^2}^2\int_\Omega u^3\psi_{u_N-u}+\frac{1}{4}\|u_N-u\|_{L^2}^4 \quad \text{(from (6.1.10))}.
\end{aligned}$$

For any $\psi_N\in X_N$, it therefore holds

$$\begin{aligned}
\|u_N-u\|_{L^2}^2 &= \langle (E''(u)-\lambda)(u_N-u),\psi_N\rangle_{X',X}+\langle (E''(u)-\lambda)(u_N-u),\psi_{u_N-u}-\psi_N\rangle_{X',X}\\
&\quad +\|u_N-u\|_{L^2}^2\int_\Omega u^3\psi_{u_N-u}+\frac{1}{4}\|u_N-u\|_{L^2}^4. \qquad (6.2.34)
\end{aligned}$$

From (6.2.29) with $w_N=\psi_N$, from (6.2.18) and (6.2.24), we obtain that for any $\psi_N\in X_N\cap u^\perp$,

$$\begin{aligned}
\left|\langle (E''(u)-\lambda)(u_N-u),\psi_N\rangle_{X',X}\right| &= (\lambda_N-\lambda)\int_\Omega (u_N-u)\psi_N-\int_\Omega (u_N-u)^2(u_N+2u)\psi_N\\
&\le C^\lambda\left(\|u_N-u\|_{H^1}^2+\|u_N-u\|_{L^2}\right)\|u_N-u\|_{L^2}\|\psi_N\|_{L^2}\\
&\quad +\|u_N-u\|_{L^\infty}^2\|u_N+2u\|_{L^2}\|\psi_N\|_{L^2}\\
&\le \left(3+C^\lambda\right)\Big(\|u_N-u\|_{L^\infty}^2\\
&\quad +(\|u_N-u\|_{H^1}^2+\|u_N-u\|_{L^2})\|u_N-u\|_{L^2}\Big)\|\psi_N\|_{H^1} \qquad (6.2.35)
\end{aligned}$$

Let $\psi_N^0\in X_N\cap u^\perp$ be such that

$$\|\psi_{u_N-u}-\psi_N^0\|_{H^1}=\min_{\psi_N\in X_N\cap u^\perp}\|\psi_{u_N-u}-\psi_N\|_{H^1}.$$

We deduce that $\|\psi_N^0\|_{H^1}\le 2\|\psi_{u_N-u}\|_{H^1}\le 2\beta^{-1}\|u_N-u\|_{L^2}$, then we obtain from (6.2.2), (6.2.24) (6.2.34) and (6.2.35) that for any $N\ge N_1$, $N\in\mathbb{N}$,

$$\begin{aligned}
\|u_N-u\|_{L^2}^2 &\le 2\beta^{-1}\left(3+C^\lambda\right)\left(\sqrt{5}\|u_N-u\|_{H^1}+\|u_N-u\|_{H^1}^2+\|u_N-u\|_{L^2}\right)\|u_N-u\|_{L^2}^2\\
&\quad +M_3\|u_N-u\|_{H^1}\|\psi_{u_N-u}-\psi_N^0\|_{H^1}+2\beta^{-1}\|u\|_{L^3}^3\|u_N-u\|_{L^2}^3+\frac{1}{4}\|u_N-u\|_{L^2}^4,
\end{aligned}$$

hence

$$\|u_N-u\|_{L^2}^2(1-2\beta^{-1}\left(3+C^\lambda\right)\left(\sqrt{5}\|u_N-u\|_{H^1}+\|u_N-u\|_{H^1}^2+\|u_N-u\|_{L^2}\right)$$
$$-2\beta^{-1}\|u\|_{L^3}^3\|u_N-u\|_{L^2}-\frac{1}{4}\|u_N-u\|_{L^2}^2)\leq M_3\|u_N-u\|_{H^1}\|\psi_{u_N-u}-\psi_N^0\|_{H^1}.$$

There exists $N_4\in\mathbb{N}$ such that for any $N\geq N_4$, $N\in\mathbb{N}$

$$2\beta^{-1}\left(3+C^\lambda\right)\left(\sqrt{5}\|u_N-u\|_{H^1}+\|u_N-u\|_{H^1}^2+\|u_N-u\|_{L^2}\right)$$
$$+2\beta^{-1}\|u\|_{L^3}^3\|u_N-u\|_{L^2}+\frac{1}{4}\|u_N-u\|_{L^2}^2\leq\frac{1}{2}.$$

Then we have for any $N\geq N_4$, $N\in\mathbb{N}$,

$$\|u_N-u\|_{L^2}^2\leq 2M_3\|u_N-u\|_{H^1}\|\psi_{u_N-u}-\psi_N^0\|_{H^1}. \tag{6.2.36}$$

Lastly, let us introduce the operator $\Pi_{X_N}^\perp v=\Pi_N v-\dfrac{(u,\Pi_N v)_{L^2}}{(u,\Pi_N u)_{L^2}}\Pi_N u$, such that $\Pi_{X_N}^\perp v\in u^\perp$ (where Π_N is defined in (6.1.12)). We have that

$$\begin{aligned}\forall v\in X\cap u^\perp,\quad \min_{v_N\in X_N\cap u^\perp}\|v_N-v\|_{H^1} &\leq \|\Pi_{X_N}^\perp v-v\|_{H^1}\\ &= \|\Pi_N v-v+\frac{(u,\Pi_N v-v)_{L^2}+(u,v)_{L^2}}{(u,\Pi_N u)_{L^2}}\Pi_N u\|_{H^1}\\ &\leq \left(1+\frac{\|\Pi_N u\|_{H^1}}{(u,\Pi_N u)_{L^2}}\right)\min_{v_N\in X_N}\|v_N-v\|_{H^1},\end{aligned} \tag{6.2.37}$$

hence (6.2.20) with $C^{L2}=2M_3\left(1+\dfrac{\|\Pi_N u\|_{H^1}}{(u,\Pi_N u)_{L^2}}\right)$, which concludes the proof of theorem 6.2.1.

□

6.3 A posteriori analysis

In this section we derive *a posteriori* estimates for the approximation of the problem (6.1.8), in order to quantify the error done during the iterative resolution (6.1.18), (6.1.19), (6.1.20) of the nonlinear eigenvalue problem; we introduce a residual measuring how close the approximate solution, obtained after — say — k iterations (u_N^k,λ_N^k) is to the exact one (u,λ).

We are in particular interested in deriving an accurate upper bound for the quantity

$$\left\|u-u_N^k\right\|_{H^1}=\max_{v\in H^1_\#}\frac{\displaystyle\int_\Omega\nabla(u-u_N^k)\cdot\nabla v+\int_\Omega(u-u_N^k)v}{\|v\|_{H^1}}, \tag{6.3.1}$$

where — of course — the direct knowledge of u is not available.

6.3.1 Preliminaries

We start by computing a crude *a posteriori* error estimator that allows to define a range of discretization parameters k and N that ensures that the discrete solution u_N^k verifies $\left\|u-u_N^k\right\|_{H^1}\leq$

c where c is a given constant $c > 0$ (that may not be so small as a target accuracy level), in particular this allows to verify that the discrete solution is close to the exact ground state. To do so, we use arguments stated in [37] or [40]. First we write the problem in an appropriate form: we define a function F being

$$F(\underline{u}) = \begin{pmatrix} -\Delta u + Vu + u^3 - \lambda u \\ \int u^2 - 1 \end{pmatrix} \tag{6.3.2}$$

for $\underline{u} = (u, \lambda) \in \mathcal{X} \equiv X \times \mathbb{R}$. Then we refer to the general theory that we recall for the sake of completeness: let $F : \mathcal{X} \to \mathcal{X}'$ be a $\mathcal{C}^1$ mapping and let $\underline{v} \in \mathcal{X}$ be such that $DF_{\underline{v}} \in \mathcal{L}(\mathcal{X}; \mathcal{X}')$ is an isomorphism. We introduce

$$\varepsilon = \|F(\underline{v})\|_{\mathcal{X}'}, \quad \gamma = \|DF_{\underline{v}}^{-1}\|_{\mathcal{X}';\mathcal{X}}, \quad L(\alpha) = \sup_{\underline{x} \in \overline{B}(\underline{v},\alpha)} \|DF_{\underline{v}} - DF_{\underline{x}}\|_{\mathcal{X};\mathcal{X}'} \tag{6.3.3}$$

where $B(\underline{v}, \alpha)$ is the ball in $\mathcal{X}$ of center $\underline{v}$ and radius α. Then, if $2\gamma L(2\gamma\varepsilon) \leq 1$, there exists a unique $\underline{u} \in \mathcal{X}$ in the ball $\overline{B}(\underline{v}, 2\gamma\varepsilon)$ such that $F(\underline{u}) = 0$ and

$$\|\underline{u} - \underline{v}\|_{\mathcal{X}} \leq 2\gamma \|F(\underline{v})\|_{\mathcal{X}'}.$$

We shall apply this result with $\underline{v} = (u_N^k, \lambda_N^k)$. In this case, because $\|u_N^k\|_{L^2} = 1$, the residual quantity: ε is (only) the part associated with the nonlinear eigenvalue problem for the approximate solution (u_N^k, λ_N^k), i.e.

$$\varepsilon = \| - \Delta u_N^k + V u_N^k + (u_N^k)^3 - \lambda_N^k u_N^k \|_{H^{-1}} \tag{6.3.4}$$

An essential element in this direction is thus the differential form DF that, computed at (u_N^k, λ_N^k) writes

$$\forall w \in X, \tau \in \mathbb{R}, \quad DF_{(u_N^k, \lambda_N^k)}(w, \tau) = \begin{pmatrix} -\Delta w + Vw + 3(u_N^k)^2 w - \tau u_N^k - \lambda_N^k w \\ 2 \int u_N^k w \end{pmatrix} \tag{6.3.5}$$

We will require that $DF_{(u_N^k, \lambda_N^k)}$ is an isomorphism. In order to check that statement, let us introduce the following linear eigenvalue problem associated with the discrete solution u_N^k : Find $v_N \in X_N$ and $\mu_N \in \mathbb{R}$ such that

$$\forall w_N \in X_N, \quad \int \nabla v_N \nabla w_N + \int V v_N w_N + \int (u_N^k)^2 v_N w_N = \mu_N \int v_N w_N, \quad \int v_N^2 = 1. \tag{6.3.6}$$

The associated eigenvalues ranked in increasing order are denoted by μ_N^i, $i \geq 1$, with

$$\mu_N^1 < \mu_N^2 \leq \mu_N^3 \leq ... \leq \mu_N^i \leq ...$$

and the corresponding normalized eigenvectors are denoted v_N^i, $i \geq 1$. Note that (v_N^i, μ_N^i) all depend on k through u_N^k. The existence of a gap

$$\delta_N^k = \mu_N^2 - \mu_N^1 > 0 \tag{6.3.7}$$

between the first and second eigenvalue is a well-known fact. The computation of μ_N^1 and μ_N^2 (that can be done with a power method) allows to provide a first coarse indicator of the convergence (in k) of the iterative algorithm since λ_N^k should be close to μ_N^1 and u_N^k should be close to $\pm v_N^1$ (and we define properly v_N^1 so that u_N^k is close to v_N^1).

With these notations, let us introduce the quantities

$$C_N = \left(\min\left(\frac{N^2}{N^2+1}, \widetilde{\beta_N^k} \right) - \frac{2\sqrt[4]{5}\|V + 3(u_N^k)^2 - \lambda_N^k\|_{L^2}}{N} \right) \tag{6.3.8}$$

with

$$\widetilde{\beta_N^k} = \frac{\tilde{\eta}_N^k}{\tilde{\eta}_N^k + \|(V + 3(u_N^k)^2 - \lambda_N^k)_-\|_{L^\infty} + 1} \quad \text{and} \quad \tilde{\eta}_N^k = \frac{1}{4}\min(\delta_N^k, 3). \tag{6.3.9}$$

We also define

$$I_N^k = \int (\nabla u_N^k)^2 + \int V(u_N^k)^2 + 3\int (u_N^k)^4 - \lambda_N^k \int (u_N^k)^2 \tag{6.3.10}$$

We can now prove the following lemma:

Lemma 6.3.1. *Let us assume that $k \in \mathbb{N}$ is large enough so that*

$$\|v_N^1 - u_N^k\|_{L^\infty} \leq \frac{1}{4}\min(1, \frac{\min(\delta_N^k, 3)}{4\|v_N^1\|_{L^\infty}}) \quad \text{and} \quad \lambda_N^k - \mu_N^1 \leq \frac{1}{2}\min(\delta_N^k, 1), \tag{6.3.11}$$

and that $N \in \mathbb{N}$ is large enough so that $C_N > 0$ (C_N defined in (6.3.8)). Then DF is a diffeomorphism, and, with the notations in (6.3.3):

$$\begin{aligned} \gamma &\leq I_N^k + C_N^{-1}\max\left(\frac{I_N^k}{2}, I_N^k C_N^{-1} + 1\right), \\ L(\alpha) &\leq \left(3\sqrt{5}(2+\alpha) + 4\right)\alpha. \end{aligned}$$

In addition, if $2\gamma L(2\gamma\epsilon) \leq 1$ (where we recall ε is defined in (6.3.4)), there exists a unique $(\tilde{u}, \tilde{\lambda})$ in the ball $B((u_N^k, \lambda_N^k), 2\gamma\epsilon)$ such that $F(\tilde{u}, \tilde{\lambda}) = 0$ and

$$\|\tilde{u} - u_N^k\|_{H^1} + |\tilde{\lambda} - \lambda_N^k| \leq 2\gamma\| - \Delta u_N^k + V u_N^k + (u_N^k)^3 - \lambda_N^k u_N^k\|_{H^{-1}} \tag{6.3.12}$$

and there exists a computable condition depending on $\|\tilde{u} - u_N^k\|_{H^1}$, $|\tilde{\lambda} - \lambda_N^k|$, u_N^k, $\lambda_N^k, \mu_N^1, \mu_N^2$ guaranteeing that $(\tilde{u}, \tilde{\lambda})$ is the ground state (u, λ) of (6.1.1).

Proof. Let us first prove that the operator

$$B_N^k = \Pi_N B^k \quad \text{where} \quad B^k = -\Delta + V + 3(u_N^k)^2 - \lambda_N^k \tag{6.3.13}$$

is elliptic, which will in turn allow to prove that $DF_{(u_N^k, \lambda_N^k)}$ is an isomorphism and also estimate the norm of $[DF_{(u_N^k, \lambda_N^k)}]^{-1}$.

Let us define the space $(v_N^1)^\perp = \{v \in X_N, \int_\Omega v_N^1 v = 0\}$. The operator $-\Delta + V + (u_N^k)^2 - \mu_N^1$ is positive definite over the space $(v_N^1)^\perp$. More precisely, we have:

$$\forall v \in (v_N^1)^\perp, \quad \langle(-\Delta + V + (u_N^k)^2 - \mu_N^1)v, v\rangle \geq (\mu_N^2 - \mu_N^1)\|v\|_{L^2}^2$$

We first note that we have:

$$\forall v \in (v_N^1)^\perp, \quad \langle(-\Delta + V + (u_N^k)^2 - \lambda_N^k)v, v\rangle \geq (\mu_N^2 - \lambda_N^k)\|v\|_{L^2}^2 \tag{6.3.14}$$

and

$$\forall v \in X_N, \quad \langle(-\Delta + V + (u_N^k)^2 - \lambda_N^k)v_N^1, v\rangle = (\mu_N^1 - \lambda_N^k)(v_N^1, v)_{L^2} \tag{6.3.15}$$

hence, for any $v \in X_N$, by decomposing v into a part in $(v_N^1)^\perp$ and a part collinear to v_N^1 : $v = [v - (v, v_N^1)v_N^1] + (v, v_N^1)v_N^1 \equiv v_\perp + v_{//}$

$$\begin{aligned}
\langle B_N^k v, v\rangle &= \langle(-\Delta + V + (u_N^k)^2 - \lambda_N^k)v, v\rangle + 2\int_\Omega (u_N^k)^2 v^2 \\
&= \langle(-\Delta + V + (u_N^k)^2 - \lambda_N^k)v_\perp, v_\perp\rangle + \langle(-\Delta + V + (u_N^k)^2 - \lambda_N^k)v_{//}, v_{//}\rangle \\
&\quad +2\langle(-\Delta + V + (u_N^k)^2 - \lambda_N^k)v_\perp, v_{//}\rangle + 2\int_\Omega (u_N^k)^2 v^2 \\
&\geq (\mu_N^2 - \lambda_N^k)\|v_\perp\|_{L^2}^2 + (\mu_N^1 - \lambda_N^k)\|v_{//}\|_{L^2}^2 + 2\int_\Omega (u_N^k)^2 v^2 \\
&\quad \text{using} \quad \mu_N^1 \leq \lambda_N^k \leq \mu_N^2 \quad \text{and} \quad (6.3.14), (6.3.15) \\
&\geq \frac{\delta_N^k}{2}\|v_\perp\|_{L^2}^2 - \frac{1}{2}\|v_{//}\|_{L^2}^2 + 2\int_\Omega (v_N^1)^2 v^2 - 2\int_\Omega (v_N^1 - u_N^k)(v_N^1 + u_N^k)v^2 \\
&\quad \text{using} \quad (6.3.11) \\
&\geq \frac{\delta_N^k}{2}\|v_\perp\|_{L^2}^2 + \frac{3}{2}\|v_{//}\|_{L^2}^2 + 2\left(\int_\Omega (v_N^1)^2 v^2 - [\int_\Omega (v_N^1 v)]^2\right) \\
&\quad -2\int_\Omega (v_N^1 - u_N^k)(v_N^1 + u_N^k)v^2 \\
&\geq \frac{1}{2}\min(\delta_N^k, 3)\|v\|_{L^2}^2 + 2[\int_\Omega (v_N^1 - u_N^k)^2 v^2 - 2\int_\Omega (v_N^1 - u_N^k)v_N^1 v^2] \\
&\geq \frac{1}{4}\min(\delta_N^k, 3)\|v\|_{L^2}^2 \quad \text{using} \quad (6.3.11).
\end{aligned}$$

Moreover, for all $v \in X_N$,

$$\langle(-\Delta + V + 3(u_N^k)^2 - \lambda_N^k)v, v\rangle \geq \|\nabla v\|_{L^2}^2 - \|(V + 3(u_N^k)^2 - \lambda_N^k)_-\|_{L^\infty}\|v\|_{L^2}^2$$

where $(V + 3(u_N^k)^2 - \lambda_N^k)_-$ is the negative part of $(V + 3(u_N^k)^2 - \lambda_N^k)$.
In conclusion for all $v \in X_N$,

$$\langle B_N^k v, v\rangle \geq \widetilde{\beta_N^k}\|v\|_{H^1}^2 \tag{6.3.16}$$

with $\widetilde{\beta_N^k}$ defined in (6.3.9), this constant being computable. Hence B_N^k is elliptic over X_N.
Let us prove the bijectivity of $DF_{(u_N^k, \lambda_N^k)}$. Let $f \in X', \mu \in \mathbb{R}$, we are looking for $w \in X, \tau \in \mathbb{R}$ such that

$$\begin{cases} -\Delta w + Vw + 3(u_N^k)^2 w - \lambda_N^k w - \tau u_N^k &= f \\ 2\int u_N^k w &= \mu \end{cases} \tag{6.3.17}$$

We first prove that there exists a computable constant $C_N > 0$, such that, for any $\varphi \in X'$, there exists $w(\varphi)$

$$-\Delta w(\varphi) + Vw(\varphi) + 3(u_N^k)^2 w(\varphi) - \lambda_N^k w(\varphi) = \varphi \quad \text{with} \quad \|w(\varphi)\|_{H^1} \leq C_N^{-1}\|\varphi\|_{H^{-1}} \tag{6.3.18}$$

To get this, (6.3.16) is not sufficient, it has to be extended from the discrete to the continuous

frame. $\forall v \in X$,

$$\begin{aligned}
\left|\int(\nabla v)^2+\int Vv^2 \;+\; 3\int(u_N^k)^2v^2-\lambda_N^k\int v^2\right| &= \left|\int\nabla(v-\Pi_N v)^2+\int\nabla(\Pi_N v)^2+\int(V+3(u_N^k)^2-\lambda_N^k)(\Pi_N v)^2\right.\\
&\quad\left.+\int(V+3(u_N^k)^2-\lambda_N^k)(v^2-(\Pi_N v)^2)\right|\\
&\geq \left|\int\nabla(v-\Pi_N v)^2\right|+\widetilde{\beta_N^k}\|\Pi_N v\|_{H^1}^2\\
&\quad -\|V+3(u_N^k)^2-\lambda_N^k\|_{L^2}\|v-\Pi_N v\|_{L^2}\|v+\Pi_N v\|_{L^\infty}\\
&\geq \left|\int\nabla(v-\Pi_N v)^2\right|+\widetilde{\beta_N^k}\|\Pi_N v\|_{H^1}^2\\
&\quad -\frac{2\sqrt[4]{5}}{N}\|V+3(u_N^k)^2-\lambda_N^k\|_{L^2}\|v\|_{H^1}^2 \quad \text{using (6.1.4) and (6.1.13)}\\
&\geq \frac{N^2}{N^2+1}\|v-\Pi_N v\|_{H^1}^2+\widetilde{\beta_N^k}\|\Pi_N v\|_{H^1}^2\\
&\quad -\frac{2\sqrt[4]{5}\|V+3(u_N^k)^2-\lambda_N^k\|_{L^2}}{N}\|v\|_{H^1}^2\\
&\geq \left(\min\left(\frac{N^2}{N^2+1},\widetilde{\beta_N^k}\right)-\frac{2\sqrt[4]{5}\|V+3(u_N^k)^2-\lambda_N^k\|_{L^2}}{N}\right)\|v\|_{H^1}^2
\end{aligned}$$

The constant C_N defined in (6.3.8) being computable and strictly positive for N large, we are able to determine N such that this bilinear form is elliptic. This form being clearly continuous, Lax-Milgram theorem holds and there exists a unique solution w to equation (6.3.18). It then follows that the solution to (6.3.17) is $w = w(f) + \tau w(u_N^k)$, and τ is the only value such that $2\int u_N^k[w(f)+\tau w(u_N^k)] = \mu$, which is possible, and achieved with

$$\tau = \frac{\frac{\mu}{2}-\int u_N^k w(f)}{\int u_N^k w(u_N^k)} \tag{6.3.19}$$

due to the fact that, from (6.3.18),

$$\int u_N^k w(u_N^k) = \int(\nabla w(u_N^k))^2+\int V(w(u_N^k))^2+3\int(u_N^k)^2(w(u_N^k))^2-\lambda_N^k\int(w(u_N^k))^2 \tag{6.3.20}$$

which is positive from the above estimates as soon as N is large enough. Hence, if the hypotheses are satisfied, DF is an isomorphism.

Let us now get a lower bound for γ. As

$$|\tau| \leq \frac{1}{|\int u_N^k w(u_N^k)|}\left(\frac{|\mu|}{2}+\|u_N^k\|_{L^2}\|w(f)\|_{L^2}\right) = \frac{1}{|\int u_N^k w(u_N^k)|}\left(\frac{|\mu|}{2}+\|w(f)\|_{L^2}\right),$$

we need to lower bound $|\int u_N^k w(u_N^k)|$. We first notice that

$$\begin{aligned}
1=\int[u_N^k]^2 &= \int\nabla u_N^k\nabla w(u_N^k)+\int Vu_N^k w(u_N^k)+3\int(u_N^k)^3w(u_N^k)-\lambda_N^k\int u_N^k w(u_N^k)\\
&\leq [\int(\nabla u_N^k)^2+\int V(u_N^k)^2+3\int(u_N^k)^4-\lambda_N^k\int(u_N^k)^2]^{1/2}[\int(\nabla w(u_N^k))^2\\
&\quad+\int V(w(u_N^k))^2+3\int(u_N^k)^2(w(u_N^k))^2-\lambda_N^k\int(w(u_N^k))^2]^{1/2}
\end{aligned} \tag{6.3.21}$$

from the Cauchy-Schwarz inequality since the operator on the right hand side is coercive, so that

$$\begin{aligned}\int(\nabla w(u_N^k))^2+\int V(w(u_N^k))^2+3\int(u_N^k)^2(w(u_N^k))^2-\lambda_N^k\int(w(u_N^k))^2\\ \geq\frac{1}{\int(\nabla u_N^k)^2+\int V(u_N^k)^2+3\int(u_N^k)^4-\lambda_N^k\int(u_N^k)^2}\end{aligned}$$

which is a quantity that can be computed exactly.

From the definition of I_N^k in (6.3.10), we get $|\tau|\leq I_N^k\left(\frac{|\mu|}{2}+\|w(f)\|_{L^2}\right)$. Hence, using (6.3.18),

$$\begin{aligned}\|w\|_{H^1} &\leq |\tau|\|w(u_N^k)\|_{H^1}+\|w(f)\|_{H^1}\\ &\leq I_N^k\left(\frac{|\mu|}{2}+\|w(f)\|_{L^2}\right)\|w(u_N^k)\|_{H^1}+\|w(f)\|_{H^1}\\ &\leq I_N^k\left(\frac{|\mu|}{2}+C_N^{-1}\|f\|_{H^{-1}}\right)C_N^{-1}\|u_N^k\|_{H^{-1}}+C_N^{-1}\|f\|_{H^{-1}}\\ &\leq C_N^{-1}\left(\frac{I_N^k}{2}|\mu|+\left(I_N^kC_N^{-1}+1\right)\|f\|_{H^{-1}}\right)\\ &\leq C_N^{-1}\max\left(\frac{I_N^k}{2},I_N^kC_N^{-1}+1\right)(|\mu|+\|f\|_{H^{-1}})\end{aligned}\tag{6.3.22}$$

hence

$$|\tau|+\|w\|_{H^1}\leq\left[I_N^k+C_N^{-1}\max\left(\frac{I_N^k}{2},I_N^kC_N^{-1}+1\right)\right](|\mu|+\|f\|_{H^{-1}})$$

which gives an upper bound for γ.

The only thing that remains is to evaluate the Lipschitz condition L again in (6.3.3). We have $\forall w\in X,\tau\in\mathbb{R}$,

$$[DF_{(u_N^k,\lambda_N^k)}-DF_{(u',\lambda')}](w,\tau)=\begin{pmatrix}3[(u_N^k)^2-(u')^2]w-\tau(u_N^k-u')-(\lambda_N^k-\lambda')w\\ 2\int(u_N^k-u')w\end{pmatrix}\tag{6.3.23}$$

hence, from the bounds

$$\begin{aligned}\int 3[(u_N^k)^2-(u')^2]ww' &- \tau(u_N^k-u')w'-(\lambda_N^k-\lambda')ww'\\ &\leq 3\sqrt{5}\|u_N^k-u'\|_{L^2}\|u_N^k+u'\|_{L^2}\|w\|_{H^1}\|w'\|_{H^1}\\ &\quad+|\tau|\|u_N^k-u'\|_{L^2}\|w'\|_{L^2}+|\lambda_N^k-\lambda'|\|w\|_{L^2}\|w'\|_{L^2}\end{aligned}$$

and

$$2\int(u_N^k-u')w\leq 2\|u_N^k-u'\|_{L^2}\|w\|_{L^2},\tag{6.3.24}$$

we easily derive that

$$L(\alpha)\leq[3\sqrt{5}(2+\alpha)+4]\alpha.\tag{6.3.25}$$

We can now provably check if $2\gamma L(2\gamma\epsilon)\leq 1$ which is the condition under which we know that there exists a unique solution $(\tilde{u},\tilde{\lambda})$ to the problem (6.1.9) in the ball $B((u_N^k,\lambda_N^k),2\gamma\epsilon)$ and (6.3.12) provides a first *a posteriori* estimate between a solution $(\tilde{u},\tilde{\lambda})$ to the problem (6.1.9) and the current iterate (u_N^k,λ_N^k).

The pair $(\tilde{u},\tilde{\lambda})$ is an eigenpair of the linear operator $A_{\tilde{u}}$, but may not be associated with the ground state of (6.1.1). In order to complete the last statement in the Lemma, we are going

to use Lemma 6.1.2. We are thus interested in the lowest eigenpair (v,μ) of the (symmetric) linear operator $A_{\tilde{u}}$. Let us assume by contradiction that $(v,\mu)\neq(\tilde{u},\tilde{\lambda})$, hence, in particular we have

$$(\tilde{u},v)=0,\quad \|v\|_{L^2}=\|\tilde{u}\|_{L^2}=1,\quad \mu<\tilde{\lambda}. \tag{6.3.26}$$

Recalling the definition of the linear, discrete eigenvalue problem (6.3.6), let us introduce the following decomposition of v :

$$v=w+(v-\Pi_N v)+(v_N^1,\Pi_N v)v_N^1,\qquad \text{with}\qquad w=\Pi_N v-(v_N^1,\Pi_N v)v_N^1.$$

Note that w is orthogonal to v_N^1. Hence

$$\begin{aligned}
\mu &= \langle A_{\tilde{u}}v,v\rangle=\langle A_{\tilde{u}}w,w\rangle+\langle A_{\tilde{u}}(v-w),v+w\rangle\\
&= \langle A_{u_N^k}w,w\rangle && (6.3.27)\\
&\quad +\langle A_{\tilde{u}}w,w\rangle-\langle A_{u_N^k}w,w\rangle && (6.3.28)\\
&\quad +\langle A_{\tilde{u}}(v-\Pi_N v),v+\Pi_N v\rangle && (6.3.29)\\
&\quad +2(\Pi_N v,v_N^1)\langle A_{\tilde{u}}v_N^1,\Pi_N v\rangle && (6.3.30)\\
&\quad -(v_N^1,\Pi_N v)^2\langle A_{\tilde{u}}v_N^1,v_N^1\rangle && (6.3.31)
\end{aligned}$$

In what follows we first prove that the L^2-norm of w is (measurably) close to 1 and then use the fact that w belongs to X_N and is orthogonal to v_N^1 to state that $\langle A_{u_N^k}w,w\rangle$ will be somehow larger than μ_N^2 and the other terms in the next lines above are all (measurably) small. The contradiction will then come from the existence of a measurable gap $\delta_N^k=\mu_N^2-\mu_N^1$ as stated in (6.3.7), and the fact that $\tilde{\lambda}$ is measurably close to λ_N^k hence close to μ_N^1.

Before estimating μ, we first bound the four following terms: $\|v_N^1-\tilde{u}\|_{L^2}$, $\|\tilde{u}^2-(u_N^k)^2\|_{L^\infty}$, $\|v\|_{H^1}$, and $\|v\|_{H^2}$, which we shall use later when analysing the terms composing μ.

- Firstly,
$$\|v_N^1-\tilde{u}\|_{L^2}\le\|v_N^1-u_N^k\|_{L^2}+\|u_N^k-\tilde{u}\|_{L^2} \tag{6.3.32}$$
- Secondly,
$$\|\tilde{u}^2-(u_N^k)^2\|_{L^\infty}^2\le\frac{1+\sqrt{5}}{2}\|\tilde{u}^2-(u_N^k)^2\|_{H^1}^2 \tag{6.3.33}$$

  Moreover,

$$\begin{aligned}
\|\tilde{u}^2-(u_N^k)^2\|_{H^1}^2 &= \int_\Omega(\tilde{u}-u_N^k)^2(\tilde{u}+u_N^k)^2\\
&\quad+\int_\Omega[(\tilde{u}-u_N^k)'(\tilde{u}+u_N^k)+(\tilde{u}-u_N^k)(\tilde{u}+u_N^k)']^2\\
&\le \|\tilde{u}-u_N^k\|_{L^\infty}^2\|\tilde{u}+u_N^k\|_{L^2}^2+2\|\tilde{u}+u_N^k\|_{L^\infty}^2\|(\tilde{u}-u_N^k)'\|_{L^2}^2\\
&\quad+2\|\tilde{u}-u_N^k\|_{L^\infty}^2\|(\tilde{u}+u_N^k)'\|_{L^2}^2\\
&\le 5\frac{1+\sqrt{5}}{2}\|\tilde{u}-u_N^k\|_{H^1}^2\|\tilde{u}+u_N^k\|_{H^1}^2\\
&\le 5\frac{1+\sqrt{5}}{2}\|\tilde{u}-u_N^k\|_{H^1}^2[\|\tilde{u}-u_N^k\|_{H^1}+2\|u_N^k\|_{H^1}]^2
\end{aligned}$$

  Hence

$$\|\tilde{u}^2-(u_N^k)^2\|_{L^\infty}^2\le 5\left(\frac{1+\sqrt{5}}{2}\right)^2\|\tilde{u}-u_N^k\|_{H^1}^2[\|\tilde{u}-u_N^k\|_{H^1}+2\|u_N^k\|_{H^1}]^2 \tag{6.3.34}$$

- Thirdly, using that $\mu < \tilde{\lambda}$, we derive

$$\begin{aligned}\|v\|_{H^1}^2 &= \mu - \int_\Omega Vv^2 - \int_\Omega \tilde{u}^2v^2 + \int_\Omega v^2 + \int_\Omega (u_N^k)^2v^2 - \int_\Omega (u_N^k)^2v^2 \\ &\leq \lambda_N^k + |\tilde{\lambda} - \lambda_N^k| + \|1 - V - (u_N^k)^2\|_{L^2}\|v\|_{L^2}\|v\|_{H^1} + \|\tilde{u}^2 - (u_N^k)^2\|_{L^\infty}\|v\|_{L^2}^2.\end{aligned}$$

Hence

$$\|v\|_{H^1}^2 - \|1 - V - (u_N^k)^2\|_{L^2}\|v\|_{H^1} - (\lambda_N^k + |\tilde{\lambda} - \lambda_N^k| + \|\tilde{u}^2 - (u_N^k)^2\|_{L^\infty}) \leq 0$$

from which we deduce that

$$\begin{aligned}\|v\|_{H^1} \leq \frac{1}{2}\|1 - V - (u_N^k)^2\|_{L^2} \\ + \frac{1}{2}\left(\|1 - V - (u_N^k)^2\|_{L^2}^2 + 4(\lambda_N^k + |\tilde{\lambda} - \lambda_N^k| + \|\tilde{u}^2 - (u_N^k)^2\|_{L^\infty})\right)^{\frac{1}{2}}.\end{aligned} \quad (6.3.35)$$

Hence $\|v\|_{H^1}$ is measurably bounded.

- Finally, let us estimate $\|v\|_{H^2}$:

$$\begin{aligned}\|\Delta v\|_{L^2}^2 &= \int_\Omega (\mu v - Vv - \tilde{u}^2v)^2 \\ &\leq \|\mu - V - \tilde{u}^2\|_{L^2}^2 \frac{1+\sqrt{5}}{2}\|v\|_{H^1}^2 \\ &\leq (\lambda_N^k + |\tilde{\lambda} - \lambda_N^k| + \|V + (u_N^k)^2\|_{L^2} + \|(u_N^k)^2 - \tilde{u}^2\|_{L^2})^2 \frac{1+\sqrt{5}}{2}\|v\|_{H^1}^2\end{aligned}$$

Hence

$$\|v\|_{H^2}^2 \leq \left[\frac{1+\sqrt{5}}{2}(\lambda_N^k + |\tilde{\lambda} - \lambda_N^k| + \|V + (u_N^k)^2\|_{L^2} + \|(u_N^k)^2 - \tilde{u}^2\|_{L^2})^2 + 2\right]\|v\|_{H^1}^2 \quad (6.3.36)$$

With (6.3.35), we deduce a computable estimation of $\|v\|_{H^2}$.

Note that the only noncomputable terms are $\|\tilde{u} - u_N^k\|_{H^1}$ and $|\tilde{\lambda} - \lambda_N^k|$ but these terms are computably bounded in (6.3.12).

Let us finish the proof starting with the estimate of the L^2-norm of w : we write

$$\begin{aligned}\|w\|_{L^2}^2 &= \|\Pi_N v - (v_N^1, \Pi_N v)v_N^1\|_{L^2}^2 \\ &= (v, \Pi_N v) - (v, v_N^1)^2 \\ &= 1 - \|v - \Pi_N v\|_{L^2}^2 - [(v, \tilde{u}) + (v, v_N^1 - \tilde{u})]^2 \\ &= 1 - \|v - \Pi_N v\|_{L^2}^2 - (v, v_N^1 - \tilde{u})^2\end{aligned}$$

where we have used (6.3.26) in the last line. Hence

$$1 - \frac{\|v\|_{H^1}^2}{N^2} - \|v_N^1 - \tilde{u}\|_{L^2}^2 \leq \|w\|_{L^2}^2 \leq 1 \quad (6.3.37)$$

We are now capable of estimating each of the five terms in equation (6.3.27) – (6.3.31).

- For the right-hand side in (6.3.27) we have

$$\begin{aligned}\langle A_{u_N^k} w, w\rangle &\geq \mu_N^2 \|w\|_{L^2}^2 \\ &\geq \mu_N^2 (1 - \frac{\|v\|_{H^1}^2}{N^2} - \|v_N^1 - \tilde{u}\|_{L^2}^2)\end{aligned}$$

- For the term in (6.3.28) we write

$$\begin{aligned}|\langle A_{\tilde{u}} w, w\rangle - \langle A_{u_N^k} w, w\rangle| &= |\int_\Omega (\tilde{u}^2 - (u_N^k)^2) w^2| \\ &\leq \|\tilde{u}^2 - (u_N^k)^2\|_{L^\infty} \|w\|_{L^2}^2 \\ &\leq \|\tilde{u}^2 - (u_N^k)^2\|_{L^\infty}\end{aligned}$$

- Let us define $C(A_{u_N^k})$ as the continuity constant of the operator $A_{u_N^k} = -\Delta + V + (u_N^k)^2$ from $H^1(\Omega)$ into its dual space. The third term of equation (6.3.29) can be bounded as follows:

$$\begin{aligned}|\langle A_{\tilde{u}}(v - \Pi_N v), v + \Pi_N v\rangle| &\leq |\langle A_{u_N^k}(v - \Pi_N v), v + \Pi_N v\rangle| \\ &\quad + |\int_\Omega ((u_N^k)^2 - \tilde{u}^2)(v - \Pi_N v)(v + \Pi_N v)| \\ &\leq C(A_{u_N^k}) \|v - \Pi_N v\|_{H^1} \|v + \Pi_N v\|_{H^1} \\ &\quad + \|(u_N^k)^2 - \tilde{u}^2\|_{L^\infty}^2 \|v - \Pi_N v\|_{L^2} \|v + \Pi_N v\|_{L^2} \\ &\leq \left(C(A_{u_N^k}) + \|(u_N^k)^2 - \tilde{u}^2\|_{L^\infty}^2\right) \|v - \Pi_N v\|_{H^1} \|v + \Pi_N v\|_{H^1} \\ &\leq \left(C(A_{u_N^k}) + \|(u_N^k)^2 - \tilde{u}^2\|_{L^\infty}^2\right) \frac{2}{N} \|v\|_{H^2} \|v\|_{H^1}\end{aligned}$$

- For the term in (6.3.30) we show that

$$\begin{aligned}|2(\Pi_N v, v_N^1)\langle A_{\tilde{u}} v_N^1, \Pi_N v\rangle| &\leq |2(v, v_N^1 - \tilde{u}) \left(\langle A_{u_N^k} v_N^1, \Pi_N v\rangle + \int_\Omega (\tilde{u}^2 - (u_N^k)^2) v_N^1 \Pi_N v\right)| \\ &\leq 2\|v_N^1 - \tilde{u}\|_{L^2} \left(C(A_{u_N^k}) + \|(u_N^k)^2 - \tilde{u}^2\|_{L^\infty}^2\right) \|v_N^1\|_{H^1} \|v\|_{H^1}\end{aligned}$$

- Bounding the last term in (6.3.31) is done as follows

$$\begin{aligned}|-(v_N^1, \Pi_N v)^2 \langle A_{\tilde{u}} v_N^1, v_N^1\rangle| &= |(v_N^1 - \tilde{u}, v)^2 \left(\langle A_{u_N^k} v_N^1, v_N^1\rangle + \int_\Omega (\tilde{u}^2 - (u_N^k)^2)(v_N^1)^2\right)| \\ &\leq \|v_N^1 - \tilde{u}\|_{L^2}^2 \left(C(A_{u_N^k}) + \|(u_N^k)^2 - \tilde{u}^2\|_{L^\infty}^2\right)\end{aligned}$$

Then gathering all the previous estimates provides a computable lower bound for μ that we call LB_N^k that is close to μ_N^2 when N and k are large : i.e.

$$\mu \geq LB_N^k = \mu_N^2 - \varepsilon(N, k) \tag{6.3.38}$$

where $\varepsilon(N, k)$ can be computed from the previous estimations of ((6.3.28),(6.3.29),(6.3.30), (6.3.31)) and $\varepsilon(N, k) \to 0$ as N and k go to infinity. Then, we deduce from (6.3.7) that

$$\mu \geq \mu_N^2 - \varepsilon(N, k) = \mu_N^1 + \delta_N^k - \varepsilon(N, k) = \tilde{\lambda} + \delta_N^k + (\mu_N^1 - \lambda_N^k) + (\lambda_N^k - \tilde{\lambda}) - \varepsilon(N, k). \tag{6.3.39}$$

Hence, as soon as the three computable quantities $\mu_N^1 - \lambda_N^k$, $\lambda_N^k - \tilde{\lambda}$ and $\varepsilon(N, k)$ are small enough, verified by the condition that they are, e.g., smaller than $\delta_N^k/6$, we shall get that $\mu > \tilde{\lambda}$, and the contradiction appears. □

Remark 6.3.2. *This approach provides some first information on the error between the exact and approximate solutions. However, as can be expected, this information is quite rough and is only used to indicate if we are in an asymptotic regime in which, the following analysis allows to propose a more accurate a posteriori error estimator. We illustrate in section 4 this feature on a series of numerical examples. In addition to this asymptotic behavior of the error, the main output from the previous analysis is to certify that the discretization u_N^k is indeed approximating the ground state of (6.1.1) .*

6.3.2 More accurate *a posteriori* estimate

In order to get a more accurate *a posteriori* estimate, as is classical, we work more carefully on the residual:

$$R_N^k = \left[F(u_N^k, \lambda_N^k)\right]_X = -\Delta u_N^k + V u_N^k + (u_N^k)^3 - \lambda_N^k u_N^k, \tag{6.3.40}$$

that evaluates in which sense the snapshot (u_N^k, λ_N^k) obtained after k iterations of the algorithm (6.1.18), (6.1.19), (6.1.20) in X_N fails to solve the problem (6.1.8) we look for.

As was said in the introduction to this paper, this global error between the exact and the approximated solutions stems from two main sources : (i) one is the finite dimension $2N+1$ of the Fourier space X_N, i.e. the discretization of the space X , (ii) the other one is the finite number of iterations k.

In order to estimate the error coming from the two sources, the discretization parameter N and the number of iterations k, we separate the global error into two components, each of them depending mainly on one parameter associated with the above sources of error. The discretization residual is based on the numerical scheme and hence can be naturally defined as

$$R_{disc} = -\Delta u_N^k + V u_N^k + (u_N^{k-1})^2 u_N^k - \|u_N^{k*}\|_{L^2}^{-1} \lambda_N^{k-1} u_N^{k-1}, \tag{6.3.41}$$

the quantity $\|R_{disc}\|_{H^{-1}}$ then measures the discretization error and depends on the finite dimension $(2N+1)$ of the Fourier space X_N on which we solve the problem.

The iteration residual is then defined such that $R_N^k = R_{disc} + R_{iter}$. Hence

$$R_{iter} = (u_N^k)^3 - (u_N^{k-1})^2 u_N^k - \lambda_N^k u_N^k + \|u_N^{k*}\|_{L^2}^{-1} \lambda_N^{k-1} u_N^{k-1}. \tag{6.3.42}$$

The quantity $\|R_{iter}\|_{H^{-1}}$ is the iteration error and depends mainly on the finite number of iterations k.

We now relate the error in the functional space X — which is here $\|u - u_N^k\|_{H^1}$ — to the error of this specific problem expressed by an upper bound of the global residual defined previously. Besides the bounds have to be *a posteriori* computable.

First we express the term $\displaystyle\int_\Omega \nabla(u - u_N^k) \cdot \nabla v$ that appears in the right-hand side of (6.3.1) with a maximum of *a posteriori* computable terms (i.e. contributions involving u_N^k and λ_N^k and not u nor λ). Then we deal with the remaining terms. Finally we gather everything and get the *a posteriori* estimates. From (6.1.9), we can write the following equalities, at least in

the distributional sense if the functions are not smooth enough.

$$\begin{aligned}
-\Delta(u-u_N^k) &= \lambda u - Vu - u^3 + \Delta u_N^k \\
&= \Delta u_N^k - Vu_N^k - (u_N^k)^3 + \lambda_N^k u_N^k + \lambda u - \lambda_N^k u_N^k + V(u_N^k - u) + (u_N^k)^3 - u^3 \\
&= -R_N^k + (\lambda - \lambda_N^k)u + \lambda_N^k(u - u_N^k) + V(u_N^k - u) \\
&\quad +(u_N^k - u)((u_N^k)^2 + uu_N^k + u^2) \\
&= -R_N^k + (\lambda - \lambda_N^k)u + \lambda_N^k(u - u_N^k) + V(u_N^k - u) \\
&\quad +3(u_N^k)^2(u_N^k - u) - (u_N^k - u)^2(2u_N^k + u).
\end{aligned}$$

Hence

$$\begin{aligned}
\left\|u - u_N^k\right\|_{H^1} &= \max_{v\in H^1_\#} \frac{\displaystyle\int_\Omega \nabla(u-u_N^k)\cdot\nabla v + \int_\Omega (u-u_N^k)v}{\|v\|_{H^1}} \\
&= \max_{v\in H^1_\#} \frac{1}{\|v\|_{H^1}} \Bigg[- \langle R_N^k, v\rangle_{X',X} + \int_\Omega (\lambda - \lambda_N^k)uv \\
&\quad +(\lambda_N^k + 1)\int_\Omega (u-u_N^k)v + \int_\Omega V(u_N^k - u)v \\
&\quad + \int_\Omega 3(u_N^k)^2(u_N^k - u)v - \int_\Omega (u_N^k - u)^2(2u_N^k + u)v \Bigg].
\end{aligned}$$

Let us now use the fact that the maximum in the first line above is achieved for $v = u - u_N^k$, the second term in the above right hand side then reads

$$(\lambda - \lambda_N^k)\int_\Omega u(u-u_N^k) = \frac{1}{2}(\lambda - \lambda_N^k)\int_\Omega (u-u_N^k)^2 \tag{6.3.43}$$

and part of the third reads

$$\int_\Omega \left((\lambda_N^k + 1) - V - 3(u_N^k)^2\right)(u-u_N^k)^2 \le \|(V + 3(u_N^k)^2 - \lambda_N^k - 1)_-\|_{L^\infty}\int_\Omega (u-u_N^k)^2, \tag{6.3.44}$$

where $(V+3(u_N^k)^2-\lambda_N^k-1)_-$ stands for the negative part of $V+3(u_N^k)^2-\lambda_N^k-1$. By summing up the above expressions, we get

$$\begin{aligned}
\|u - u_N^k\|_{H^1} &\le \|R_N^k\|_{H^{-1}} + \frac{1}{2}|\lambda_N^k - \lambda|\|u - u_N^k\|_{L^2} + \sqrt{\frac{1+\sqrt{5}}{2}}\|u_N^k - u\|_{L^2}^2\|2u_N^k + u\|_{L^\infty} \\
&\quad +\|(V + 3(u_N^k)^2 - \lambda_N^k - 1)_-\|_{L^\infty}\frac{\displaystyle\int_\Omega (u-u_N^k)^2}{\|u-u_N^k\|_{H^1}}
\end{aligned} \tag{6.3.45}$$

In this expression the only term in the right hand-side that is neither *a posteriori* computable nor small compared to the left hand-side is $\frac{\int_\Omega (u-u_N^k)^2}{\|u-u_N^k\|_{H^1}}$. In order to bound this quantity, let us introduce $\tilde\Psi_v \in (v_N^1)^\perp$ as being the solution of $\forall w_N \in (v_N^1)^\perp$, $\left\langle \widetilde{B_N^k}\tilde\Psi_v, w_N\right\rangle = \int v w_N$ where $\widetilde{B_N^k} = \Pi_N \widetilde{B^k}$ and $\widetilde{B^k} = -\Delta + V + 3(u_N^k)^2 - \mu_N^1$, i.e.

$$\forall w_N \in (v_N^1)^\perp, \quad \int \nabla\tilde\Psi_v\nabla w_N + \int V\tilde\Psi_v w_N + \int 3(u_N^k)^2\tilde\Psi_v w_N - \mu_N^1\int \tilde\Psi_v w_N = \int v w_N \tag{6.3.46}$$

We then define $\widetilde{u_N^k - \Pi_N u} = [u_N^k - \Pi_N u] - (u_N^k - \Pi_N u, v_N^1)v_N^1$ that belongs to $(v_N^1)^\perp$. Note that $(v_N^1)^\perp \subset X_N$ so that

$$\|u - u_N^k\|_{L^2}^2 = \int (u_N^k - u)(\widetilde{u_N^k - \Pi_N u}) + \int (u_N^k - u)\delta u_N^k \tag{6.3.47}$$

with $\delta u_N^k = (u_N^k - \Pi_N u) - (\widetilde{u_N^k - \Pi_N u}) - (u - \Pi_N u) = (u_N^k - \Pi_N u, v_N^1)v_N^1 - (u - \Pi_N u)$. In order to bound the second term of (6.3.47), we first state from (6.1.13) that

$$\|u - \Pi_N u\|_{H^{-1}} = \|(u - u_N^k) - \Pi_N(u - u_N^k)\|_{H^{-1}} \leq \frac{1}{N^2}\|u - u_N^k\|_{H^1}$$

Second

$$(u_N^k - \Pi_N u, v_N^1) = (u_N^k - v_N^1, v_N^1) - (u - v_N^1, v_N^1) = -\frac{1}{2}\|u_N^k - v_N^1\|_{L^2}^2 + \frac{1}{2}\|u - v_N^1\|_{L^2}^2$$

so

$$|(u_N^k - \Pi_N u, v_N^1)| \leq \frac{3}{2}\|u_N^k - v_N^1\|_{L^2}^2 + \|u - u_N^k\|_{L^2}^2 \tag{6.3.48}$$

Hence

$$|\int (u_N^k - u)\delta u_N^k| \leq \|u - u_N^k\|_{L^2}\left(\frac{3}{2}\|u_N^k - \nu_N^1\|_{L^2}^2 + \|u_N^k - u\|_{L^2}^2\right) + \frac{1}{N^2}\|u_N^k - u\|_{H^1}^2 \tag{6.3.49}$$

Let us now focus on the first term of (6.3.47).

$$\begin{aligned}
\int (u_N^k - u)(\widetilde{u_N^k - \Pi_N u}) &= \langle \widetilde{B_N^k}\tilde{\Psi}_{u_N^k - u}, \widetilde{u_N^k - \Pi_N u}\rangle \\
&= \langle \widetilde{B^k}\tilde{\Psi}_{u_N^k - u}, \widetilde{u_N^k - \Pi_N u}\rangle \\
&= \langle \tilde{\Psi}_{u_N^k - u}, \widetilde{B^k}(\widetilde{u_N^k - \Pi_N u})\rangle \\
&= \langle \tilde{\Psi}_{u_N^k - u}, \widetilde{B^k}(u_N^k - \Pi_N u)\rangle - \langle u_N^k - \Pi_N u, \nu_N^1\rangle\langle \tilde{\Psi}_{u_N^k - u}, \widetilde{B^k}\nu_N^1\rangle \\
&= \langle \tilde{\Psi}_{u_N^k - u}, \widetilde{B^k}(u_N^k - u)\rangle + \langle \tilde{\Psi}_{u_N^k - u}, (V + 3(u_N^k)^2)(u - \Pi_N u)\rangle \\
&\quad -\langle u_N^k - \Pi_N u, \nu_N^1\rangle\langle \tilde{\Psi}_{u_N^k - u}, \widetilde{B^k}\nu_N^1\rangle \\
&= \int (A_{u_N^k} - \lambda_N^k)u_N^k\tilde{\Psi}_{u_N^k - u} + \int [2(u_N^k)^2 + (\lambda_N^k - \mu_N^1)]u_N^k\tilde{\Psi}_{u_N^k - u} \\
&\quad - \int (A_u - \lambda)u\tilde{\Psi}_{u_N^k - u} + \int (u^2 - 3(u_N^k)^2)u\tilde{\Psi}_{u_N^k - u} \\
&\quad -(\lambda - \mu_N^1)\int u\tilde{\Psi}_{u_N^k - u} + \langle \tilde{\Psi}_{u_N^k - u}, (V + 3(u_N^k)^2)(u - \Pi_N u)\rangle \\
&\quad -\langle u_N^k - \Pi_N u, \nu_N^1\rangle\langle \tilde{\Psi}_{u_N^k - u}, \widetilde{B^k}\nu_N^1\rangle \\
&= (\Pi_N(R_N^k), \tilde{\Psi}_{u_N^k - u}) + \int [2(u_N^k)^3 + u^3 - 3(u_N^k)^2 u]\tilde{\Psi}_{u_N^k - u} \\
&\quad + \int [(\lambda_N^k - \mu_N^1)(u_N^k - v_N^1) - (\lambda - \mu_N^1)(u - v_N^1)]\tilde{\Psi}_{u_N^k - u} \\
&\quad +\langle \tilde{\Psi}_{u_N^k - u}, (V + 3(u_N^k)^2)(u - \Pi_N u)\rangle \\
&\quad -\langle u_N^k - \Pi_N u, \nu_N^1\rangle\langle \tilde{\Psi}_{u_N^k - u}, \widetilde{B^k}\nu_N^1\rangle
\end{aligned}$$

where, in the last but one line above, we have used the fact that $\tilde{\Psi}_{u_N^k-u} \in (v_N^1)^\perp$.
Let us bound each term of the right-hand side. For the first term we have

$$|(\Pi_N(R_N^k), \tilde{\Psi}_{u_N^k-u})| \leq \|\Pi_N R_N^k\|_{H^{-1}} \|\tilde{\Psi}_{u_N^k-u}\|_{H^1} \leq \frac{1}{\beta_N^k} \|\Pi_N R_N^k\|_{H^{-1}} \|u_N^k - u\|_{L^2}$$

where β_N^k is such that $\|\tilde{\Psi}_{u_N^k-u}\|_{H^1} \leq \frac{1}{\beta_N^k}\|u_N^k - u\|_{L^2}$.

In particular, since $\tilde{\Psi}_{u_N^k-u} \in (v_N^1)^\perp$, we can use the fact that

$$\forall \Psi \in (v_N^1)^\perp, \quad \langle \widetilde{B_N^k}\Psi, \Psi\rangle \geq (\mu_N^2 - \mu_N^1)\|\Psi\|_{L^2}^2,$$

and

$$\forall \Psi \in (v_N^1)^\perp, \quad \langle \widetilde{B_N^k}\Psi, \Psi\rangle \geq \|\Psi\|_{H^1}^2 - \|(V + (u_N^k)^2 - \mu_N^1 - 1)_-\|_{L^\infty}\|\Psi\|_{L^2}^2$$

to show that

$$\forall \Psi \in (v_N^1)^\perp, \quad \langle \widetilde{B_N^k}\Psi, \Psi\rangle \geq \frac{\mu_N^2 - \mu_N^1}{\mu_N^2 - \mu_N^1 + \|(V + (u_N^k)^2 - \mu_N^1 - 1)_-\|_{L^\infty}}\|\Psi\|_{H^1}^2$$

and so

$$\forall \Psi \in (v_N^1)^\perp, \quad \langle \widetilde{B_N^k}\Psi, \Psi\rangle \geq \frac{\mu_N^2 - \mu_N^1}{\sqrt{\mu_N^2 - \mu_N^1 + \|(V + (u_N^k)^2 - \mu_N^1 - 1)_-\|_{L^\infty}}}\|\Psi\|_{H^1}\|\Psi\|_{L^2}$$

Therefore we can take $\beta_N^k = \dfrac{\mu_N^2 - \mu_N^1}{\sqrt{\mu_N^2 - \mu_N^1 + \|(V + (u_N^k)^2 - \mu_N^1 - 1)_-\|_{L^\infty}}}$.

For the second term we have

$$\begin{aligned}
\int [2(u_N^k)^3 + u^3 - 3(u_N^k)^2 u]\tilde{\Psi}_{u_N^k-u} &= \int (u - u_N^k)^2(u + 2u_N^k)\tilde{\Psi}_{u_N^k-u} \\
&\leq \|u - u_N^k\|_{L^\infty}^2 \|u + 2u_N^k\|_{L^2}\|\tilde{\Psi}_{u_N^k-u}\|_{L^2} \\
&\leq \frac{1}{\beta_N^k}\|u - u_N^k\|_{L^\infty}^2 \|u + 2u_N^k\|_{L^2}\|u - u_N^k\|_{L^2}
\end{aligned}$$

For the third term we have

$$\begin{aligned}
&\int [(\lambda_N^k - \mu_N^1)(u_N^k - v_N^1) - (\lambda - \mu_N^1)(u - v_N^1)]\tilde{\Psi}_{u_N^k-u} \\
&\leq \frac{1}{\beta_N^k}\|u - u_N^k\|_{L^2}\left[|\lambda_N^k - \mu_N^1|\|u_N^k - v_N^1\|_{L^2} + |\lambda - \mu_N^1|\|u - v_N^1\|_{L^2}\right] \\
&\leq \frac{1}{\beta_N^k}\|u - u_N^k\|_{L^2}\big[2|\lambda_N^k - \mu_N^1|\|u_N^k - v_N^1\|_{L^2} + |\lambda_N^k - \mu_N^1|\|u_N^k - u\|_{L^2} \\
&+ |\lambda_N^k - \lambda|\|u_N^k - v_N^1\|_{L^2} + |\lambda_N^k - \lambda|\|u_N^k - u\|_{L^2}\big]
\end{aligned}$$

For the fourth term we have on the one hand

$$\begin{aligned}
|\langle \tilde{\Psi}_{u_N^k-u}, (V + 3(u_N^k)^2)(u - \Pi_N u)\rangle| &\leq \|\tilde{\Psi}_{u_N^k-u}\|_{L^\infty}\|V + 3(u_N^k)^2\|_{L^2}\|u - \Pi_N u\|_{L^2} \\
&\leq 5^{\frac{1}{4}}\|\tilde{\Psi}_{u_N^k-u}\|_{H^1}\|V + 3(u_N^k)^2\|_{L^2}\|u - \Pi_N u\|_{L^2}
\end{aligned}$$

and on the other hand, if V belongs to H^1, we have

$$\begin{aligned}|\langle\tilde{\Psi}_{u_N^k-u},(V+3(u_N^k)^2)(u-\Pi_N u)\rangle| &\leq \|\tilde{\Psi}_{u_N^k-u}(V+3(u_N^k)^2)\|_{H^1}\|u-\Pi_N u\|_{H^{-1}}\\ &\leq \sqrt{2\sqrt{5}}\|\tilde{\Psi}_{u_N^k-u}\|_{H^1}\|V+3(u_N^k)^2\|_{H^1}\|u-\Pi_N u\|_{H^{-1}}.\end{aligned}$$

Then, by an interpolation argument in Hilbert spaces we get depending on the regularity of V for $0\leq s\leq 1$,

$$\begin{aligned}|\langle\tilde{\Psi}_{u_N^k-u},(V+3(u_N^k)^2)(u-\Pi_N u)\rangle| \leq& (2\sqrt{5})^{\frac{s}{2}}5^{\frac{1-s}{4}}\|\tilde{\Psi}_{u_N^k-u}\|_{H^1}\|V+3(u_N^k)^2\|_{H^s}\|u-\Pi_N u\|_{H^{-s}}\\ \leq& \frac{C_s}{N^{1+s}}\|\tilde{\Psi}_{u_N^k-u}\|_{H^1}\|V+3(u_N^k)^2\|_{H^s}\|u-u_N^k\|_{H^1}\end{aligned}$$

with $C_s=(2\sqrt{5})^{\frac{s}{2}}5^{\frac{1-s}{4}}$.

For the last term, thanks to (6.3.48) and the fact that $B\nu_N^1=2(u_N^k)^2\nu_N^1$ we have

$$\begin{aligned}|\langle u_N^k-\Pi_N u,\nu_N^1\rangle\langle\tilde{\Psi}_{u_N^k-u},B\nu_N^1\rangle| &\leq \frac{1}{\beta_N^k}\|u-u_N^k\|_{L^2}\|2(u_N^k)^2\nu_N^1\|_{H^{-1}}\\ &\quad\times\left(\frac{3}{2}\|u_N^k-\nu_N^1\|_{L^2}^2+\|u-u_N^k\|_{L^2}^2\right)\end{aligned}$$

By summing up the five above bounds, we derive

$$\begin{aligned}|\int(u_N^k-u)(\widetilde{u_N^k-\Pi_N u})| \leq\ & \frac{1}{\beta_N^k}\|\Pi_N R_N^k\|_{H^{-1}}\|u_N^k-u\|_{L^2}\\ &+\frac{1}{\beta_N^k}\|u-u_N^k\|_{L^\infty}^2\|u+2u_N^k\|_{L^2}\|u-u_N^k\|_{L^2}\\ &+\frac{1}{\beta_N^k}\|u-u_N^k\|_{L^2}\big[2|\lambda_N^k-\mu_N^1|\|u_N^k-v_N^1\|_{L^2}\\ &+|\lambda_N^k-\mu_N^1|\|u_N^k-u\|_{L^2}+|\lambda_N^k-\lambda|\|u_N^k-v_N^1\|_{L^2}\\ &+|\lambda_N^k-\lambda|\|u_N^k-u\|_{L^2}\big]\\ &+\frac{C_s}{\beta_N^k}\frac{1}{N^{1+s}}\|u_N^k-u\|_{L^2}\|V+3(u_N^k)^2\|_{H^s}\|u-u_N^k\|_{H^1}\\ &+\frac{1}{\beta_N^k}\|u-u_N^k\|_{L^2}\|2(u_N^k)^2\nu_N^1\|_{H^{-1}}\\ &\times\left(\frac{3}{2}\|u_N^k-\nu_N^1\|_{L^2}^2+\|u-u_N^k\|_{L^2}^2\right)\end{aligned}$$

Finally, from (6.3.47), (6.3.49) and the above inequality, we get the following estimate :

$$\begin{aligned}\frac{\|u-u_N^k\|_{L^2}^2}{\|u-u_N^k\|_{H^1}} \leq\ & \frac{1}{\beta_N^k}\|\Pi_N R_N^k\|_{H^{-1}}+\frac{\sqrt{5}}{\beta_N^k}\|u-u_N^k\|_{H^1}^2\|u+2u_N^k\|_{L^2}\\ &+\frac{1}{\beta_N^k}\big[2|\lambda_N^k-\mu_N^1|\|u_N^k-v_N^1\|_{L^2}+|\lambda_N^k-\mu_N^1|\|u_N^k-u\|_{L^2}\\ &+|\lambda_N^k-\lambda|\|u_N^k-v_N^1\|_{L^2}+|\lambda_N^k-\lambda|\|u_N^k-u\|_{L^2}\big]\\ &+\frac{C_s}{\beta_N^k}\frac{1}{N^{1+s}}\|u_N^k-u\|_{L^2}\|V+3(u_N^k)^2\|_{H^s}\\ &+\frac{1}{\beta_N^k}\|2(u_N^k)^2\nu_N^1\|_{H^{-1}}\left(\frac{3}{2}\|u_N^k-\nu_N^1\|_{L^2}^2+\|u-u_N^k\|_{L^2}^2\right)\\ &+\left(\frac{3}{2}\|u_N^k-\nu_N^1\|_{L^2}^2+\|u_N^k-u\|_{L^2}^2\right)+\frac{1}{N^2}\|u_N^k-u\|_{H^1}. \qquad (6.3.50)\end{aligned}$$

Final step: From (6.3.45) and (6.3.50) we write that

$$\begin{aligned}\|u-u_N^k\|_{H^1} \;\leq\; & \|R_N^k\|_{H^{-1}}+\frac{1}{2}|\lambda_N^k-\lambda|\|u-u_N^k\|_{L^2}+\|u_N^k-u\|_{L^2}^2\|2u_N^k+u\|_{L^\infty}\\ & +\|(V+3(u_N^k)^2-\lambda_N^k-1)_-\|_{L^\infty}\Bigg[\frac{1}{\beta_N^k}\|\Pi_N R_N^k\|_{H^{-1}}\\ & +\frac{\sqrt{5}}{\beta_N^k}\|u-u_N^k\|_{H^1}^2\|u+2u_N^k\|_{L^2}\\ & +\frac{1}{\beta_N^k}\big[2|\lambda_N^k-\mu_N^1|\|u_N^k-v_N^1\|_{L^2}+|\lambda_N^k-\mu_N^1|\|u_N^k-u\|_{L^2}\\ & +|\lambda_N^k-\lambda|\|u_N^k-v_N^1\|_{L^2}+|\lambda_N^k-\lambda|\|u_N^k-u\|_{L^2}\big]\\ & +\frac{C_s}{\beta_N^k}\frac{1}{N^{1+s}}\|u_N^k-u\|_{L^2}\|V+3(u_N^k)^2\|_{H^s}\\ & +\frac{1}{\beta_N^k}\|2(u_N^k)^2\nu_N^1\|_{H^{-1}}\left(\frac{3}{2}\|u_N^k-\nu_N^1\|_{L^2}^2+\|u-u_N^k\|_{L^2}^2\right)\\ & +\left(\frac{3}{2}\|u_N^k-\nu_N^1\|_{L^2}^2+\|u_N^k-u\|_{L^2}^2\right)+\frac{1}{N^2}\|u_N^k-u\|_{H^1}\Bigg]\end{aligned}$$

Following the same lines as in the *a priori* analysis (see (6.2.25)) and using (6.1.20) instead of (6.1.17) we can write

$$\begin{aligned}\lambda_N^k-\lambda &= \langle(A_u-\lambda)(u_N^k-u),(u_N^k-u)\rangle_{X',X}+\int_\Omega(u_N^k)^2(u_N^k+u)(u_N^k-u)\\ &= \langle(A_u-\lambda)(u_N^k-u),(u_N^k-u)\rangle_{X',X}+\int_\Omega 2(u_N^k)^3(u_N^k-u)-\int_\Omega(u_N^k)^2(u_N^k-u)^2.\end{aligned}$$

From (6.2.1) we get

$$|\lambda-\lambda_N^k|\leq M_1\|u-u_N^k\|_{H^1}^2+2\|(u_N^k)^3\|_{L^2}\|u-u_N^k\|_{L^2}+\|u_N^k\|_{L^\infty}^2\|u-u_N^k\|_{L^2}^2. \tag{6.3.51}$$

Using (6.3.51), we get

$$\begin{aligned}\|u-u_N^k\|_{H^1}\Bigg(1-\frac{1}{2}|\lambda_N^k-\lambda|-\|u_N^k-u\|_{L^2}\|2u_N^k+u\|_{L^\infty}&\\ -\|(V+3(u_N^k)^2-\lambda_N^k-1)_-\|_{L^\infty}\left[\frac{\sqrt{5}}{\beta_N^k}\|u-u_N^k\|_{H^1}\|u+2u_N^k\|_{L^2}\right.&\\ +\frac{1}{\beta_N^k}\big[|\lambda_N^k-\mu_N^1|+\|u_N^k-v_N^1\|_{L^2}(M_1\|u-u_N^k\|_{H^1}+2\|(u_N^k)^3\|_{L^2}+\|u_N^k\|_{L^\infty}^2\|u-u_N^k\|_{L^2})&\\ \left.+|\lambda_N^k-\lambda|\big]+\|u_N^k-u\|_{L^2}\left(1+\frac{\|2(u_N^k)^2\nu_N^1\|_{H^{-1}}}{\beta_N^k}\right)+\frac{1}{N^2}+\frac{C_s}{\beta_N^k}\frac{1}{N^{1+s}}\|V+3(u_N^k)^2\|_{H^s}\right]\Bigg)&\\ \leq\|R_N^k\|_{H^{-1}}+\|(V+3(u_N^k)^2-\lambda_N^k-1)_-\|_{L^\infty}\left[\frac{1}{\beta_N^k}\|\Pi_N R_N^k\|_{H^{-1}}\right.&\\ \left.+\frac{2}{\beta_N^k}|\lambda_N^k-\mu_N^1|\|u_N^k-v_N^1\|_{L^2}+\frac{3}{2}\|u_N^k-\nu_N^1\|_{L^2}^2\left(1+\frac{\|2(u_N^k)^2\nu_N^1\|_{H^{-1}}}{\beta_N^k}\right)\right]&\end{aligned}$$

Since $\|u-u_N^k\|_{H^1}\underset{N,k\to\infty}{\to}0$, $|\lambda-\lambda_N^k|\underset{N,k\to\infty}{\to}0$, $|\lambda_N^k-\mu_N^1|\underset{N,k\to\infty}{\to}0$, $\|u_N^k-v_N^1\|_{L^2}\underset{N,k\to\infty}{\to}0$, for any $\alpha>1$ we can identify, thanks to the analysis in subsection 6.3.1, two quantities : $N_\alpha\in\mathbb{N}$

and $k_\alpha \in \mathbb{N}$ such that for any $N \geq N_\alpha$, and for any $k \geq k_\alpha$,

$$\frac{1}{2}|\lambda_N^k-\lambda|+\|u_N^k-u\|_{L^2}\|2u_N^k+u\|_{L^\infty}+\|(V+3(u_N^k)^2-\lambda_N^k-1)_-\|_{L^\infty}\left[\frac{\sqrt{5}}{\beta_N^k}\|u-u_N^k\|_{H^1}\|u+2u_N^k\|_{L^2}\right.$$
$$+\frac{1}{\beta_N^k}\big[|\lambda_N^k-\mu_N^1|+\|u_N^k-v_N^1\|_{L^2}(M_1\|u-u_N^k\|_{H^1}+2\|(u_N^k)^3\|_{L^2}+\|u_N^k\|_{L^\infty}^2\|u-u_N^k\|_{L^2})$$
$$\left.+|\lambda_N^k-\lambda|\big]+\|u_N^k-u\|_{L^2}\left(1+\frac{\|2(u_N^k)^2\nu_N^1\|_{H^{-1}}}{\beta_N^k}\right)+\frac{1}{N^2}+\frac{C_s}{\beta_N^k}\frac{1}{N^{1+s}}\|V+3(u_N^k)^2\|_{H^s}\right] \leq 1-\frac{1}{\alpha} \tag{6.3.52}$$

Note that the contribution $\frac{C_s}{\beta_N^k}\frac{1}{N^{1+s}}\|V+3(u_N^k)^2\|_{H^s}$ tends to 0 when $N \to +\infty$ for $V \in L^2(\Omega)$, that is $s=0$ but is smaller with more regularity. Then we have for $N \geq N_\alpha$ and $k \geq k_\alpha$

$$\begin{aligned}\frac{1}{\alpha}\|u-u_N^k\|_{H^1} \;\leq\; & \|R_N^k\|_{H^{-1}}+\|(V+3(u_N^k)^2-\lambda_N^k-1)_-\|_{L^\infty}\left[\frac{1}{\beta_N^k}\|\Pi_N R_N^k\|_{H^{-1}}\right.\\ & \left.+\frac{2}{\beta_N^k}|\lambda_N^k-\mu_N^1|\|u_N^k-v_N^1\|_{L^2}+\frac{3}{2}\|u_N^k-\nu_N^1\|_{L^2}^2\left(1+\frac{\|2(u_N^k)^2\nu_N^1\|_{H^{-1}}}{\beta_N^k}\right)\right]\end{aligned}$$

Let us now notice that, in the limit we can choose α as close as we wish to 1 (at the price of choosing N_α and k_α large enough).

Theorem 6.3.1. *Let $\alpha > 1$. Let $N \geq N_\alpha$ and $k \geq k_\alpha$ such that the inequality (6.3.52) is verified. Then the following a posteriori error bound holds:*

$$\begin{aligned}\|u-u_N^k\|_{H^1} \;\leq\; & \alpha\Big(\|R_N^k\|_{H^{-1}}+\|(V+3(u_N^k)^2-\lambda_N^k-1)_-\|_{L^\infty}\left[\frac{1}{\beta_N^k}\|\Pi_N R_N^k\|_{H^{-1}}\right.\\ & \left.+\frac{2}{\beta_N^k}|\lambda_N^k-\mu_N^1|\|u_N^k-v_N^1\|_{L^2}+\frac{3}{2}\|u_N^k-\nu_N^1\|_{L^2}^2\left(1+\frac{\|2(u_N^k)^2\nu_N^1\|_{H^{-1}}}{\beta_N^k}\right)\right]\Big)\end{aligned} \tag{6.3.53}$$

This expression shows the *a posteriori* relation between the error in H^1-norm and the global residual R_N^k.

We can now split the global residual into its two components in order to make the iteration and discretization errors explicit. From the definition of the discretization and iteration residuals (6.3.41) and (6.3.42) we write (with $\alpha > 1$)

$$\begin{aligned}\|u-u_N^k\|_{H^1} \;\leq\; & \alpha\Big(\|R_{disc}+R_{iter}\|_{H^{-1}}\\ & +\|(V+3(u_N^k)^2-\lambda_N^k-1)_-\|_{L^\infty}\left[\frac{1}{\beta_N^k}\|\Pi_N(R_{disc}+R_{iter})\|_{H^{-1}}\right.\\ & \left.+\frac{2}{\beta_N^k}|\lambda_N^k-\mu_N^1|\|u_N^k-v_N^1\|_{L^2}+\frac{3}{2}\|u_N^k-\nu_N^1\|_{L^2}^2\left(1+\frac{\|2(u_N^k)^2\nu_N^1\|_{H^{-1}}}{\beta_N^k}\right)\right]\Big)\end{aligned}$$

As the discretization residual R_{disc} has been defined from the numerical scheme described in the introduction, the projection of the residue on X_N is zero, that is $\Pi_N(R_{disc})=0$. Then, for any $\alpha > 1$, as close to 1 as we wish when the convergence is acknowledged, we have the

following decoupled upper bound

$$\begin{aligned}\|u-u_N^k\|_{H^1} \quad \leq \quad & \alpha\Bigg(\|R_{disc}\|_{H^{-1}}+\|R_{iter}\|_{H^{-1}}\left(1+\frac{\|(V+3(u_N^k)^2-\lambda_N^k-1)_-\|_{L^\infty}}{\beta_N^k}\right)\\ & +\|(V+3(u_N^k)^2-\lambda_N^k-1)_-\|_{L^\infty}\Bigg[\frac{2}{\beta_N^k}|\lambda_N^k-\mu_N^1|\|u_N^k-v_N^1\|_{L^2}\\ & +\frac{3}{2}\|u_N^k-\nu_N^1\|_{L^2}^2\left(1+\frac{\|2(u_N^k)^2\nu_N^1\|_{H^{-1}}}{\beta_N^k}\right)\Bigg]\Bigg). \end{aligned} \tag{6.3.54}$$

Thus the H^1-error can be *a posteriori* bounded by fully computable terms. Note that for planewave discretization the norm H^{-1} is computable.

At this point, let us remark that in all the performed numerical simulations, the term $\|(V+3(u_N^k)^2-\lambda_N^k-1)_-\|_{L^\infty}$ is zero, simplifying a lot the *a posteriori* bounds found in the analysis. We have not found any straightforward theoretical reason for it nevertheless we think that this is an interesting enough result *per se* and state

Corollary 6.3.1. *If $\|(V+3(u_N^k)^2-\lambda_N^k-1)_-\|_{L^\infty}=0$, we have the following error estimate, valid for any $\alpha>1$ and as close to 1 as we wish as the error goes to zero.*

$$\|u-u_N^k\|_{H^1}\leq\alpha\|R_N^k\|_{H^{-1}}. \tag{6.3.55}$$

This error estimate can be split into two parts

$$\|u-u_N^k\|_{H^1}\leq\alpha\Big(\|R_{disc}\|_{H^{-1}}+\|R_{iter}\|_{H^{-1}}\Big). \tag{6.3.56}$$

6.4 Numerical results

In this section we gather some numerical results that illustrate the statements proven above, in particular the *a posteriori* analysis of the nonlinear eigenvalue problem (6.1.9). First we provide some results on the properties of the "inverse power" method (6.1.18), (6.1.19), (6.1.20). Then we show that the numerical results are coherent with the *a posteriori* analysis and that the separation of the error components is relevant. We show that the crude a posteriori estimator of subsection 6.3.1 is reliable and converges but provides a much too large estimate. The strong reliability of this first estimator allows to certify the domain of validity for the second, sharper, estimator that we have presented in section 6.3.2. We also study the influence of the potential regularity on the H^1-convergence and the effect of the variation of the nonlinearity on the convergence of the algorithm used to solve the problem.

In the next subsections 6.4.2, 6.4.3 and 6.4.4, we evaluate the *a posteriori* estimators found in the previous section and perform numerical simulations with a potential V given by its Fourier coefficients

$$\widehat{V}_k=-\frac{1}{\sqrt{2\pi}}\frac{1}{|k|^4-\frac{1}{4}}, \tag{6.4.1}$$

from which we deduce that $V\in L^\infty$ hence $V\in L^p$ for any $p\geq 2$. The approximate potential we consider is calculated with 801 coefficients i.e. $|k|\leq 400$.

6.4.1 General framework

We first verify the convergence of the "inverse power" method (6.1.18), (6.1.19), (6.1.20) for different "strength" of the nonlinear contribution. Let us define $\omega \in \mathbb{R}^+$ the coefficient of nonlinearity. Equation (6.1.9) becomes

$$\left\{ \begin{array}{l} \forall v \in X, \int_\Omega \nabla u \cdot \nabla v + \int_\Omega V uv + \omega \int_\Omega u^3 v = \lambda \int_\Omega uv \\ \int_\Omega u^2 = 1. \end{array} \right. \tag{6.4.2}$$

Taking this coefficient into account, the algorithm we use to solve the equation in the space X_N is similar to (6.1.18), (6.1.19), (6.1.20).

This algorithm converges numerically for small values of ω. However for large values of ω the algorithm does not converge anymore. This non-convergence starts for ω in the range of 10. In order to cope with this problem, two strategies have been considered. First the convergence is improved for larger dimension of the space X_N. Hence we can increase the dimension in the numerical "exact" space and in the approximate spaces X_N. Another solution is to introduce a relaxation coefficient η, $0 < \eta \leq 1$, such that for each k a relaxation step is added in the algorithm as we define $\widetilde{u_N^k} = \eta u_N^k + (1-\eta) u_N^{k-1}$. This improves the convergence of the algorithm. For example for $N = 80$ and $\eta = 0.3$ the algorithm converges in less than 100 iterations for ω up to 15 and the number of iterations required to verify the condition $\|u_N^k - u_N^{k-1}\|_{H^1} < 10^{-12}$ increases from 27 to 80 when ω increases from 4 to 15.

In all what follows, ω is fixed equal to 1. The numerical "exact solution" is computed in the space $X_{300} = \text{Span}\{e_k, |k| \leq 300\}$, and the number of iterations is pushed so that the global residual defined above for this numerical "exact solution" is:

$$\|R_N^{k_{max}}\|_{H^{-1}} = \| - \Delta u_N^{k_{max}} + V u_N^{k_{max}} + (u_N^{k_{max}})^3 - \lambda_N^{k_{max}} u_N^{k_{max}} \|_{H^{-1}} = 1.7\ 10^{-12}$$

which is achieved with $k_{max} = 32$.

The total error is given by the *a posteriori* bound (6.3.53), that is

$$\begin{aligned} err_{total} &= \|R_N^k\|_{H^{-1}} + \|(V + 3(u_N^k)^2 - \lambda_N^k - 1)_-\|_{L^\infty} \left[\frac{1}{\beta_N^k} \|\Pi_N R_N^k\|_{H^{-1}} \right. \\ &\quad + \frac{2}{\beta_N^k} |\lambda_N^k - \mu_N^1| \|u_N^k - v_N^1\|_{L^2} \\ &\quad \left. + \frac{3}{2} \|u_N^k - \nu_N^1\|_{L^2}^2 \left(\frac{1}{N} + \frac{\|2(u_N^k)^2 \nu_N^1\|_{H^{-1}}}{\beta_N^k} \right) \right] \end{aligned} \tag{6.4.3}$$

Two error components are defined from the bound (6.3.54): the k-error and the N-error which depend respectively mainly on k and N. More precisely we define the k-error by

$$\begin{aligned} err_k &= \|R_{iter}\|_{H^{-1}} \left(1 + \frac{\|(V + 3(u_N^k)^2 - \lambda_N^k - 1)_-\|_{L^\infty}}{\beta_N^k} \right) \\ &\quad + \|(V + 3(u_N^k)^2 - \lambda_N^k - 1)_-\|_{L^\infty} \left[\frac{2}{\beta_N^k} |\lambda_N^k - \mu_N^1| \|u_N^k - v_N^1\|_{L^2} \right. \\ &\quad \left. + \frac{3}{2} \|u_N^k - \nu_N^1\|_{L^2}^2 \left(\frac{1}{N} + \frac{\|2(u_N^k)^2 \nu_N^1\|_{H^{-1}}}{\beta_N^k} \right) \right] \end{aligned} \tag{6.4.4}$$

and the N-error by

$$err_N = \|R_{disc}\|_{H^{-1}} \tag{6.4.5}$$

Let us notice that the total error is not exactly the sum of the two error components. Indeed

$$err_{total} \leq err_k + err_N,$$

hence err_{total} is a sharper estimate than the sum of the two contributions $err_k + err_N$.

6.4.2 With a large number of iterations

In this subsection, we compute different approximate solutions using a given large number of iterations and varying the dimension of the Fourier space X_N (see table 6.1). The number of iterations is $k_{max} = 32$ in our case for N between 15 and 100. Recall that this value of k_{max} corresponds to the minimum of iterations required to complete the condition the residual is less than 10^{-12} for $N = 300$. The eighth and ninth columns are given by $R_1, R_2 = \dfrac{err_{total}}{\|u_{exact} - u_N^k\|_{H^1}}$, R_1 being computed with the exact value of $\|(V + 3(u_N^k)^2 - \lambda_N^k - 1)_-\|_{L^\infty}$ (which in our case is zero) and R_2 being computed with $\|(V+3(u_N^k)^2-\lambda_N^k-1)_-\|_{L^\infty} \neq 0$, we force this contribution to be e.g. equal to 1, in order to see the effect of this additional contribution in the *a posteriori* bound. The second to last column indicates whether or not the conditions for the first a posteriori bound are verified. When this bound is guaranteed, the corresponding error bound is given in the last column.

Let us first remark that the total error (6.4.3) (with α chosen equal to 1) is larger than $\|u_{exact} - u_N^k\|_{H^1}$ which confirms the fact that the convergence of α to 1 is very fast, since then the total error is an upper bound for $\|u_{exact} - u_N^k\|_{H^1}$). Secondly the k-error obtained is close to 5.10^{-15} and almost constant which depicts the fact that the k-error is independent of N and almost zero when the algorithm has converged. The N-error is then the main component of the total error and decreases from 10^{-6} to 10^{-10}.

The total error is very close to $\|u_{exact} - u_N^k\|_{H^1}$ which shows that the *a posteriori* bounds found in the previous analysis seem to be close to optimal when the algorithm has converged.

In the last columns, we see that the error is guaranteed only for N larger than 45. Moreover the first guaranteed bound is larger than the real error by a factor between 1000 for relatively small values of N to 100 when N is large. Hence this first bound is actually much more coarse than the second a posteriori bound derived in the section 6.3.2.

6.4.3 In large dimension for the discretization space

In this subsection, we compute the approximate solution using a large dimension for the discretization space ($N = 100$) and varying the number of iterations. The number of iterations varies from 1 to 31. As in the previous subsection we remark that the total error is larger than $\|u_{exact} - u_N^k\|_{H^1}$. The N-error obtained is close to 10^{-10} and almost constant which depicts the fact that the N-error is independent of k and almost zero when N is large. The k-error is then the main component of the total error and decreases from 10^{-2} to 10^{-14}. The factor between the total error and $\|u_{exact} - u_N^k\|_{H^1}$ decreases from 1.3 to 1.0 when $\|(V+3(u_N^k)^2-\lambda_N^k-1)_-\|_{L^\infty}$ is computed exactly and is in always less than 5 even if we take $\|(V+3(u_N^k)^2-\lambda_N^k-1)_-\|_{L^\infty} = 1$. Hence the estimate seems to be close to optimal.

This illustrates the fact that the k-error estimator can be used as a stopping criterion for the convergence of the "inverse power" iterative technique (6.1.18), (6.1.19), (6.1.20).

The error is guaranteed for k larger than 13 and we remark as in the previous subsection that the first a posteriori bound given in the last column of the table is always worse than the

N	k	err_N	err_k	err_{total}	$\|u_{exact} - u_N^k\|_{H^1}$	R_1	R_2	C.-R. conditions guaranteed?	Guaranteed bound
15	32	1.1e-06	5.7e-15	1.1e-06	1.1e-06	1.006	1.006	No	
20	32	3.1e-07	5.8e-15	3.1e-07	3.1e-07	1.003	1.003	No	
25	32	1.2e-07	5.6e-15	1.2e-07	1.2e-07	1.002	1.002	No	
30	32	5.2e-08	5.8e-15	5.2e-08	5.2e-08	1.001	1.001	No	
35	32	2.6e-08	6.0e-15	2.6e-08	2.6e-08	1.001	1.001	No	
40	32	1.5e-08	5.5e-15	1.5e-08	1.5e-08	1.001	1.001	No	
45	32	8.6e-09	5.7e-15	8.6e-09	8.6e-09	1.000	1.000	Yes	1.1e-05
50	32	5.4e-09	6.1e-15	5.4e-09	5.4e-09	1.000	1.000	Yes	4.5e-06
55	32	3.5e-09	5.9e-15	3.5e-09	3.5e-09	1.000	1.000	Yes	2.2e-06
60	32	2.4e-09	5.8e-15	2.4e-09	2.4e-09	1.000	1.000	Yes	1.2e-06
65	32	1.7e-09	6.5e-15	1.7e-09	1.7e-09	1.000	1.000	Yes	7.3e-07
70	32	1.2e-09	6.1e-15	1.2e-09	1.2e-09	1.000	1.000	Yes	4.6e-07
75	32	8.8e-10	6.6e-15	8.8e-10	8.8e-10	1.000	1.001	Yes	3.1e-07
80	32	6.6e-10	5.8e-15	6.6e-10	6.6e-10	1.000	1.001	Yes	2.1e-07
85	32	5.1e-10	5.6e-15	5.1e-10	5.1e-10	1.000	1.001	Yes	1.5e-07
90	32	3.9e-10	5.5e-15	3.9e-10	3.9e-10	1.000	1.002	Yes	1.1e-07
95	32	3.1e-10	6.0e-15	3.1e-10	3.1e-10	1.000	1.002	Yes	8.1e-08
100	32	2.4e-10	5.8e-15	2.4e-10	2.4e-10	1.000	1.003	Yes	6.2e-08

Table 6.1 – Error components evolution with large number of iterations

precise a posteriori bound of section 6.3.2. The real error is about 1000 to 100 times less than the coarse a posteriori error bound.

Hence, the first a posteriori bound obtained with Caloz-Rappaz theory appears quite crude. It could probably be improved by analyzing more carefully the constant *gamma* that appears in the Caloz & Rappaz bound, leading to a better rough a posteriori bound. However, an improvement of a factor 10 (that may be obtain) on this bound would not be enough compared to the finer bound obtained in section 6.3.2. Therefore we did not push this estimation any further here.

6.4.4 Error balance

We would now like to minimize the computational cost necessary to reach a given acceptable total error. So given a small number ϵ, we want to find N and k as small as possible such that the total error is not larger than ϵ. As the total error is less than the sum of two error components depending respectively mainly on k or N, we try to balance the two error components.

Let ϵ be the total acceptable error. The aim is to find (N, k) such that $err_{total} < \epsilon$. The algorithm used is the following:

1. Set $N = 10$ and $k = 1$. Choose an initial pair (u_N^0, λ_N^0).
2. From $(u_N^{k-1}, \lambda_N^{k-1})$, find (u_N^k, λ_N^k) solution of (6.1.18) and (6.1.20).
3. Compute err_{total}

N	k	err_N	err_k	err_{total}	$\|u_{exact} - u_N^k\|_{H^1}$	R_1	R_2	C-R conditions guaranteed?	bound Guaranteed bound
100	1	2.4e-10	8.2e-02	8.2e-02	6.3e-02	1.30	3.62	No	
100	3	2.4e-10	1.1e-02	1.1e-02	8.9e-03	1.27	3.10	No	
100	5	2.4e-10	1.6e-03	1.6e-03	1.3e-03	1.26	3.00	No	
100	7	2.4e-10	2.2e-04	2.2e-04	1.8e-04	1.26	2.99	No	
100	9	2.4e-10	3.2e-05	3.2e-05	2.5e-05	1.26	2.98	No	
100	11	2.4e-10	4.5e-06	4.5e-06	3.6e-06	1.26	2.98	No	
100	13	2.4e-10	6.4e-07	6.4e-07	5.1e-07	1.26	2.98	Yes	1.6e-04
100	15	2.4e-10	9.1e-08	9.1e-08	7.2e-08	1.26	2.98	Yes	2.3e-05
100	17	2.4e-10	1.3e-08	1.3e-08	1.0e-08	1.26	2.98	Yes	3.3e-06
100	19	2.4e-10	1.9e-09	1.9e-09	1.5e-09	1.25	2.95	Yes	4.7e-07
100	21	2.4e-10	2.6e-10	3.6e-10	3.2e-10	1.11	2.23	Yes	9.0e-08
100	23	2.4e-10	3.7e-11	2.5e-10	2.5e-10	1.00	1.21	Yes	6.2e-08
100	25	2.4e-10	5.2e-12	2.4e-10	2.4e-10	1.00	1.03	Yes	6.2e-08
100	27	2.4e-10	7.4e-13	2.4e-10	2.4e-10	1.00	1.00	Yes	6.2e-08
100	29	2.4e-10	1.1e-13	2.4e-10	2.4e-10	1.00	1.00	Yes	6.2e-08
100	31	2.4e-10	1.5e-14	2.4e-10	2.4e-10	1.00	1.00	Yes	6.2e-08

Table 6.2 – Error components evolution in large dimensional space

- If $err_{total} < \epsilon$, then return (N, k)
- Else compute err_N and err_k
  - If $err_k > err_N$ then back to 2 with $k = k + 1$.
  - Else back to 1 with $N = N + 2$.

The table 6.3 shows the results of this study. Both the number of iterations and the dimension necessary to achieve a given accuracy increase when the accuracy increases.

It is interesting to note that the way k and N increase as the error decreases in the following manner $k = -2.39x - 0.46$ while $log(N) = -0.19x + 0.09$, where x is the logarithm of the acceptable error. The figure 6.1 illustrates this point. Note that N is 10 at the beginning of the error balance which explains the left part of the bottom figure.

Note that we obtain similar results for different potentials. For example, table 6.4 shows the results of the error balance with a less regular potential V with Fourier coefficients

$$\widehat{V}_k = -\frac{1}{\sqrt{2\pi}} \frac{1}{|k|^2 - \frac{1}{4}}. \tag{6.4.6}$$

The regularity of the potential affects the convergence of the algorithm, i.e. for $N = 100$, the N-error is only 3.10^{-5}, but the rate between the real error and the *a posteriori* boundis less than 4 for acceptable errors from $10^{-\frac{1}{2}}$ to 10^{-4}, which shows the sharpness of the *a posteriori* bounds.

6.5 Conclusions

In this paper we have performed a completely new *a posteriori* analysis for the solution technique applied to a simple but representative nonlinear eigenvalue problem. We first propose

N	k	err_k	err_N	err_{total}	$\|u_{exact} - u_N^k\|_{H^1}$	$\frac{err_{total}}{\|u_{exact}-u_N^k\|_{H^1}}$
10	1	8.2e-02	6.3e-06	8.2e-02	6.3e-02	1.30
10	3	1.6e-02	6.3e-06	1.6e-02	1.2e-02	1.28
10	4	6.8e-03	6.30e-06	6.8e-03	5.3e-03	1.27
10	5	2.9e-03	6.3e-06	2.9e-03	2.3e-03	1.26
10	7	5.4e-04	6.3e-06	5.4e-04	4.3e-04	1.26
10	8	2.3e-04	6.3e-06	2.3e-04	1.9e-04	1.26
10	10	4.4e-05	6.3e-06	4.4e-05	3.5e-05	1.25
10	11	1.9e-05	6.3e-06	2.0e-05	1.6e-05	1.22
12	12	3.1e-06	2.9e-06	4.2e-06	3.8e-06	1.12
14	13	1.3e-06	1.5e-06	2.0e-06	1.8e-06	1.10
18	14	2.2e-07	4.9e-07	5.4e-07	5.2e-07	1.03
22	15	9.3e-08	2.1e-07	2.3e-07	2.2e-07	1.04
28	16	1.5e-08	7.1e-08	7.2e-08	7.2e-08	1.01
34	17	6.6e-09	3.0e-08	3.1e-08	3.0e-08	1.01
44	18	4.0e-10	9.5e-09	9.6e-09	9.5e-09	1.00
58	20	1.1e-11	2.8e-09	2.8e-09	2.8e-09	1.00
74	21	1.7e-12	9.4e-10	9.4e-10	9.4e-10	1.00

Table 6.3 – Error balance with a potential having the Fourier coefficients $\widehat{V}_k = -\frac{1}{\sqrt{2\pi}}\frac{1}{|k|^4-\frac{1}{4}}$

N	k	err_k	err_N	err_{total}	$\|u_{exact} - u_N^k\|_{H^1}$	$\frac{err_{total}}{\|u_{exact}-u_N^k\|_{H^1}}$
10	4	2.6e-01	7.3e-03	2.6e-01	7.0e-02	3.63
10	6	8.4e-02	7.3e-03	8.5e-02	2.4e-02	3.52
10	8	2.8e-02	7.3e-03	2.9e-02	1.1e-02	2.77
16	8	5.5e-03	2.4e-03	6.0e-03	2.8e-03	2.14
22	10	1.8e-03	1.1e-03	2.1e-03	1.2e-03	1.79
28	12	6.2e-04	6.0e-04	8.6e-04	6.2e-04	1.38
38	13	1.2e-04	2.8e-04	3.1e-04	2.9e-04	1.08
58	14	4.6e-06	9.9e-05	9.9e-05	9.9e-05	1.00
94	16	2.0e-08	3.0e-05	3.0e-05	3.0e-05	1.00

Table 6.4 – Error balance with a potential having the Fourier coefficients $\widehat{V}_k = -\frac{1}{\sqrt{2\pi}}\frac{1}{|k|^2-\frac{1}{4}}$

a coarse estimator that allows to certify that we are indeed rather close to a solution of the nonlinear eigenvalue problem and also that this solution is the ground state, associated to the energy minimization problem (6.1.1). We propose also a refined estimator that is guaranteed only if we are close enough to the solution. We show that this refined estimator can be written as a sum of two contributions, each one dedicated to characterize the source of the main error, either due to the number of degrees of freedom used to discretize the problem or the number of iterations for the fixed point approach allowing to solve the nonlinear problem.

Note that in practice, we first adaptively refine the parameters N and k using the second a posteriori bound as an indicator. Indeed for rough parameters, the second bound is not guaranteed. However, when the solution is accurate enough, the first bound will certify that the error is below a certain quantity, which allows to use the second a posteriori bound as a guaranteed bound.

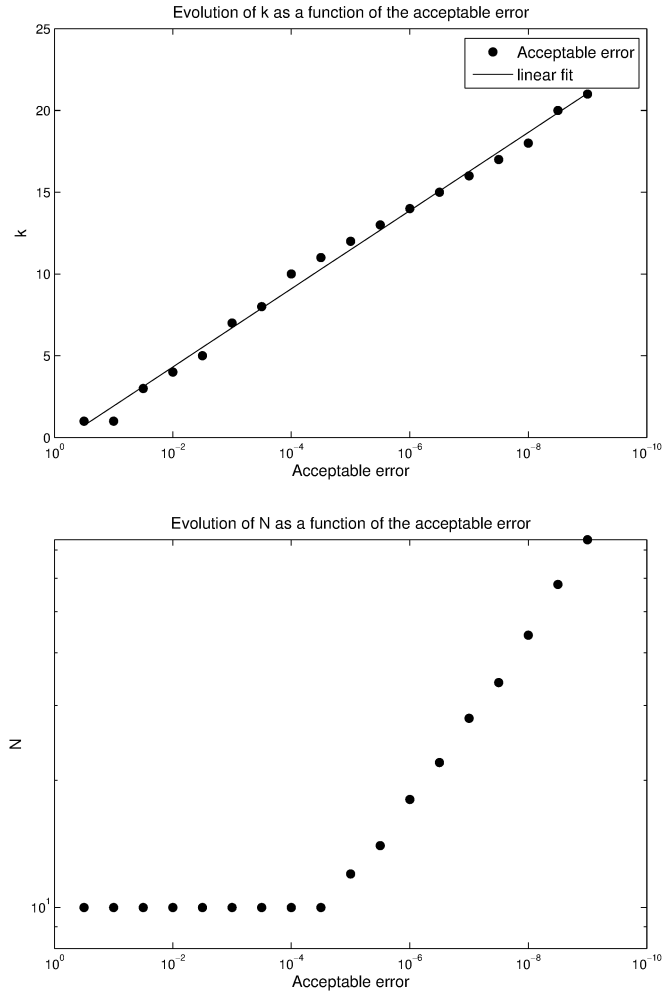

Figure 6.1 – Top: k evolution as the acceptable error decreases (linear fit in log-scale for the acceptable error). Bottom: N evolution as the acceptable error decreases (linear fit in loglog-scale)

The numerical simulations that have been provided show that the very precise analysis that can be done on this one dimensional example is optimal since the ratio between the error estimate and the exact error appears close to 1.

The extension of these ideas and techniques to a more complex framework of the Kohn-Sham problem is under consideration, following the techniques presented in [44].

Acknowledgements

The authors are grateful to Eric Cancès, Benjamin Stamm and Martin Vohralik so as the two unknown reviewers for helpful discussions and comments on a preliminary version of the present article. Financial support from the French state funds managed by CALSIMLAB and the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02 together with the ANR Manif are also acknowledged.

Part V

Post-processing for the planewave approximation of linear and nonlinear Schrödinger operators

Chapter 7

A perturbation-method-based post-processing for the planewave discretization of Kohn–Sham models

We expose in this chapter the results of [50]. This work was done in collaboration with with Eric Cancès, Yvon Maday, Benjamin Stamm and Martin Vohralík.

Abstract

In this article, we propose a post-processing of the planewave solution of the Kohn–Sham LDA model with pseudopotentials. This post-processing is based upon the fact that the exact solution can be interpreted as a perturbation of the approximate solution, allowing us to compute corrections for both the eigenfunctions and the eigenvalues of the problem in order to increase the accuracy. Indeed, this post-processing only requires the computation of the residual of the solution on a finer grid so that the additional computational cost is negligible compared to the initial cost of the planewave-based method needed to compute the approximate solution. Theoretical estimates certify an increased convergence rate in the asymptotic convergence range. Numerical results confirm the low computational cost of the post-processing and show that this procedure improves the energy accuracy of the solution even in the pre-asymptotic regime which comprises the target accuracy of practitioners.

7.1 Introduction

First-principle molecular simulation is nowadays a major tool in different fields such as chemistry, condensed matter physics and materials science. Its use is motivated by the fact that it enables one to understand and predict the properties of a molecular system, without any empirical parameter except a few fundamental constants of physics (the reduced Planck constant $\hbar$, the Boltzmann constant k_B, the mass of the electron m_e, the elementary charge e, the dielectric permittivity of the vacuum ε_0) and the masses and atomic numbers of the nuclei contained in the system under investigation.

At this level, matter is described as a system of interacting nuclei and electrons. Within the Born-Oppenheimer approximation [31] (made in almost all molecular simulations), nuclei are considered as classical point-like particles and electrons are assumed to be, at each time t, in their ground state. As a consequence, the nuclei dynamics is governed by a classical Hamiltonian

$$H\left(\{\mathbf{R}_k\}_{1\leq k\leq M}, \{\mathbf{P}_k\}_{1\leq k\leq M}\right) = \sum_{k=1}^{M} \frac{|\mathbf{P}_k|^2}{2m_k} + W(\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_M),$$

where m_k, $\mathbf{R}_k \in \mathbb{R}^3$ and $\mathbf{P}_k \in \mathbb{R}^3$ are respectively the mass, the position and the momentum of the k^{th} nucleus, where M is the total number of nuclei and where W is an effective potential taking into account the presence of the electrons. The bottleneck in first-principle molecular simulation is the evaluation of the potential W for a given nuclear configuration which requires computing the ground state energy of the electrons in the electrostatic potential generated by the nuclei. This quantity can, in principle, be computed by solving the electronic Schrödinger equation. However, as this equation is a (linear) $3\mathcal{N}$-dimensional partial differential equation, where $\mathcal{N}$ is the number of electrons in the system, this cannot be done by brute force numerical methods when $\mathcal{N}$ exceeds two or three due to the curse of dimensionality. Different approaches have been proposed to compute the electronic ground state energy. The most popular of them can be classified in three groups:

- wavefunction methods, among which the Hartree-Fock and multiconfiguration self-consistent-field (MCSCF) models (see [45] for a mathematical introduction);
- methods originating from the density functional theory (DFT), namely orbital-free and Kohn–Sham models, that are used and presented in details hereafter;
- quantum Monte Carlo methods [166, 165].

The Kohn–Sham models [93, 152] are the most popular approach to date as they offer a good compromise between accuracy and computational cost; they are among the most widely used models in physics and chemistry [39].

The purpose of this article is to present a new post-processing method for periodic Kohn–Sham calculations in planewave bases, leading to a significant gain in accuracy at a very limited extra computational cost. This method is based on the observation that the exact Kohn–Sham ground state can be considered as a perturbation of the approximate Kohn–Sham ground state computed in a finite basis set, and in applying first and second-order perturbation theory to the eigenvectors and eigenvalues of the Kohn–Sham operator respectively, in order to improve their accuracies. The specific structure of the problem and the *a priori* error estimates in [44] allow us to identify the leading terms in these first and second-order contributions, which turn

out to be easy to evaluate, and discard the other terms, which are very costly or impossible to evaluate, but negligible since proven to be small.

Our approach strongly relies on the fact that the kinetic energy operator which, from a mathematical point of view, is the leading term in the Kohn-Sham Hamiltonian commutes with the orthogonal projection on the discretization space. This is not the case for atomic orbital basis sets methods, but this is the case for other discretizations such as some wavelet methods. The extension of our approach to approximation settings that do not satisfy this commutation property requires additional theoretical investigations and is work in progress.

This article is organized as follows. We first recall in Section 7.2.1 the mathematical formulation of the Kohn–Sham models for isolated molecular systems. We then present the supercell Kohn–Sham model used in condensed phase modeling and simulation (Section 7.2.2), and the concept of pseudopotential (Section 7.2.3). The planewave discretization method for the supercell Kohn–Sham model with pseudopotential is discussed in Section 7.3.1 and the iterative algorithms used to solve the resulting Kohn–Sham equations are detailed in Section 7.3.2. The *a priori* error estimates our analysis is based upon are reviewed in Section 7.3.3. We then introduce the post-processing in Section 7.4. For pedagogical reasons, we derive the expressions of the post-processed eigenfunctions and eigenvalues under the assumptions that all the eigenvalues of the Kohn–Sham operator are non-degenerate. The general case, as well as the proof of Theorem 7.4.1, which quantifies the improvement of the Kohn–Sham ground state energy obtained by the post-processing in the asymptotic regime, will be detailed in a mathematical analysis oriented paper [97]. In Section 7.5, we report numerical simulations on a simple system, an alanine molecule, obtained with the KSSOLV package [252], showing that our post-processing method leads to significant gain in accuracy (typically one order of magnitude on the energy) for a small extra cost (a few percent of the overall cost). Numerical simulations with the CO_2 and the benzene molecules were also performed and led to very similar results, and therefore are not presented in this paper.

7.2 DFT Kohn–Sham models

7.2.1 Introduction to Kohn–Sham models

Throughout this article, we adopt the system of atomic units for which $\hbar = 1$, $m_e = 1$, $e = 1$, $4\pi\epsilon_0 = 1$. In this system of units, the charge of the electron is -1 and the charges of the nuclei are positive integers.

Let us first consider an isolated molecular system *in vacuo*, consisting of M nuclei of charges $(z_1, \cdots, z_M) \in (\mathbb{N} \setminus \{0\})^M$ located at the positions $(\mathbf{R}_1, \cdots, \mathbf{R}_M) \in (\mathbb{R}^3)^M$ of the physical space, and of $\mathcal{N}$ electrons. The electrostatic potential generated by the nuclei and felt by the electrons is then given by

$$V_{\rm nuc}(\mathbf{r}) = -\sum_{k=1}^{M} \frac{z_k}{|\mathbf{r} - \mathbf{R}_k|}. \tag{7.2.1}$$

In the spin-restricted Kohn–Sham model, the electronic state of a closed-shell system with an even number $\mathcal{N} = 2N$ of electrons is described by N Kohn–Sham orbitals $\Phi = (\phi_1, \cdots, \phi_N)^T \in \left[H^1(\mathbb{R}^3)\right]^N$ satisfying the orthonormality conditions

$$\int_{\mathbb{R}^3} \phi_i \phi_j = \delta_{ij}, \qquad \forall\, i, j = 1, \ldots, N.$$

The associated electronic density

$$\rho_{[\Phi]}(\mathbf{r}) := 2 \sum_{i=1}^{N} |\phi_i(\mathbf{r})|^2$$

plays a key-role in DFT. The factor 2 in the above expression accounts for the spin. Indeed, in the spin-restricted Kohn–Sham model, each orbital is occupied by two electrons, one with spin up and one with spin down. The Kohn–Sham ground state is then obtained by solving

$$\inf \left\{ \mathcal{E}_0^{\mathrm{KS}}(\Phi),\ \Phi = (\phi_1, \cdots, \phi_N)^T \in \left[H^1(\mathbb{R}^3)\right]^N,\ \int_{\mathbb{R}^3} \phi_i \phi_j = \delta_{ij} \right\}, \tag{7.2.2}$$

where the Kohn–Sham energy functional reads

$$\mathcal{E}_0^{\mathrm{KS}}(\Phi) := \sum_{i=1}^{N} \int_{\mathbb{R}^3} |\nabla \phi_i|^2 + \int_{\mathbb{R}^3} V_{\mathrm{nuc}}\, \rho_{[\Phi]} + \frac{1}{2} D(\rho_{[\Phi]}, \rho_{[\Phi]}) + E_{\mathrm{xc}}(\rho_{[\Phi]}). \tag{7.2.3}$$

In the right-hand side of (7.2.3), the first term approximates the kinetic energy of the electrons, the second term accounts for the interactions between nuclei and electrons, and the third term is a (crude) approximation of the interaction between electrons. The bilinear form $D(\cdot,\cdot)$ is in fact the Coulomb energy functional *in vacuo*:

$$D(\rho, \rho') = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(\mathbf{r})\, \rho'(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\, d\mathbf{r}\, d\mathbf{r}', \tag{7.2.4}$$

so that $\frac{1}{2} D(\rho_{[\Phi]}, \rho_{[\Phi]})$ is the Coulomb energy of a *classical* charge distribution of density $\rho_{[\Phi]}$. The fourth term in the right-hand side of (7.2.3), called the exchange-correlation functional, is a correction term, which is essential to describe quantitatively, and sometimes even qualitatively, the physics and chemistry of the system. The exchange-correlation functional collects the errors made in the approximations of the kinetic energy and of the interactions between electrons by respectively the first and third terms of the Kohn–Sham functional. It follows from the Hohenberg–Kohn theorem [139, 167, 169, 237] that there exists an *exact* exchange-correlation functional, that is a functional of the electronic density $\rho_{[\Phi]}$ for which solving (7.2.2) provides the ground state electronic energy and density of the $\mathcal{N}$-body electronic Schrödinger equation. Unfortunately, no mathematical expression of the exchange-correlation functional amenable to numerical simulations is known. It therefore has to be approximated in practice. The local density approximation (LDA) consists in approximating the exchange-correlation functional by

$$E_{\mathrm{xc}}(\rho_{[\Phi]}) = \int_{\mathbb{R}^3} e_{\mathrm{xc}}^{\mathrm{LDA}}(\rho_{[\Phi]}(\mathbf{r}))\, d\mathbf{r}, \tag{7.2.5}$$

where $e_{\mathrm{xc}}^{\mathrm{LDA}}(\overline{\rho})$ is an approximation of the exchange-correlation energy per unit volume in a homogeneous electron gas with density $\overline{\rho}$.

7.2.2 Periodic Kohn–Sham models

In the sequel, we will focus on the *periodic versions* of the Kohn–Sham LDA model. In the periodic setting, the nuclear configuration is supposed to be $\mathcal{R}$-periodic, where $\mathcal{R}$ is a discrete periodic lattice of $\mathbb{R}^3$, and the simulation domain, sometimes referred to as the supercell, no longer consists of the whole space $\mathbb{R}^3$ (as in (7.2.2)–(7.2.3)) but is a unit cell, denoted here by Ω, of the periodic lattice $\mathcal{R}$. Periodic boundary conditions (PBC) are imposed to the Kohn–Sham

orbitals (Born–von Karman PBC) at the boundary $\partial\Omega$ of the simulation cell Ω. This is the standard method to compute condensed phase properties with a limited number of atoms in the simulation cell, hence at a moderate computational cost. In most applications in solid state physics and materials science, the periodic Kohn–Sham models are discretized in Fourier modes, more commonly referred to as planewave basis sets in the physics and chemistry literature. This is the reason why we focus on this particular discretization method in the present work.

As a consequence, the domain of integration in (7.2.3) and (7.2.5) is now Ω instead of $\mathbb{R}^3$, and the Coulomb energy is now defined as

$$D_\Omega(\rho,\rho') = \int_\Omega \int_\Omega G_\Omega(\mathbf{r},\mathbf{r}')\rho(\mathbf{r})\rho'(\mathbf{r}')d\mathbf{r}d\mathbf{r}' = \int_\Omega \rho(\mathbf{r}')[V_{\mathrm{coul}}(\rho')](\mathbf{r}')d\mathbf{r}',$$

where the Green's function G_Ω and the periodic Coulomb potential $V_{\mathrm{coul}}(\rho')$ are respectively solutions to the following problems

$$\begin{cases} -\Delta G_\Omega = 4\pi\left(\displaystyle\sum_{\mathbf{k}\in\mathcal{R}}\delta_{\mathbf{k}} - \frac{1}{|\Omega|}\right) & \text{in} \quad \mathbb{R}^3, \\ G_\Omega \quad \mathcal{R}-\text{periodic}, \\ \displaystyle\int_\Omega G_\Omega = 0, \end{cases} \quad \text{and} \quad \begin{cases} -\Delta V_{\mathrm{coul}}(\rho') = 4\pi\left(\rho' - \dfrac{1}{|\Omega|}\displaystyle\int_\Omega \rho'\right) \text{in}\,\mathbb{R}^3, \\ V_{\mathrm{coul}}(\rho') \quad \mathcal{R}-\text{periodic}, \\ \displaystyle\int_\Omega V_{\mathrm{coul}}(\rho') = 0. \end{cases}$$

The exchange-correlation functional in this periodic setting is given by

$$E_{\mathrm{xc},\Omega}(\rho) = \int_\Omega e_{\mathrm{xc}}^{\mathrm{LDA}}(\rho(\mathbf{r}))\,d\mathbf{r},$$

and the orthonormality constraints read

$$\int_\Omega \phi_i\,\phi_j = \delta_{ij}.$$

7.2.3 Pseudopotentials

The core electrons of an atom are barely affected by the chemical environment. In pseudopotential methods, the all-electron model is replaced by a reduced model explicitly dealing with valence electrons only, while core electrons are frozen in some reference state. The valence electrons are described by valence pseudo-orbitals, and the interaction between the valence electrons and the ionic cores (an ionic core consists of a nucleus and of the associated core electrons) is modelled by a nonlocal operator called a pseudopotential, constructed once and for all from single-atom reference calculations. The reduction of dimensionality obtained by getting rid of the core electrons results in a much less computationally expensive approach since only the valence pseudo-orbitals need to be computed. In addition, pseudopotentials are constructed in such a way that the valence pseudo-orbitals oscillate much less than the valence orbitals in the core region, hence can be approximated using smaller planewave bases, or discretized on coarser grids. Lastly, pseudopotentials are used to take into account, in the nonrelativistic framework of the Kohn–Sham model, some relativistic effects which play an important role in the simulation of heavy atoms.

The resulting model for the pseudo-orbitals is similar to (7.2.2)–(7.2.3), but presents some differences:

(i) N now denotes the number of *valence* electron pairs;

(ii) Φ now denotes the set of the pseudo-orbitals of the valence electrons;

(iii) the nuclear potential V_{nuc} is replaced by a *pseudopotential* operator V_{ion} modelling the interaction between the valence electrons on the one hand, and the nuclei and the core electrons on the other hand.

More precisely, the pseudopotential consists of two terms: a local component V_{local} (whose associated operator is the multiplication by the $\mathcal{R}$-periodic function V_{local}) and a nonlocal component V_{nl} given by

$$V_{\text{nl}}\phi = \sum_{j=1}^{J} \left(\int_{\Omega} \xi_j(\mathbf{r})\phi(\mathbf{r})\, d\mathbf{r} \right) \xi_j, \tag{7.2.6}$$

where ξ_j are regular enough $\mathcal{R}$-periodic functions and J is an integer depending on the chemical nature of the ions in the unit cell. As a consequence, the second term in the Kohn–Sham energy functional (7.2.3) is replaced by

$$\int_{\Omega} \rho_{[\Phi]} V_{\text{local}} + 2\sum_{i=1}^{N} \langle \phi_i | V_{\text{nl}} | \phi_i \rangle, \tag{7.2.7}$$

where the Dirac bra-ket notation is used to represent the non-local part of the operator V_{ion}.

Further, a correction to the exchange-correlation energy due to the introduction of pseudpotential is done by setting

$$E^c_{\text{xc},\Omega}(\rho_{[\Phi]}) = \int_{\Omega} e^{\text{LDA}}_{\text{xc}}(\rho_c(\mathbf{r}) + \rho_{[\Phi]}(\mathbf{r}))\, d\mathbf{r},$$

where $\rho_c \geq 0$ is a nonlinear core-correction.

To summarize, we are therefore considering the following energy functional

$$\mathcal{E}^{\text{KS}}_{0,\Omega}(\Phi) = \sum_{i=1}^{N} \int_{\Omega} |\nabla \phi_i|^2 + \int_{\Omega} V_{\text{local}}\, \rho_{[\Phi]} + 2\sum_{i=1}^{N} \langle \phi_i | V_{\text{nl}} | \phi_i \rangle + \frac{1}{2} D_{\Omega}(\rho_{[\Phi]}, \rho_{[\Phi]}) + E^c_{\text{xc},\Omega}(\rho_{[\Phi]}), \tag{7.2.8}$$

and the set of admissible states

$$\mathcal{M} = \left\{ \Phi = (\phi_1, \cdots, \phi_N)^T \in \left[H^1_{\#}(\Omega) \right]^N \;\middle|\; \int_{\Omega} \phi_i \phi_j = \delta_{ij} \right\},$$

where

$$H^1_{\#}(\Omega) = \left\{ \phi \in H^1_{\text{loc}}(\mathbb{R}^3) \;\middle|\; \phi \quad \mathcal{R} - \text{periodic} \right\}$$

is the $\mathcal{R}$-periodic H^1-space.

The ground state energy is then defined by

$$I_0^{KS} = \inf \left\{ \mathcal{E}^{\text{KS}}_{0,\Omega}(\Phi),\ \Phi \in \mathcal{M} \right\}. \tag{7.2.9}$$

It can be proven that, under reasonable assumptions on $V_{\text{nl}}, V_{\text{local}}$, and $E^c_{\text{xc},\Omega}$ (see [44]), (7.2.9) has a minimizer $\Phi^0 = (\phi^0_1, \ldots, \phi^0_N) \in \mathcal{M}$ (see Section 7.3.3).

The first optimality conditions read

$$\forall 1 \leq i \leq N, \quad \mathcal{H}^0\, \phi^0_i = \sum_{j=1}^{N} \lambda^0_{ij} \phi^0_j, \quad \int_{\Omega} \phi^0_i \phi^0_j = \delta_{ij}, \tag{7.2.10}$$

where the $N \times N$ matrix $\Lambda^0 = (\lambda_{ij}^0)$, which is the Lagrange multiplier of the matrix constraint $\int_\Omega \phi_i \phi_j = \delta_{ij}$, is symmetric, and where the Hamiltonian $\mathcal{H}^0$ is defined as follows:

$$\mathcal{H}^0 = -\frac{1}{2}\Delta + V_{\text{ion}} + V_{\text{coul}}(\rho^0) + V_{\text{xc}}(\rho^0),$$

with $\rho^0 = \rho_{[\Phi^0]}$, $V_{\text{ion}} = V_{\text{local}} + V_{\text{nl}}$, and where

$$V_{\text{xc}}(\rho)(\mathbf{r}) = \frac{de_{\text{xc}}^{\text{LDA}}}{d\rho}(\rho_c(\mathbf{r}) + \rho(\mathbf{r})).$$

Note that $\mathcal{H}^0$ is the Kohn–Sham operator

$$\mathcal{H}_{[\rho]} = -\frac{1}{2}\Delta + V_{\text{ion}} + V_{\text{coul}}(\rho) + V_{\text{xc}}(\rho), \tag{7.2.11}$$

in the case where ρ is the ground state density ρ^0.

In fact, (7.2.9) has an infinity of minimizers since any unitary transform of the Kohn–Sham orbitals Φ^0 in the sense of (7.2.12) below is also a minimizer of the Kohn–Sham energy. This is a consequence of the following invariance property:

$$\forall \Phi \in \mathcal{M}, \quad \forall U \in \mathcal{U}(N), \quad U\Phi \in \mathcal{M}, \quad \rho_{[U\Phi]} = \rho_{[\Phi]} \quad \text{and} \quad \mathcal{E}_{0,\Omega}^{\text{KS}}(U\Phi) = \mathcal{E}_{0,\Omega}^{\text{KS}}(\Phi), \tag{7.2.12}$$

where $\mathcal{U}(N)$ is the group of orthogonal matrices:

$$\mathcal{U}(N) = \left\{ U \in \mathbb{R}^{N\times N} \mid U^T U = 1_N \right\},$$

1_N denoting the identity matrix of rank N. This invariance can be exploited to diagonalize the matrix of the Lagrange multipliers of the orthonormality constraints (see e.g. [93]), yielding the existence of a minimizer (still denoted by Φ^0) with the same density ρ^0, such that

$$\mathcal{H}^0 \phi_i^0 = \epsilon_i^0 \phi_i^0, \quad \int_\Omega \phi_i^0 \phi_j^0 = \delta_{ij}, \tag{7.2.13}$$

for some $\epsilon_1^0 \leq \epsilon_2^0 \leq \cdots \leq \epsilon_N^0$.

As discussed in [44], it is not known whether Kohn–Sham ground states satisfy the so-called *Aufbau* principle, that is whether $\epsilon_1^0, \cdots, \epsilon_N^0$ are the lowest N eigenvalues of the Kohn–Sham Hamiltonian $\mathcal{H}^0$. However, this property seems to be satisfied in practice for most systems, and it is always satisfied for the extended Kohn–Sham model (see [44] for details). We will assume here that the molecular system under consideration does satisfy the *Aufbau* principle. This allows us to solve the Kohn–Sham equations using iterative algorithms such as the one described in Section 7.3.2, which implicitly rely on the *Aufbau* principle.

7.3 Discretization and resolution of the Kohn–Sham model

7.3.1 Planewave discretization

In order to approximate the solution of (7.2.9), we first discretize the variational set $\mathcal{M}$ using a planewave basis set. This approximation setting is the state-of-the-art method for Kohn–Sham simulations in solid state physics and materials science. Thus the computational domain Ω equipped with periodic boundary conditions can be a cubic box, or more generally the unit

cell of a periodic lattice $\mathcal{R} \subset \mathbb{R}^3$. The valence pseudo-orbitals are expanded in terms of the functions $e_{\mathbf{k}}(\mathbf{r}) := |\Omega|^{-1/2} e^{i\mathbf{k}\cdot\mathbf{r}}$, which are the Fourier modes with wavevectors $\mathbf{k} \in \mathcal{R}^*$, where $\mathcal{R}^*$ denotes the dual lattice of $\mathcal{R}$. The lattice $\mathcal{R}^*$ indeed consists of all wavevectors $\mathbf{k}$ such that $e_{\mathbf{k}}$ is $\mathcal{R}$-periodic. In this article, we assume for simplicity that $\Omega = [0, L)^3$ ($L > 0$) but our arguments can be easily extended to the general case. In this case, $\mathcal{R} = L\mathbb{Z}^3$ and $\mathcal{R}^* = \frac{2\pi}{L}\mathbb{Z}^3$.

Recall that the family $(e_{\mathbf{k}})_{\mathbf{k}\in\mathcal{R}^*}$ forms an orthonormal basis of

$$L^2_{\#}(\Omega, \mathbb{C}) := \left\{ u \in L^2_{\mathrm{loc}}(\mathbb{R}^3, \mathbb{C}) \mid u \text{ is } \mathcal{R}\text{-periodic} \right\},$$

and that for all $v \in L^2_{\#}(\Omega, \mathbb{C})$,

$$v(\mathbf{r}) = \sum_{\mathbf{k}\in\mathcal{R}^*} \widehat{v}_{\mathbf{k}}\, e_{\mathbf{k}}(\mathbf{r}) \qquad \text{with} \qquad \widehat{v}_{\mathbf{k}} = (e_{\mathbf{k}}, v)_{L^2_{\#}} = |\Omega|^{-1/2} \int_\Omega v(\mathbf{r}) e^{-i\mathbf{k}\cdot\mathbf{r}}\, d\mathbf{r}.$$

For each $s \in \mathbb{R}$, we denote by

$$H^s_{\#}(\Omega) := \left\{ v = \sum_{\mathbf{k}\in\mathcal{R}^*} \widehat{v}_{\mathbf{k}} e_{\mathbf{k}} \;\middle|\; \forall \mathbf{k}, \quad \widehat{v}_{-\mathbf{k}} = \widehat{v}^*_{\mathbf{k}}, \quad \|v\|^2_{H^s_{\#}} := \sum_{\mathbf{k}\in\mathcal{R}^*} (1 + |\mathbf{k}|^2)^s |\widehat{v}_{\mathbf{k}}|^2 < \infty \right\}$$

the Sobolev space of real-valued periodic distributions with regularity H^s.

The kinetic energy of a basis function $e_{\mathbf{k}}$ is given by $\frac{1}{2}|\mathbf{k}|^2$, where $|\cdot|$ denotes the Euclidean norm. We introduce some energy cutoff $E_{\mathrm{c}} > 0$ and consider all basis functions whose kinetic energy is smaller than E_{c}, i.e. $|\mathbf{k}| \le \sqrt{2E_{\mathrm{c}}}$, to define the approximation space. That is, for each cutoff E_{c}, we set $N_{\mathrm{c}} = \sqrt{\frac{E_{\mathrm{c}}}{2}} \frac{L}{\pi}$ and consider the finite-dimensional discretization space

$$X_{N_{\mathrm{c}}} := \left\{ \sum_{\mathbf{k}\in\mathcal{R}^*, |\mathbf{k}| \le \frac{2\pi}{L} N_{\mathrm{c}}} \widehat{v}_{\mathbf{k}}\, e_{\mathbf{k}} \;\middle|\; \forall \mathbf{k}, \quad \widehat{v}_{-\mathbf{k}} = \widehat{v}^*_{\mathbf{k}} \right\} \subset \bigcap_{s\in\mathbb{R}} H^s_{\#}(\Omega).$$

We also denote by $\Pi_{N_{\mathrm{c}}}$, the linear operator on the space of $\mathcal{R}$-periodic distributions defined by

$$\Pi_{N_{\mathrm{c}}} \left(\sum_{\mathbf{k}\in\mathcal{R}^*} \widehat{v}_{\mathbf{k}} e_{\mathbf{k}} \right) = \sum_{\mathbf{k}\in\mathcal{R}^*, |\mathbf{k}| \le \frac{2\pi}{L} N_{\mathrm{c}}} \widehat{v}_{\mathbf{k}} e_{\mathbf{k}}.$$

The operator $\Pi_{N_{\mathrm{c}}}|_{H^s_{\#}(\Omega)}$ (which we shall also denote by $\Pi_{N_{\mathrm{c}}}$ for convenience) is in fact the orthogonal projector from $H^s_{\#}(\Omega)$ to $X_{N_{\mathrm{c}}}$ for any $s \in \mathbb{R}$, and we denote by $\Pi^{\perp}_{N_{\mathrm{c}}} = (1 - \Pi_{N_{\mathrm{c}}})$ the orthogonal projector on $X^{\perp}_{N_{\mathrm{c}}}$, the orthogonal of $X_{N_{\mathrm{c}}}$ in $H^s_{\#}(\Omega)$ (again for any $s \in \mathbb{R}$). Finally, the variational approximation to the ground state energy in $X_{N_{\mathrm{c}}}$ is defined as

$$I^{\mathrm{KS}}_{0,N_{\mathrm{c}}} = \inf \left\{ \mathcal{E}^{\mathrm{KS}}_0(\Phi_{N_{\mathrm{c}}}),\ \Phi_{N_{\mathrm{c}}} \in \mathcal{M} \cap [X_{N_{\mathrm{c}}}]^N \right\}. \tag{7.3.1}$$

Using once again the invariance property (7.2.12), the Euler equations of this minimization problem can be diagonalized and therefore reduced to find the pairs $(\phi_{j,N_{\mathrm{c}}}, \epsilon_{j,N_{\mathrm{c}}})_{j=1,\dots,N}$ satisfying

$$\mathcal{H}_{N_{\mathrm{c}},\mathrm{proj}}\, \phi_{j,N_{\mathrm{c}}} = \epsilon_{j,N_{\mathrm{c}}}\, \phi_{j,N_{\mathrm{c}}}, \qquad \int_\Omega \phi_{i,N_{\mathrm{c}}}\, \phi_{j,N_{\mathrm{c}}} = \delta_{ij}, \qquad \epsilon_{1,N_{\mathrm{c}}} \le \epsilon_{2,N_{\mathrm{c}}} \le \ldots \le \epsilon_{N,N_{\mathrm{c}}}, \tag{7.3.2}$$

for all $i, j = 1, \ldots, N$. Here we define $\mathcal{H}_{N_\mathrm{c},\mathrm{proj}} : X_{N_\mathrm{c}} \to X_{N_\mathrm{c}}$ by

$$\mathcal{H}_{N_\mathrm{c},\mathrm{proj}} = \Pi_{N_\mathrm{c}} \mathcal{H}_{[\rho_{N_\mathrm{c}}]} \Pi_{N_\mathrm{c}} = -\frac{1}{2} \Pi_{N_\mathrm{c}} \Delta \Pi_{N_\mathrm{c}} + \Pi_{N_\mathrm{c}} \Big[V_\mathrm{ion} + V_\mathrm{coul}(\rho_{N_\mathrm{c}}) + V_\mathrm{xc}(\rho_{N_\mathrm{c}}) \Big] \Pi_{N_\mathrm{c}},$$

with $\rho_{N_\mathrm{c}} = \rho_{[\Phi_{N_\mathrm{c}}]}$, $\Phi_{N_\mathrm{c}} = (\phi_{1,N_\mathrm{c}}, \ldots, \phi_{N,N_\mathrm{c}})^T$ and where $\mathcal{H}_{[\rho_{N_\mathrm{c}}]}$ is defined by (7.2.11) for the approximate ground state density ρ_{N_c}. A key observation is that $\mathcal{H}_{N_\mathrm{c},\mathrm{proj}}$ is the restriction to X_{N_c} of the self-adjoint operator $\mathcal{H}_{N_\mathrm{c}}$ on $L^2_\#(\Omega)$ with domain $H^2_\#(\Omega)$ defined by

$$\mathcal{H}_{N_\mathrm{c}} = -\frac{1}{2} \Delta + \Pi_{N_\mathrm{c}} \Big[V_\mathrm{ion} + V_\mathrm{coul}(\rho_{N_\mathrm{c}}) + V_\mathrm{xc}(\rho_{N_\mathrm{c}}) \Big] \Pi_{N_\mathrm{c}},$$

and that $X_{N_\mathrm{c}} \oplus X_{N_\mathrm{c}}^\perp$ is a $\mathcal{H}_{N_\mathrm{c}}$-stable decomposition of $L^2_\#(\Omega)$. More precisely, the operator $\mathcal{H}_{N_\mathrm{c}}$ can be decomposed as follows:

$$\mathcal{H}_{N_\mathrm{c}} = \begin{pmatrix} \mathcal{H}_{N_\mathrm{c},\mathrm{proj}} & 0 \\ 0 & -\frac{1}{2}\Delta \end{pmatrix} \begin{matrix} \Big\} X_{N_\mathrm{c}} \\ \Big\} X_{N_\mathrm{c}}^\perp \end{matrix}$$

$$\underbrace{\qquad\qquad}_{X_{N_\mathrm{c}}} \quad \underbrace{\qquad\qquad}_{X_{N_\mathrm{c}}^\perp}$$

Note that, as we are using a planewave discretization, the operator $-\frac{1}{2}\Delta$ is diagonal on $X_{N_\mathrm{c}}^\perp$ with smallest eigenvalue larger than $\frac{1}{2}\left(\frac{LN_\mathrm{c}}{\pi}\right)^2$. Thus, as soon as

$$\epsilon_{N,N_\mathrm{c}} < \frac{1}{2}\left(\frac{LN_\mathrm{c}}{\pi}\right)^2, \tag{7.3.3}$$

$\epsilon_{N,N_\mathrm{c}}$ being the N^{th} eigenvalue of the operator $\mathcal{H}_{N_\mathrm{c},\mathrm{proj}}$, Φ_{N_c} is also solution to the following eigenvalue problem

$$\mathcal{H}_{N_\mathrm{c}} \phi_{j,N_\mathrm{c}} = \epsilon_{j,N_\mathrm{c}} \phi_{j,N_\mathrm{c}}, \qquad \int_\Omega \phi_{i,N_\mathrm{c}}\, \phi_{j,N_\mathrm{c}} = \delta_{ij}, \qquad \epsilon_{1,N_\mathrm{c}} \le \epsilon_{2,N_\mathrm{c}} \le \ldots \le \epsilon_{N,N_\mathrm{c}}, \tag{7.3.4}$$

$\epsilon_{k,N_\mathrm{c}}$ being the k-th eigenvalue of $\mathcal{H}_{N_\mathrm{c}}$.

Further, the energy can then alternatively, but equivalently to (7.2.8), be obtained by

$$\mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\Phi_{N_\mathrm{c}}) = 2\sum_{i=1}^{N} \epsilon_{j,N_\mathrm{c}} - \frac{1}{2} D_\Omega(\rho_{N_\mathrm{c}}, \rho_{N_\mathrm{c}}) + E^c_\mathrm{xc}(\rho_{N_\mathrm{c}}) - \int_\Omega V_\mathrm{xc}(\rho_{N_\mathrm{c}})\, \rho_{N_\mathrm{c}}, \tag{7.3.5}$$

where the right-hand side only depends on the eigenvalues and the electron density [129]. This energy is called double-counting energy in the following.

7.3.2 SCF-iterations

In order to solve the nonlinear eigenvalue problem (7.3.2), a Self-Consistent Field (SCF) procedure is employed [252]. It consists of solving a linear eigenvalue problem at each step, at which the Hamiltonian is computed from the density found at the previous step. Moreover, a linear charge mixing is performed in order to improve the convergence of the algorithm. The algorithm can be written as follows:

1. **Initialization:** Take an initial guess of the orbitals $\Phi_{N_c}^{(0)} = (\phi_{1,N_c}^{(0)}, \ldots, \phi_{N,N_c}^{(0)})^T$ with associated density $\rho_{N_c}^{(0)} = \rho_{[\Phi_{N_c}^{(0)}]}$, a memory parameter $m \in \mathbb{N}$ and a tolerance $\eta > 0$.

2. **Iterations:** For $i = 1, 2, ...$ until convergence

   (a) Compute the Hamiltonian $\mathcal{H}_{N_c,\mathrm{proj}}^{(i-1)} := \Pi_{N_c} \mathcal{H}_{[\rho_{N_c}^{(i-1)}]} \Pi_{N_c}$.

   (b) Solve the linear eigenvalue problem

   $$\mathcal{H}_{N_c,\mathrm{proj}}^{(i-1)} \phi_{j,N_c}^{(i)} = \epsilon_{j,N_c}^{(i)} \phi_{j,N_c}^{(i)}, \quad \int_\Omega \phi_{j,N_c}^{(i)} \phi_{k,N_c}^{(i)} = \delta_{ik},$$

   for $j = 1, \ldots, N$ to obtain a set of orbitals $\Phi_{N_c}^{(i)} = (\phi_{1,N_c}^{(i)}, \ldots, \phi_{N,N_c}^{(i)})^T$, by selecting the lowest N eigenvalues $\epsilon_{1,N_c}^{(i)}, \epsilon_{2,N_c}^{(i)}, \ldots, \epsilon_{N,N_c}^{(i)}$, counted with their multiplicities, and corresponding eigenfunctions $(\phi_{1,N_c}^{(i)}, \phi_{2,N_c}^{(i)}, \ldots, \phi_{N,N_c}^{(i)})^T \in \mathcal{M}$, following the *Aufbau principle*.

   (c) Compute the new density $\check{\rho}_{N_c}^{(i)} = \rho_{[\Phi_{N_c}^{(i)}]}$.

   (d) Charge mixing: replace the charge density $\check{\rho}_{N_c}^{(i)}$ with a linear combination of previously computed charge densities, i.e.,

   $$\rho_{N_c}^{(i)} = \alpha_0 \check{\rho}_{N_c}^{(i)} + \sum_{k=1}^{\min(i,m)} \alpha_k \rho_{N_c}^{(i-k)}$$

   with appropriately chosen mixing parameters satisfying $\sum_{k=0}^{\min(i,m)} \alpha_k = 1$.

3. **Output:** If $\|\rho_{N_c}^{(i)} - \rho_{N_c}^{(i-1)}\| < \eta$, where $\|.\|$ is a given norm, stop, else go back to step 2.

Note that several points in this algorithm need to be specified for its practical implementation. Indeed, several linear eigenvalue solvers and several charge mixing procedures are available, and different norms can be used for the convergence test. Every possible choice could be used, without affecting the results presented in the next section. The choices made for the numerical tests reported in Section 7.5 will be described there.

7.3.3 Smoothness assumptions and *a priori* results

In order to guarantee the existence of minimizers of problem (7.2.9) and to study the convergence of the solutions to the discretized problem (7.3.1) to those of the continuous problem (7.2.9), some assumptions on the data are needed. However, to avoid technicalities, these assumptions are not detailed in the present paper and the interested reader is referred to [44].

First, under sufficient regularity assumptions, problem (7.2.9) with energy functional (7.2.8) has a minimizer Φ^0 satisfying (7.2.13). Note that the Xα exchange-correlation functional defined by $e_{\mathrm{xc}}^{\mathrm{X}\alpha}(\rho) = -C_{\mathrm{X}} \rho^{4/3}$, where $C_{\mathrm{X}} > 0$ is a given constant, satisfies these assumptions. They are also satisfied by the exact exchange-correlation functional.

Second, the following *a priori* estimates hold.

Theorem 7.3.1 ([44]). *Under sufficient regularity assumptions, there exists $r^0 > 0$ and N_c^0 such that for $N_c \geq N_c^0$,* (7.3.1) *has a unique local minimizer $\Phi_{N_c}^0$ in the set*

$$\left\{ \Phi_{N_c} \in (X_{N_c})^N \cap \mathcal{M}^{\Phi^0} \;\middle|\; \|\Phi_{N_c} - \Phi^0\|_{H^1_\#} \leq r^0 \right\},$$

with

$$\mathcal{M}^{\Phi^0} = \left\{ \Psi \in \mathcal{M} \;\middle|\; \|\Psi - \Phi^0\|_{L^2_\#} = \min_{U \in \mathcal{U}(N)} \|U\Psi - \Phi^0\|_{L^2_\#} \right\},$$

where $\mathcal{U}(N)$ is the set of all unitary transforms in $\mathbb{R}^N$. Besides,

$$\begin{aligned} \|\Phi^0_{N_c} - \Phi^0\|_{H^s_\#} &\leq C_s N_c^{-(2-s)} \|\Pi_{N_c}\Phi^0 - \Phi^0\|_{H^2_\#}, \\ |\epsilon^0_{i,N_c} - \epsilon^0_i| &\xrightarrow[N_c\to\infty]{} 0, \\ \gamma \, \|\Phi^0_{N_c} - \Phi^0\|^2_{H^1_\#} \leq I^{\mathrm{KS}}_{0,N_c} - I^{\mathrm{KS}}_0 &\leq C \, \|\Phi^0_{N_c} - \Phi^0\|^2_{H^1_\#}, \end{aligned}$$

for all $0 \leq s \leq 2$, and for some constants $\gamma > 0$, $C_s \geq 0$, and $C \geq 0$ independent of N_c, where the ϵ^0_i, $\epsilon^0_{i,N_c}, 1 \leq i \leq N$, are the N lowest eigenvalues (counting multiplicities) of the Hamiltonians $\mathcal{H}^0 = H_{\rho_{[\Phi^0]}}$ and $\mathcal{H}_{N_c} = H_{\rho_{[\Phi_{N_c}]}}$ respectively.

Stronger results can be obtained under additional assumptions on the exchange-correlation functional. We refer to [44] for the details.

7.4 A post-processing based on perturbation theory

In this section, we propose a post-processing that is based upon the fact that the exact solution of the Kohn–Sham problem can be interpreted as a perturbation of the approximate solution. We assume that the tolerance parameter η defined in the SCF-procedure is sufficiently small. Given the result of the converged SCF procedure, i.e. given the eigenfunctions $\Phi_{N_c} = (\phi_{1,N_c}, \ldots, \phi_{N,N_c})^T$ and eigenvalues $(\epsilon_{j,N_c})_{j=1,\ldots,N}$ with density ρ_{N_c} of the discretized nonlinear eigenvalue problem in the space X_{N_c}, it is then possible to compute corrections for both the eigenfunctions and the eigenvalues of the problem in order to increase the accuracy.

The key observation is that the exact solution $(\phi^0_j, \epsilon^0_j)_{j=1,\ldots,N}$ satisfies

$$(\mathcal{H}_{N_c} + \mathcal{V}_{N_c} + \mathcal{W}_{N_c})\,\phi^0_j = \epsilon^0_j \, \phi^0_j, \qquad\qquad \int_\Omega \phi^0_i \, \phi^0_j = \delta_{ij},$$

where

$$\begin{aligned} \mathcal{V}_{N_c} &= \Big[V_{\mathrm{ion}} + V_{\mathrm{coul}}(\rho_{N_c}) + V_{\mathrm{xc}}(\rho_{N_c})\Big] - \Pi_{N_c}\Big[V_{\mathrm{ion}} + V_{\mathrm{coul}}(\rho_{N_c}) + V_{\mathrm{xc}}(\rho_{N_c})\Big]\Pi_{N_c}, \\ \mathcal{W}_{N_c} &= \Big[V_{\mathrm{coul}}(\rho^0) + V_{\mathrm{xc}}(\rho^0)\Big] - \Big[V_{\mathrm{coul}}(\rho_{N_c}) + V_{\mathrm{xc}}(\rho_{N_c})\Big]. \end{aligned}$$

With these definitions, we obtain that

$$\mathcal{H}_{[\rho_{N_c}]} = \mathcal{H}_{N_c} + \mathcal{V}_{N_c} \quad \text{and} \quad \mathcal{H}^0 = \mathcal{H}_{N_c} + \mathcal{V}_{N_c} + \mathcal{W}_{N_c}.$$

We then apply the nonlinear Rayleigh–Schrödinger perturbation method (see e.g. [119] and [57] for a mathematical analysis) using $(\phi_{j,N_c}, \epsilon_{j,N_c})_{j=1,\ldots,N}$ as the reference solution and $(\phi^0_j, \epsilon^0_j)_{j=1,\ldots,N}$ as the perturbed solution, in order to build improved approximations $(\widetilde{\phi}_{j,N_c}, \widetilde{\epsilon}_{j,N_c})_{j=1,\ldots,N}$ based upon perturbation arguments. The use of perturbation theory in this setting seems to be new.

For the sake of simplicity, we will explain the argument assuming that all the eigenvalues under consideration are simple, so that non-degenerate perturbation theory applies. The general case can be dealt with the density matrix formalism (see [97]).

Since $\mathcal{V}_{N_c}+\mathcal{W}_{N_c}$ is a compact perturbation of the Hamiltonian $\mathcal{H}_{N_c}$, we can define for $\nu \in [0,1]$ $\mathcal{H}(\nu) = \mathcal{H}_{N_c} + \nu(\mathcal{V}_{N_c} + \mathcal{W}_{N_c})$ so that $\mathcal{H}(0) = \mathcal{H}_{N_c}$ and $\mathcal{H}(1) = \mathcal{H}^0$. Expanding the orbitals and eigenvalues in terms of powers of the perturbation parameter ν and applying Kato's regular perturbation theory results in

$$\phi_j(\nu) = \sum_{\ell=0}^{\infty} \nu^\ell \, \phi_{j,N_c}^{(\ell)}, \qquad \epsilon_j(\nu) = \sum_{\ell=0}^{\infty} \nu^\ell \, \epsilon_{j,N_c}^{(\ell)}, \qquad \text{for all} \quad j = 1, \dots, N,$$

so that (taking $\nu = 1$)

$$\phi_j^0 = \sum_{\ell=0}^{\infty} \phi_{j,N_c}^{(\ell)}, \qquad \epsilon_j^0 = \sum_{\ell=0}^{\infty} \epsilon_{j,N_c}^{(\ell)}, \qquad \text{for all} \quad j = 1, \dots, N,$$

with $(\phi_{j,N_c}^{(0)}, \epsilon_{j,N_c}^{(0)}) = (\phi_{j,N_c}, \epsilon_{j,N_c})$ being the solution of the unperturbed ($\nu = 0$) nonlinear eigenvalue problem (7.3.4). Indeed, one can show [97] that $\mathcal{V}_{N_c}+\mathcal{W}_{N_c}$ is not only $\mathcal{H}_{N_c}$-bounded but that the bound tends to 0 when $N_c \to +\infty$. In consequence, the convergence radii of the above series tend to infinity when N_c increases, so that we can guarantee convergence for $\nu = 1$ and N_c sufficiently large.

Further, incorporating *a priori* results stating that the H^1-norm of the first-order perturbation of the orbitals generated by $\mathcal{W}_{N_c}$ is negligible with respect to the one generated by $\mathcal{V}_{N_c}$ allows us to consider only the latter, called $\phi_{j,N_c}^{(1,1)}$. Then, a simple calculation shows that the first-order correction to the eigenfunctions $(\phi_{j,N_c})_{j=1,\dots,N}$ due to $\mathcal{V}_{N_c}$ is well-defined provided that equation (7.3.3) is verified, and is given by (see [97] for details)

$$\phi_{j,N_c}^{(1,1)} = -(-\frac{1}{2}\Delta|_{X_{N_c}^\perp} - \epsilon_{j,N_c})^{-1} r_j \quad \text{for all} \quad j = 1, \dots, N, \tag{7.4.1}$$

where

$$r_j = \mathcal{H}_{[\rho_{N_c}]}\phi_{j,N_c} - \epsilon_{j,N_c}\phi_{j,N_c} = (\mathcal{H}_{N_c} + \mathcal{V}_{N_c})\phi_{j,N_c} - \epsilon_{j,N_c}\phi_{j,N_c} \in X_{N_c}^\perp,$$

is the residual of the eigenvalue problem, which can also be written as

$$r_j = \left(-\frac{1}{2}\Delta + V_{\text{ion}} + V_{\text{coul}}(\rho_{N_c}) + V_{\text{xc}}(\rho_{N_c}) - \epsilon_{j,N_c}\right)\phi_{j,N_c}. \tag{7.4.2}$$

Note that as $\phi_{j,N_c}^{(1,1)} \in X_{N_c}^\perp$, and $-(-\frac{1}{2}\Delta|_{X_{N_c}^\perp} - \epsilon_{j,N_c})^{-1}$ is a diagonal operator in the Fourier basis, the first-order correction to the j-th eigenvector $\phi_{j,N_c}^{(1,1)}$ is easy to compute.

The first order correction to the eigenvalue originating from $\mathcal{V}_{N_c}$ vanishes, and the one originating from $\mathcal{W}_{N_c}$ is small from *a priori* results. The computable part of the second order correction is (using Dirac's bra-ket notation)

$$\epsilon_{j,N_c}^{(2,1)} = (\phi_{j,N_c}^{(1,1)}, r_j)_{L^2_\#} = -\langle r_j|(-\frac{1}{2}\Delta|_{X_{N_c}^\perp} - \epsilon_{j,N_c})^{-1}|r_j\rangle.$$

Hence $\phi_{j,N_c}^{(1,1)}$ and $\epsilon_{j,N_c}^{(2,1)}$ are well-defined (provided that N_c is large enough).

We then define the post-processed solution as

$$\widetilde{\phi}_{j,N_c} = \phi_{j,N_c} + \phi_{j,N_c}^{(1,1)}, \qquad \widetilde{\epsilon}_{j,N_c} = \epsilon_{j,N_c} + \epsilon_{j,N_c}^{(2,1)}, \qquad \text{and} \qquad \widetilde{\rho}_{N_c} = \rho_{[\widetilde{\Phi}_{N_c}]}, \tag{7.4.3}$$

with $\widetilde{\Phi}_{N_c} = (\widetilde{\phi}_{1,N_c}, \ldots, \widetilde{\phi}_{N,N_c})^T$. Therefore, the post-processed ground state energy can be provided, either following (7.2.8) by

$$\tilde{I}^{\mathrm{KS}}_{0,N_c} = \sum_{i=1}^{N} \int_\Omega |\nabla \widetilde{\phi}_{i,N_c}|^2 + \int_\Omega V_{\mathrm{local}}\, \widetilde{\rho}_{N_c} + 2 \sum_{i=1}^{N} \langle \widetilde{\phi}_{i,N_c} | V_{\mathrm{nl}} | \widetilde{\phi}_{i,N_c} \rangle + \frac{1}{2} D_\Omega(\widetilde{\rho}_{N_c}, \widetilde{\rho}_{N_c}) + E^c_{\mathrm{xc},\Omega}(\widetilde{\rho}_{N_c}), \tag{7.4.4}$$

or following the double counting formula (7.3.5) by

$$\tilde{I}^{\mathrm{KS,DC}}_{0,N_c} = 2 \sum_{i=1}^{N} \widetilde{\epsilon}_{j,N_c} - \frac{1}{2} D_\Omega(\widetilde{\rho}_{N_c}, \widetilde{\rho}_{N_c}) + E^c_{\mathrm{xc}}(\widetilde{\rho}_{N_c}) - \int_\Omega V_{\mathrm{xc}}(\widetilde{\rho}_{N_c})\, \widetilde{\rho}_{N_c}. \tag{7.4.5}$$

Even though definitions (7.2.8) and (7.3.5) are equivalent for the discrete ground states, i.e. for the exact solutions to problem (7.3.4), the two perturbed energies computed by (7.4.4) and (7.4.5) are not equal, and lead to different numerical results. From a computational viewpoint, the time needed to compute two energies (7.4.4) and (7.4.5) is almost the same. But it seems that (7.4.4) gives better approximations for the energy than the double counting formula (7.4.5). The numerical performance of these two formulas will be discussed in Section 7.5.

From a theoretical viewpoint, we can state the following result for the perturbed energy (7.4.4).

Theorem 7.4.1 ([97]). *Let I^{KS}_{0,N_c} be the planewave approximation of the ground state energy be defined by* (7.3.1) *and $\tilde{I}^{\mathrm{KS}}_{0,N_c}$ the post-processed approximation given by* (7.4.4). *Then, under sufficient regularity assumptions (see [97]), there exists a constant $C > 0$, independent of N_c, such that*

$$\left| \tilde{I}^{\mathrm{KS}}_{0,N_c} - I^{\mathrm{KS}}_0 \right| \leq C\, N_c^{-2} \left| I^{\mathrm{KS}}_{0,N_c} - I^{\mathrm{KS}}_0 \right|,$$

where I^{KS}_0 is the exact ground state energy, defined in (7.2.9).

Although the above inequality might not be not optimal (some numerical results seem to show that the improvement factor is better than N_c^{-2}), this result clearly indicates that the perturbation method leads to a substantial improvement of the accuracy of the energy in the asymptotic regime, that is when N_c goes to infinity. Note, though, that most calculations are performed in practice in the pre-asympotic regime, with moderate values of N_c. It is therefore important to check numerically that the perturbation method performs well also in the pre-asymptotic regime, which will be done in the next section. Let us also mention that the above inequality is concerned with the energy only. Obtaining similar estimates for other properties, and in particular for atomic forces, is a difficult task, which is work in progress.

From an implementational viewpoint, the residual r_j, defined in (7.4.2), which is an infinite-dimensional object belonging to $\Pi^{\perp}_{N_c}$, is represented on a discrete space $X_{N_c,\mathrm{res}}$ based on some $E_{c,\mathrm{res}} \geq E_c$ which, in turn, corresponds to a certain $N_{c,\mathrm{res}} \geq N_c$. Further, observing that the density ρ_{N_c} belongs to X_{2N_c} (corresponding to $4E_c$ and $2N_c$), the potential $V_{\mathrm{coul}}(\rho_{N_c})$ therefore belongs to X_{2N_c} as well. The post-processing requires that the potentials V_{ion} and $V_{\mathrm{xc}}(\rho_{N_c})$ can be expressed in the larger space with cutoff $E_{c,\mathrm{res}}$ so that full knowledge of all the modes of the residual lying in $X_{N_c,\mathrm{res}}$ are accessible. In practice, it might be simpler to obtain V_{ion} and $V_{\mathrm{xc}}(\rho_{N_c})$ as elements of $X_{N_c,\mathrm{res}}$ directly in order to avoid too many data structures associated with different cutoffs. The computation of this residual requires additional Fast Fourier Transforms (FFT) on the finer grid corresponding to $E_{\mathrm{cut,res}}$. Indeed, applying the Hamiltonian $\mathcal{H}_{[\rho_{N_c}]}$ to the orbitals ϕ_{j,N_c} requires two additional FFT's on the fine grid for each orbital.

7.5 Numerical results

We present here some results on a small molecule as a proof of concept. The alanine molecule which has 18 valence electron pairs is considered. The computation of the planewave approximation is based on KSSOLV [252], a Matlab toolbox for solving the Kohn–Sham equations. Troullier–Martins pseudopotentials [235] are considered, in combination with the Perdew–Zunger (PZ81) LDA-functional [152, 201]. Note that we obtained qualitatively the same results with the CO_2 molecule as well as the benzene molecule.

In all what follows, the computed solutions are compared to a reference solution, which is a solution computed on a very fine grid with a kinetic energy cutoff $E_{\mathrm{c}}^{\mathrm{ref}}$ of 400 atomic units (a.u.). A coarse solution (labelled "without perturbation") is computed on a grid with cutoff E_{c}, and the post-processed approximations given by the perturbation theory are computed on a grid with fine cutoff $E_{\mathrm{c,res}}$ (labelled "with perturbation") (7.4.3). Note that the components of the Kohn–Sham orbitals on the coarse grid are not modified by the post-processing. The coefficients computed through the perturbation process correspond to basis functions with wave numbers larger than E_{c}. We denote by λ the relation between the coarse and fine cutoffs, i.e. $\lambda = \dfrac{E_{\mathrm{c,res}}}{E_{\mathrm{c}}}$.

In the following, we simulate three scenarios. In each case, we compare the errors on the energies of the solutions with perturbation computed with either formula (7.4.4) or formula (7.4.5) to the errors on the energies of the solutions without perturbation. First, we fix a coarse cutoff E_{c}, and we vary the fine cutoff $E_{\mathrm{c,res}}$, which corresponds to varying the parameter λ. In the second case, we fix a fine cutoff $E_{\mathrm{c,res}}$ and we vary the coarse cutoff E_{c}, which also corresponds to varying the parameter λ. Finally, we fix λ and we vary simultaneously the cutoffs E_{c} and $E_{\mathrm{c,res}}$. Thus we can observe the improvement in the energy when applying the perturbation theory in these different frameworks. This enables us to find the best compromises between accuracy and computational resources, both in time and memory.

7.5.1 Simulations with a constant E_{c}

Here, we fix E_{c}, hence the dimension of the space in which we compute the coarse solution. We then compute the post-processed corrections in spaces of different sizes characterized by different $E_{\mathrm{c,res}}$ to obtain a more accurate solution. We observe on Figure 7.1 that for all the tested $E_{\mathrm{c,res}}$, the energy of the solution with perturbative corrections is more precise than the coarse solution as the relative energy error is improved by one to two orders of magnitude, depending on the energy formula used to compute the perturbed energy. Indeed, the formula (7.4.4) gives an energy always closer to the exact energy than the formula (7.4.5) based on double counting.

The relative time increases when the parameter λ increases, as it is more expensive to compute the perturbed solution on a larger grid than on a smaller one. However, the time necessary to compute the perturbative corrections corresponds approximately to 1% to 12% of the cost to compute the coarse solution, which is indeed very little. The cost to compute the perturbed energy is a little higher for the energy based on (7.4.4) than for the double counting energy (7.4.5), but the difference in negligible compared to the total computing time. Note that the cost in memory is also higher when $E_{\mathrm{c,res}}$ is large, as the perturbed solution is then computed on a larger grid.

The improvement of the solution seems to be constant for $\lambda \geq 3$, which means that the main

improvement of the post-processing is due to coefficients corresponding to planewaves with kinetic energy slightly larger than E_c. We therefore conclude that it seems useless to use a large λ and hence a large $E_{c,res}$, to post-process the solution as the improvement is negligible while the cost in time and memory increases. The best choice of λ seems to vary, but a good choice seems to be around 2.

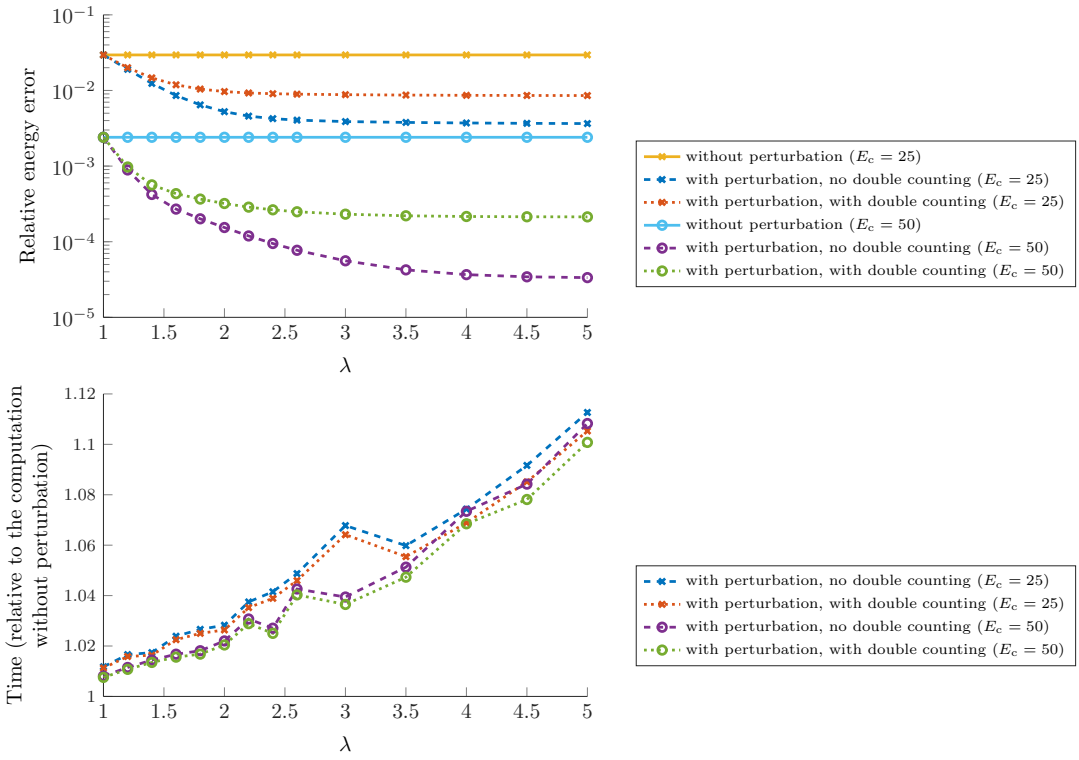

Figure 7.1 – Simulations with a constant E_c (alanine molecule).

7.5.2 Simulations with a constant $E_{c,res}$

In this case, we fix the fine grid used for the computation of the perturbed solution but we compute it from different coarse solutions obtained with different values of E_c. This corresponds to a case where the limiting parameter is the memory, and so the fine grid in which we can compute the perturbed solution is fixed. It is shown on Figure 7.2 that it is possible to get the same accuracy in energy by doing a small computation on a coarse grid and then post-processing the coarse solution on a grid with parameter $E_{c,res}$. The computational cost is in this case much reduced if we compare it to a classical approach consisting of computing the solution on the grid with cutoff $E_{c,res}$.

First, the computational time needed to compute the perturbed solution from the unperturbed one is once again negligible compared to the time needed to compute the coarse solution, at least for small values of the parameter λ. The cost in terms of CPU time is also negligible using either (7.4.4) or (7.4.5) for the post-processed energy computation.

Second, we observe that for $E_{c,res} = 100$ a.u. and $E_{c,res} = 200$ a.u., the relative energy error for the solution with perturbation is always smaller than for the solution without perturbation by a factor of about 10 to 50. When λ increases, the error becomes larger, which is expected since E_c decreases. Moreover it seems that the improvement is better for $E_{c,res} = 200$ than for

$E_{\mathrm{c,res}} = 100$.

The post-processed energy $\tilde{I}^{\mathrm{KS}}_{0,N_{\mathrm{c}}}$ given by (7.4.4) is in this case always closer to the exact energy than the energy $\tilde{I}^{\mathrm{KS,DC}}_{0,N_{\mathrm{c}}}$ given by (7.4.5). Indeed, the energy error for $\tilde{I}^{\mathrm{KS,DC}}_{0,N_{\mathrm{c}}}$ is often 2 to 5 times larger than the energy error based on $\tilde{I}^{\mathrm{KS}}_{0,N_{\mathrm{c}}}$. Hence, the post-processed energy $\tilde{I}^{\mathrm{KS}}_{0,N_{\mathrm{c}}}$ improves the energy error of the solution on the coarse grid by a factor 10 to 100 whereas the post-processed energy based on double counting formula (7.4.5) improves the energy error only by a factor between 4 and 12.

Moreover, it is interesting to compare the time needed to reach a given accuracy for the method with or without perturbation. For example, for $\lambda = 3$ and $E_{\mathrm{c,res}} = 200$ a.u., the time needed to compute the perturbed solution is about 2600 seconds which is 7 times less than the time needed to compute the self-consistent solution on this grid (corresponding to $\lambda = 1$), whereas the energy error $\tilde{I}^{\mathrm{KS}}_{0,N_{\mathrm{c}}}$ given by (7.4.4) is of the same order of magnitude. Hence, when the solution on the coarse grid is not too crude, the perturbation theory enables us to significantly improve the solution at very low extra cost.

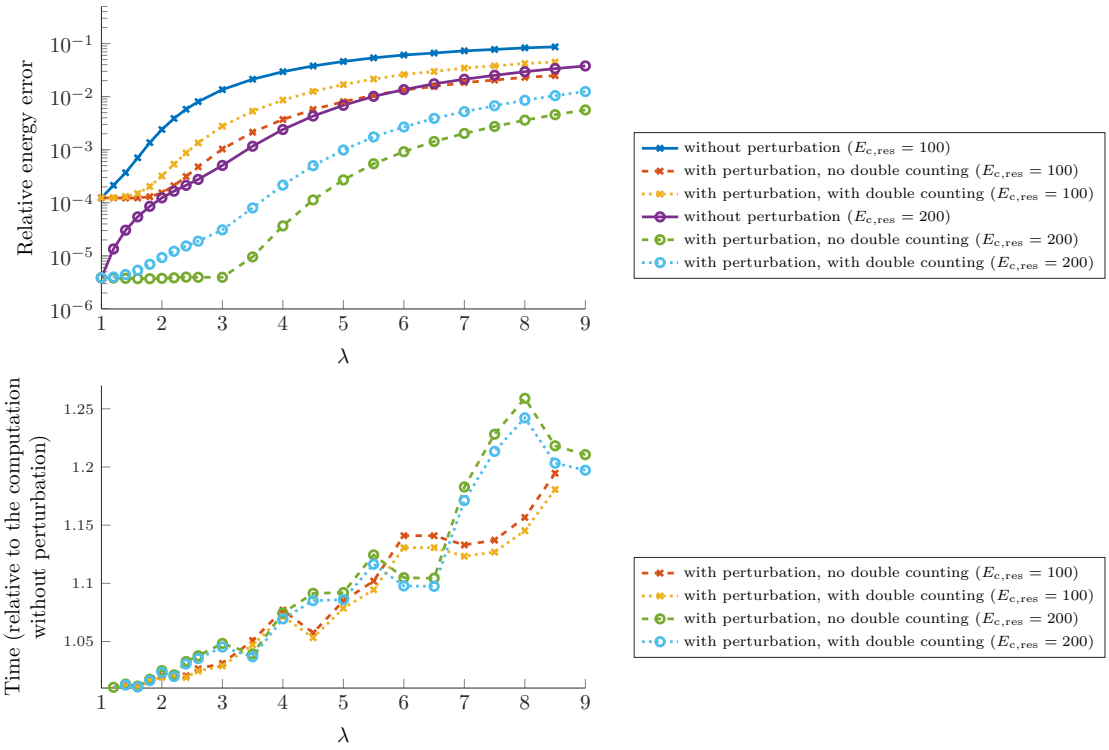

Figure 7.2 – Simulations with a constant $E_{\mathrm{c,res}}$ (alanine molecule).

7.5.3 Simulations with a constant λ

In this case, we fix the proportionality constant between E_{c} and $E_{\mathrm{c,res}}$. For different values of E_{c}, we consider the relative energy error without or with the perturbation corrections (Figure 7.3).

For all E_{c} greater than 20 a.u., there is an improvement in energy when performing the perturbation corrections, using one or the other energy formula. The relative energy error of the solutions with perturbation is indeed lower than for the solutions without perturbation by a

factor between 2 and 100. Once again, the energy formula (7.4.4) gives better results than the formula (7.4.5). The computational time increases as E_c increases and the time necessary to compute the solution with perturbation is still very small compared to the computation of the coarse solution, and is less than 3% of the total computational time for $\lambda = 1.4$ and $\lambda = 2$, and no more than 8% of the total computational time for $\lambda = 3$. The computational time is always smaller for the energy formula (7.4.5) based on double counting than for the energy formula (7.4.4).

For all the values of E_c tested, the best improvement is given either by $\lambda = 2$ or $\lambda = 1.4$, which drives us into using mostly small values of λ, as the computations are also less expensive in memory.

Concerning the two possible energy formulas for computing the post-processed energy, the energy formula (7.4.4) gives better results than (7.4.5). Note that the theoretical results obtained so far are only concerned with formula (7.4.4). However the formula (7.4.5) based on double counting is a little cheaper to compute to evaluate and also leads to an improvement of the energy.

In conclusion, it seems that whenever the coarse solution is accurate enough, the use of the perturbation method to post-process the solution improves the energy by a factor typically of more than 10. Since the computational cost of this post-process is negligible, it is possible to get the same accuracy as a solution computed on a large grid by first doing a smaller computation on a coarse grid and then post-processing the coarse solution using the perturbation method. This approach is much cheaper than the computation of the solution directly on a grid of size $E_{c,res}$.

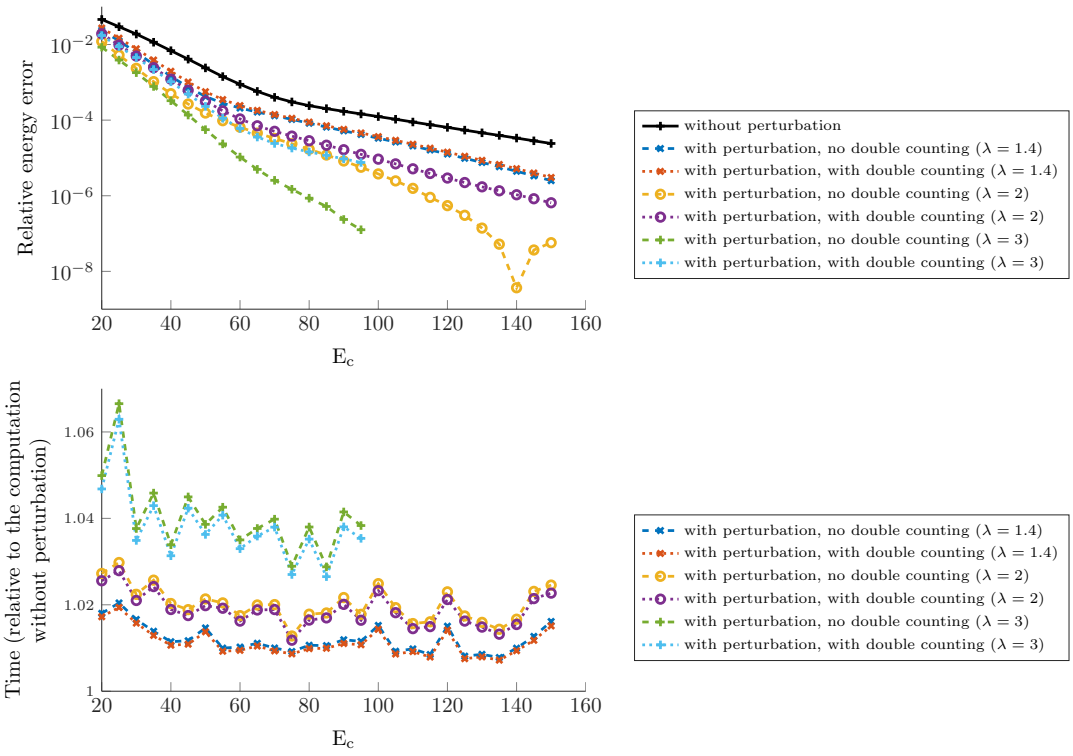

Figure 7.3 – Simulations with a constant λ (alanine molecule).

Acknowledgements

This work was partially undertaken in the framework of CALSIMLAB, supported by the public grant ANR-11-LABX- 0037-01 overseen by the French National Research Agency (ANR) as part of the Investissements d'avenir program (reference: ANR-11-IDEX-0004-02). Financial support by the ANR grants Manif and Becasim is also acknowledged.

Chapter 8

Post-processing of the plane-wave approximation of linear Schrödinger equations

We expose in this chapter the results of [97]. This work was done in collaboration with with Eric Cancès, Yvon Maday, Benjamin Stamm and Martin Vohralík.

Abstract

In this article, we prove *a priori* error estimates for the perturbation-based post-processing of the plane-wave approximation of Schrödinger equations introduced and tested numerically in previous works [49, 50]. We consider here a Schrödinger operator $\mathcal{H} = -\frac{1}{2}\Delta + \mathcal{V}$ on $L^2(\Omega)$, where Ω is a cubic box with periodic boundary conditions, and where $\mathcal{V}$ is a multiplicative operator by a regular-enough function $\mathcal{V}$. The quantities of interest are, on the one hand, the ground-state energy defined as the sum of the lowest N eigenvalues of $\mathcal{H}$, and, on the other hand, the ground-state density matrix, that is the spectral projector on the vector space spanned by the associated eigenvectors. Such a problem is central in first-principle molecular simulation, since it corresponds to the so-called linear subproblem in Kohn–Sham density functional theory (DFT). Interpreting the exact eigenpairs of $\mathcal{H}$ as perturbations of the numerical eigenpairs obtained by a variational approximation in a plane-wave (i.e. Fourier) basis, we compute first-order corrections for the eigenfunctions, which are turned into corrections on the ground-state density matrix. This allows us to increase the accuracy of both the ground-state energy and the ground-state density matrix at a low computational extra-cost. Indeed, the computation of the corrections only requires the computation of the residual of the solution in a larger plane-wave basis and two Fast Fourier Transforms per eigenvalue.

8.1 Introduction

First-principle molecular simulation is a major tool to predict the properties of matter from the atomic to the macroscopic scales. It is widely used in different fields such as chemistry, condensed matter physics, or materials science. As a main advantage, it requires no empirical parameter except a few fundamental constants of physics (the reduced Planck constant $\hbar$, the electron mass m_e, the elementary charge e, the dielectric permittivity of the vacuum ε_0, the Boltzmann constant k_B), as well as the masses and atomic numbers of the nuclei contained in the system under consideration.

At this scale, matter is described as a system of nuclei and electrons, whose dynamics is modeled by a time-dependent many-body Schrödinger equation. This equation, which is an evolution partial differential equation on a $3^{(M+N)}$-dimensional space, where M is the number of nuclei and N the number of electrons, is way too costly to be solved in practice when $(M + N)$ exceeds 2 or 3. Hence approximations have to be resorted to. First, in almost all molecular simulations, the nuclei, which are thousands times heavier than electrons, are considered as point-like classical particles, and the electrons are supposed to be, at each time t, in their ground-state. This is called the Born–Oppenheimer approximation [31].

Many different approaches have then been proposed to compute the electronic ground-state. The most popular ones can be classified into three main classes:

- wavefunction methods, among which the Hartree–Fock, post Hartree–Fock, and multi-reference methods (see [135], and [45] for a mathematical introduction);
- density functional theory (DFT) methods, consisting of orbital-free and Kohn–Sham models;
- quantum Monte Carlo methods [166, 165].

The Kohn–Sham models [93, 152] are the most widely used in physics and chemistry, as they provide a good compromise between accuracy and computational cost. In condensed matter physics and materials science, most Kohn–Sham calculations are performed in a rectangular box Ω, called a supercell, with periodic boundary conditions (Born–von Karman PBC). The most common method to discretize the Kohn–Sham equations then is to use a variational approximation in a plane-wave (Fourier) basis. For very large systems, unfortunately, using a fine discretization basis is too expensive, while using a coarse discretization basis leads to insufficiently accurate results.

In order to limit the computational cost of the method while preserving the quality of the numerical results, several post-processing methods have been proposed. Usually, the approach is to perform a full computation in a coarse basis, for which the computational cost is not excessively high, and then to make some not-too-expensive computation in a finer basis leading to a substantial improvement in accuracy.

In [50], we introduced a new post-processing method for periodic Kohn–Sham calculations in a plane-wave basis, leading to a significant gain in accuracy at a very limited computational extra-cost. This approach is based on the Rayleigh–Schrödinger perturbation method, considering the exact Kohn–Sham ground-state as a perturbation of the approximate ground-state computed in a finite basis set. Theoretical estimates in the asymptotic regime for the energy were announced and illustrated by numerical simulations. These simulations showed that the accuracy on the

ground-state energy could be improved by a factor of 10 to 100 at a very limited extra-cost (about 3%).

In this contribution, we focus on the linear subproblem of the Kohn–Sham model. We apply the post-processing of the plane-wave approximation of Schrödinger equations introduced in [49, 50] and present the proof of the improved accuracy. The linear subproblem consists in computing the rank–N ground-state density matrix γ_0 of a linear Schrödinger operator $\mathcal{H} = -\frac{1}{2}\Delta + \mathcal{V}$, acting on the space $L^2_\#(\Omega)$ of real-valued square-integrable periodic functions on $\mathbb{R}^3$ with Ω as a periodic cell. For $\mathcal{V} \in L^2_\#(\Omega)$, the Hamiltonian $\mathcal{H}$ is diagonalizable in an orthonormal basis and its eigenvalues $\lambda_1^0 \leq \lambda_2^0 \leq \cdots$ (counting multiplicities) form a non-decreasing sequence of real numbers that tends to $+\infty$. Denoting by $(\phi_i^0)_{i\geq 1}$ an orthonormal basis of associated eigenvectors, and assuming that there is a gap $g := \lambda_{N+1}^0 - \lambda_N^0 > 0$ between the N^{th} and the $(N+1)^{\text{st}}$ eigenvalues of $\mathcal{H}$, the rank–N ground-state density matrix is defined, using Dirac's bra-ket notation, as

$$\gamma_0 = \sum_{i=1}^{N} |\phi_i^0\rangle\langle\phi_i^0|.$$

It is therefore the orthogonal projector on the N-dimensional vector subspace of $L^2_\#(\Omega)$ spanned by the eigenvectors associated with the lowest N eigenvalues of $\mathcal{H}$. The ground-state energy is then defined as the scalar quantity

$$\mathcal{E}_0 = \sum_{i=1}^{N} \lambda_i^0.$$

Note that $\mathcal{E}_0$ is not equal to the Kohn–Sham ground-state energy when $\mathcal{H}$ is the Kohn–Sham Hamiltonian due to nonlinear effects, called double-counting in the chemistry and physics literatures.

Our main result is summarized in Theorem 8.4.1. We show that, in the asymptotic regime where the discretization space is large enough, the convergence rates of both the post-processed ground-state density matrix (built from the post-processed eigenvectors defined in [49]) and the post-processed ground-state energy are improved. Note that we do not make any non-degeneracy assumption on the lowest N eigenvalues of $\mathcal{H}$; only the presence of a positive gap between the N^{th} and the $(N+1)^{\text{st}}$ eigenvalues of $\mathcal{H}$ is required. As in [49, 50], our approach strongly relies on the fact that, in plane-wave approximations, the kinetic energy operator $-\frac{1}{2}\Delta$, which is the leading term in the Hamiltonian $\mathcal{H}$, commutes with the orthogonal projector on the discretization space.

In general, our approach falls into the category of two-grid methods, which have been applied to the case of a linear eigenvalue problem in [251]. It has then been extended to a class of elliptic nonlinear eigenvalue problems in [49]. In this paper, the *nonlinear* eigenvalue problem is first solved on a coarse grid, and then a *linear* eigenvalue problem or a boundary problem is solved on a finer grid. Other post-processing methods have been proposed for nonlinear eigenvalue problems, for example in [184] for the Hartree–Fock problem (see also the references therein).

This article is organized as follows. In Section 8.2.1, we present in detail the linear subproblem of the Kohn–Sham model, as well as the characterization of γ_0 as the unique solution to some constrained optimization problem, and some other useful classical results. In Section 8.2.3, we describe the plane-wave discretization of this optimization problem. In Section 8.2.4, we translate the *a priori* error analysis results of [44] into the density matrix formalism. Our post-processing method based on Rayleigh–Schrödinger perturbation theory is described in Section

8.3. In Section 8.4.1, we present the main results of this paper, i.e.an improved convergence rate on the post-processed ground-state density matrix and energy. The proofs are given in Section 8.4.2. Some numerical simulations are presented in Section 8.5. The case of the nonlinear Kohn–Sham model will be dealt with in a forthcoming paper [97].

8.2 Post-processing for the Kohn–Sham linear subproblem

In order to simplify the notation, we consider a cubic lattice $\mathcal{R} = L\mathbb{Z}^3$ ($L > 0$) corresponding to the supercell $\Omega = [0, L)^3$, but all our arguments straightforwardly apply to the general case of a lattice with lower or no point symmetry. For $1 \leq p \leq \infty$ and $s \in \mathbb{R}_+$, we denote by

$$L^p_\#(\Omega) := \left\{u \in L^p_{\rm loc}(\mathbb{R}^3, \mathbb{R}) \mid u \text{ is } \mathcal{R}\text{-periodic}\right\},$$
$$H^s_\#(\Omega) := \left\{u \in H^s_{\rm loc}(\mathbb{R}^3, \mathbb{R}) \mid u \text{ is } \mathcal{R}\text{-periodic}\right\},$$

the spaces of real-valued $\mathcal{R}$-periodic L^p and H^s functions, and by $\mathcal{L}(L^2_\#)$ the vector space of the bounded linear operators on $L^2_\#(\Omega)$.

8.2.1 Problem setting

Let $N \in \mathbb{N}^*$ and $\mathcal{V} \in L^2_\#(\Omega)$. In Kohn–Sham models, N is the number of electrons (or of electron pairs in closed-shell models) in the simulation cell, and $\mathcal{V}$ is an approximation of the Kohn–Sham effective potential. Let $\mathcal{H}$ be the operator on $L^2_\#(\Omega)$ with domain $H^2_\#(\Omega)$ defined by

$$\forall u \in H^2_\#(\Omega), \quad \mathcal{H}\, u = -\frac{1}{2}\Delta u + \mathcal{V} u.$$

It is well-known that the operator $\mathcal{H}$ is self-adjoint, bounded below, with compact resolvent. It can therefore be diagonalized in an orthonormal basis: there exists a non-decreasing sequence $(\lambda^0_i)_{i\geq 1}$ of real numbers and an orthonormal basis $(\phi^0_i)_{i\geq 1}$ of $L^2_\#(\Omega)$ consisting of functions of $H^2_\#(\Omega)$ such that

$$\forall i \geq 1, \quad \mathcal{H}\,\phi^0_i = \lambda^0_i \phi^0_i.$$

We denote by $\Phi^0 = (\phi^0_1, \cdots, \phi^0_N)^T$ and $\Lambda^0 = \mathrm{diag}(\lambda^0_1, \cdots, \lambda^0_N)$.

A key assumption for our analysis is the following:

Assumption 8.2.1. *There is a gap between the $N^{\rm th}$ and the $(N+1)^{\rm st}$ eigenvalues of $\mathcal{H}$, i.e.*

$$g := \lambda^0_{N+1} - \lambda^0_N > 0.$$

We denote by $\epsilon_{\rm F} := \frac{\lambda^0_N + \lambda^0_{N+1}}{2}$ the Fermi level. Note that, in this setting, any real number in the range $(\lambda^0_N, \lambda^0_{N+1})$ is an admissible Fermi level.

As already mentioned in the introduction, the purpose of the linear subproblem is to compute two quantities of interest:

1. the ground-state density matrix

$$\boxed{\gamma_0 := \mathbb{1}_{(-\infty, \epsilon_{\rm F}]}(\mathcal{H}) = \sum_{i=1}^{N} |\phi^0_i\rangle\langle\phi^0_i|;}$$

2. the ground-state energy

$$\boxed{\mathcal{E}_0 := \mathrm{Tr}\,(\mathcal{H}\,\gamma_0) = \sum_{i=1}^{N} \lambda_i^0,}$$

where Tr denotes the trace, and will be properly introduced in Section 8.2.2.

The linear subproblem can be formulated as a variational problem in several ways. First, introducing the quadratic form

$$H^1_{\#}(\Omega) \ni \psi \mapsto \langle\psi|\mathcal{H}\,|\psi\rangle := \frac{1}{2}\int_\Omega |\nabla\psi|^2 + \int_\Omega \mathcal{V}\,|\psi|^2 \in \mathbb{R}$$

associated with $\mathcal{H}$, the energy functional $\mathcal{E}$ defined by

$$\forall \Psi = (\psi_1, \ldots, \psi_N)^T \in \left[H^1_{\#}(\Omega)\right]^N, \quad \mathcal{E}(\Psi) := \sum_{i=1}^{N} \langle\psi_i|\mathcal{H}\,|\psi_i\rangle = \sum_{i=1}^{N} \left(\frac{1}{2}\int_\Omega |\nabla\psi_i|^2 + \int_\Omega \mathcal{V}\,|\psi_i|^2\right), \tag{8.2.1}$$

and the (infinite-dimensional) Stiefel manifold

$$\mathcal{M} = \left\{ \Psi = (\psi_1, \ldots, \psi_N)^T \in \left[H^1_{\#}(\Omega)\right]^N \;\middle|\; \forall i,j = 1,\ldots,N,\ \int_\Omega \psi_i\psi_j = \delta_{ij} \right\}, \tag{8.2.2}$$

we have

$$\mathcal{E}_0 = \inf\left\{\mathcal{E}(\Psi),\ \Psi \in \mathcal{M}\right\}. \tag{8.2.3}$$

Besides, $\Phi^0 = (\phi_1^0, \ldots, \phi_N^0)^T$ is a minimizer of (8.2.3). Note that Φ^0 is not the unique minimizer of (8.2.3). Indeed, denoting by $O(N) := \left\{U \in \mathbb{R}^{N\times N} \mid U^TU = 1_N\right\}$ the orthogonal group in dimension N (1_N is the identity matrix of rank N), we have

$$\forall \Psi \in \mathcal{M}, \quad \forall U \in O(N), \quad U\Psi \in \mathcal{M}, \quad \text{and} \quad \mathcal{E}(U\Psi) = \mathcal{E}(\Psi). \tag{8.2.4}$$

Therefore, $U\Phi^0$ is a minimizer of (8.2.3) for all $U \in O(N)$. In fact, under Assumption 8.2.1, the set of minimizers of (8.2.3) is exactly equal to $O(N)\Phi^0$. For the sake of completeness, let us recall the proof of this elementary, but key, property. Let $\Psi = (\psi_1, \cdots, \psi_N)^T$ be a critical point of (8.2.3). The first-order optimality conditions satisfied by Ψ read

$$\forall i,j = 1,\ldots,N, \quad \mathcal{H}\,\psi_i = \sum_{j=1}^{N} \lambda_{ij}\psi_j, \quad \int_\Omega \psi_i\psi_i = \delta_{ij}.$$

The $N \times N$ symmetric matrix $\Lambda = (\lambda_{ij})_{i,j=1,\ldots,N}$ is the Lagrange multiplier of the matrix constraint $\int_\Omega \psi_i\psi_j = \delta_{ij}$. It is not diagonal in general. On the other hand, since it is symmetric, there exists $U \in O(N)$ such that $U\Lambda U^T = \mathrm{diag}(\lambda_1, \cdots, \lambda_N)$ with $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_N$. Then, $\Phi = (\phi_1, \cdots, \phi_N)^T := U\Psi$ also is a critical point of (8.2.3) with the same energy as Ψ, and we have

$$\forall 1 \le i,j \le N, \quad \mathcal{H}\,\phi_i = \lambda_i\phi_i, \quad \int_\Omega \phi_i\phi_j = \delta_{ij} \quad \text{and} \quad \mathcal{E}(\Phi) = \mathcal{E}(\Psi) = \sum_{i=1}^{N} \lambda_i.$$

For Ψ to be a minimizer of (8.2.3), we must have $\lambda_i = \lambda_i^0$ for all $1 \le i \le N$. Under Assumption 8.2.1, we have in addition $\mathrm{Span}(\psi_1, \cdots, \psi_N) = \mathrm{Span}(\phi_1, \cdots, \phi_N) = \mathrm{Span}(\phi_1^0, \cdots, \phi_N^0) = \mathrm{Ran}(\gamma_0)$. Therefore, there exists $U \in O(N)$ such that $\Psi = U\Phi_0$.

To get rid of the gauge invariance (8.2.4), it is convenient to reformulate problem (8.2.3) in terms of density matrices. Introducing the (infinite-dimensional) Grassmann manifold

$$\Upsilon = \left\{\gamma \in \mathcal{L}(L^2_\#) \;\middle|\; \gamma^* = \gamma,\ \gamma^2 = \gamma,\ \mathrm{Tr}\,(\gamma) = N,\ \mathrm{Tr}\,(-\Delta\gamma) < \infty\right\}, \tag{8.2.5}$$

its convex hull

$$\mathcal{K} = \left\{\gamma \in \mathcal{L}(L^2_\#) \;\middle|\; \gamma^* = \gamma,\ 0 \le \gamma \le 1,\ \mathrm{Tr}\,(\gamma) = N,\ \mathrm{Tr}\,(-\Delta\gamma) < \infty\right\}, \tag{8.2.6}$$

and the energy functional E defined on $\mathcal{K}$ by

$$\forall \gamma \in \mathcal{K}, \quad E(\gamma) = \mathrm{Tr}\,(\mathcal{H}\,\gamma), \tag{8.2.7}$$

it holds

$$\mathcal{E}_0 = \inf\left\{E(\gamma),\ \gamma \in \Upsilon\right\} \tag{8.2.8}$$

and

$$\mathcal{E}_0 = \inf\left\{E(\gamma),\ \gamma \in \mathcal{K}\right\}. \tag{8.2.9}$$

Besides, under Assumption 8.2.1, γ_0 is the unique minimizer of both (8.2.8) and (8.2.9). Here γ^* denotes the adjoint of γ, $0 \le \gamma \le 1$ means $\forall u \in L^2_\#(\Omega)$, $0 \le \langle u|\gamma u\rangle \le \|u\|^2_{L^2_\#}$. The precise meanings of the terms $\mathrm{Tr}\,(-\Delta\gamma)$ and $\mathrm{Tr}\,(\mathcal{H}\,\gamma)$, as well as the proof of the fact that γ_0 is the unique minimizer of (8.2.8) and (8.2.9), will be given in the next section.

8.2.2 Functional setting

We denote by $\|\cdot\|$ the operator norm on $\mathcal{L}(L^2_\#)$, the space of bounded linear operators on $L^2_\#(\Omega)$. We also need to introduce the Banach space $\mathfrak{S}_1(L^2_\#)$ of trace-class operators on $L^2_\#(\Omega)$ and the Hilbert space $\mathfrak{S}_2(L^2_\#)$ of Hilbert–Schmidt operators on $L^2_\#(\Omega)$, respectively endowed with the norm defined by $\|A\|_{\mathfrak{S}_1(L^2_\#)} := \mathrm{Tr}\,(|A|) = \mathrm{Tr}\,(\sqrt{A^*A})$ and the inner product defined by $(A,B)_{\mathfrak{S}_2(L^2_\#)} := \mathrm{Tr}\,(A^*B)$. We refer to [211, Chapter VI] for an introduction to trace-class and Hilbert–Schmidt operators. Let us just recall here the properties which will be used in the sequel:

- for any orthonormal basis $(e_n)_{n\in\mathbb{N}}$ of $L^2_\#(\Omega)$, we have

$$\forall A \in \mathfrak{S}_1(L^2_\#), \quad \mathrm{Tr}\,(A) = \sum_{n\in\mathbb{N}} \langle e_n|Ae_n\rangle,$$

$$\forall A \in \mathfrak{S}_2(L^2_\#), \quad \|A\|_{\mathfrak{S}_2(L^2_\#)} = \mathrm{Tr}\,(A^*A)^{1/2} = \left(\sum_{n\in\mathbb{N}} \|Ae_n\|^2_{L^2_\#}\right)^{1/2}.$$

  If $A \in \mathcal{L}(L^2_\#)$ is a *positive* operator, that is if for all $u \in L^2_\#(\Omega)$, there holds that $\langle u|Au\rangle \ge 0$, then the value of the sum

$$\mathrm{Tr}\,(A) := \sum_{n\in\mathbb{N}} \langle e_n|Ae_n\rangle \in \mathbb{R}_+ \cup \{+\infty\}$$

  is independent of the choice of the orthonormal basis $(e_n)_{n\in\mathbb{N}}$. If $A \in \mathcal{L}(L^2_\#)$ is *positive and self-adjoint*, then $A \in \mathfrak{S}_1(L^2_\#)$ if and only if $\mathrm{Tr}\,(A) < \infty$;

- $\mathfrak{S}_1(L^2_\#) \subset \mathfrak{S}_2(L^2_\#) \subset \mathcal{L}(L^2_\#)$ and for all $A \in \mathfrak{S}_1(L^2_\#)$,
$$\|A\| \leq \|A\|_{\mathfrak{S}_2(L^2_\#)} \leq \|A\|_{\mathfrak{S}_1(L^2_\#)}; \tag{8.2.10}$$
- for all $A \in \mathfrak{S}_1(L^2_\#)$ and $B \in \mathcal{L}(L^2_\#)$, we have $AB \in \mathfrak{S}_1(L^2_\#)$, $BA \in \mathfrak{S}_1(L^2_\#)$,
$$\mathrm{Tr}\,(AB) = \mathrm{Tr}\,(BA), \quad \|AB\|_{\mathfrak{S}_1(L^2_\#)} \leq \|A\|\,\|B\|_{\mathfrak{S}_1(L^2_\#)}, \quad \|BA\|_{\mathfrak{S}_1(L^2_\#)} \leq \|A\|\,\|B\|_{\mathfrak{S}_1(L^2_\#)}; \tag{8.2.11}$$
- for all $A \in \mathfrak{S}_1(L^2_\#)$, there exists a unique function $\rho_A \in L^1_\#(\Omega)$, called the density associated with the operator A, such that for all $V \in L^\infty_\#(\Omega)$,
$$\mathrm{Tr}\,(AV) = \int_\Omega \rho_A V,$$
where on the left–hand side of the above equality, $V \in \mathcal{L}(L^2_\#)$ denotes the multiplication operator by the function V;
- for all $A \in \mathfrak{S}_2(L^2_\#)$ and $B \in \mathcal{L}(L^2_\#)$, we have $AB \in \mathfrak{S}_2(L^2_\#)$, $BA \in \mathfrak{S}_2(L^2_\#)$,
$$\|AB\|_{\mathfrak{S}_2(L^2_\#)} \leq \|A\|\,\|B\|_{\mathfrak{S}_2(L^2_\#)}, \quad \|BA\|_{\mathfrak{S}_2(L^2_\#)} \leq \|A\|\,\|B\|_{\mathfrak{S}_2(L^2_\#)}; \tag{8.2.12}$$
- for all $A \in \mathfrak{S}_2(L^2_\#)$ and $B \in \mathfrak{S}_2(L^2_\#)$, $AB \in \mathfrak{S}_1(L^2_\#)$, $BA \in \mathfrak{S}_1(L^2_\#)$,
$$\mathrm{Tr}\,(AB) = \mathrm{Tr}\,(BA) \leq \|A\|_{\mathfrak{S}_2(L^2_\#)} \|B\|_{\mathfrak{S}_2(L^2_\#)}. \tag{8.2.13}$$

We set
$$\forall \Psi = (\psi_1, \cdots, \psi_N)^T \in \left[L^2_\#(\Omega)\right]^N, \quad \|\Psi\|_{L^2_\#} := \left(\sum_{i=1}^N \|\psi_i\|^2_{L^2_\#}\right)^{1/2},$$
$$\forall \Psi = (\psi_1, \cdots, \psi_N)^T \in \left[H^1_\#(\Omega)\right]^N, \quad \|\Psi\|_{H^1_\#} := \|(1-\Delta)^{1/2}\Psi\|_{L^2_\#} = \left(\sum_{i=1}^N \|\psi_i\|^2_{H^1_\#}\right)^{1/2},$$
and more generally, for any operator A on $L^2_\#(\Omega)$ with domain $D(A)$,
$$\forall \Psi \in [D(A)]^N, \quad \|A\Psi\|_{L^2_\#} := \left(\sum_{i=1}^N \|A\psi_i\|^2_{L^2_\#}\right)^{1/2}.$$

Let $\mathcal{R}^* = \frac{2\pi}{L}\mathbb{Z}^3$ be the dual lattice of the periodic lattice $\mathcal{R} = L\mathbb{Z}^3$. For $\mathbf{k} \in \mathcal{R}^*$, we denote by $e_\mathbf{k}$ the plane-wave with wavevector $\mathbf{k}$, defined by
$$\begin{aligned} e_\mathbf{k} : \quad & \mathbb{R}^3 \to \mathbb{C} \\ & \mathbf{x} \mapsto |\Omega|^{-1/2} e^{i\mathbf{k}\cdot\mathbf{x}}, \end{aligned}$$
where $|\Omega| = L^3$. The family $(e_\mathbf{k})_{\mathbf{k}\in\mathcal{R}^*}$ forms an orthonormal basis of the complex Hilbert space
$$L^2_\#(\Omega, \mathbb{C}) := \left\{u \in L^2_{\mathrm{loc}}(\mathbb{R}^3, \mathbb{C}) \mid u \text{ is } \mathcal{R}\text{-periodic}\right\},$$
endowed with the scalar product
$$\forall u, v \in L^2_\#(\Omega, \mathbb{C}), \quad \langle u|v\rangle = \int_\Omega \overline{u(\mathbf{r})}\, v(\mathbf{r})\, d\mathbf{r},$$

where $\overline{u(\mathbf{r})}$ denotes the complex conjugate of $u(\mathbf{r})$, and for all $v \in L^2_\#(\Omega,\mathbb{C})$,

$$v(\mathbf{r}) = \sum_{\mathbf{k}\in\mathcal{R}^*} \widehat{v}_\mathbf{k}\, e_\mathbf{k}(\mathbf{r}) \qquad \text{with} \qquad \widehat{v}_\mathbf{k} = \langle e_\mathbf{k}|v\rangle = |\Omega|^{-1/2}\int_\Omega v(\mathbf{r})e^{-i\mathbf{k}\cdot\mathbf{r}}\,d\mathbf{r}.$$

Recall that the periodic Sobolev spaces $H^s_\#(\Omega)$ can be characterized in a simple way using Fourier series: for $s\in\mathbb{R}$, we have

$$H^s_\#(\Omega) := \left\{ v = \sum_{\mathbf{k}\in\mathcal{R}^*} \widehat{v}_\mathbf{k} e_\mathbf{k} \;\middle|\; \forall\,\mathbf{k},\quad \widehat{v}_{-\mathbf{k}} = \overline{\widehat{v}_\mathbf{k}},\quad \|v\|^2_{H^s_\#} := \sum_{\mathbf{k}\in\mathcal{R}^*} (1+|\mathbf{k}|^2)^s |\widehat{v}_\mathbf{k}|^2 < \infty \right\},$$

where the $H^s_\#$ inner product is defined by

$$\forall u,v\in H^s_\#(\Omega),\quad (u,v)_{H^s_\#} := \sum_{\mathbf{k}\in\mathcal{R}^*} (1+|\mathbf{k}|^2)^s \overline{u_\mathbf{k}} v_\mathbf{k}.$$

Let us now clarify the meaning of the terms $\mathrm{Tr}\,(-\Delta\gamma)$ and $\mathrm{Tr}\,(\mathcal{H}\,\gamma)$ appearing in (8.2.5)–(8.2.7). Let $\gamma\in\mathcal{L}(L^2_\#)$ be self-adjoint and positive. Since $|\nabla|$ (i.e.the multiplication operator by $|\mathbf{k}|$ in Fourier representation) is a bounded linear operator from $H^s_\#(\Omega)$ to $H^{s-1}_\#(\Omega)$, $|\nabla|\gamma|\nabla|$ defines a bounded linear operator from $H^1_\#(\Omega)$ to $H^{-1}_\#(\Omega)$. If in addition, $\mathrm{Ran}(|\nabla|\gamma|\nabla|)\subset L^2_\#(\Omega)$ and

$$\exists C\in\mathbb{R}_+ \quad \text{such that} \quad \forall u\in H^1_\#(\Omega),\quad \||\nabla|\gamma|\nabla|u\|_{L^2_\#} \le C\|u\|_{L^2_\#},$$

then $|\nabla|\gamma|\nabla|$ can be uniquely extended to a bounded, self-adjoint, positive operator on $L^2_\#(\Omega)$, also denoted by $|\nabla|\gamma|\nabla|$ for simplicity. In this case, $\mathrm{Tr}\,(|\nabla|\gamma|\nabla|)$ is well-defined in $\mathbb{R}_+\cup\{+\infty\}$. In view of the fact that $-\Delta = |\nabla|^2$, the notation

$$\mathrm{Tr}\,(-\Delta\gamma) := \mathrm{Tr}\,(|\nabla|\gamma|\nabla|)$$

is commonly used in the mathematical physics literature. Let us emphasize that $\mathrm{Tr}\,(-\Delta\gamma)<\infty$ only means that $\mathrm{Tr}\,(|\nabla|\gamma|\nabla|)<\infty$; in particular, it does *not* imply that the operator $-\Delta\gamma$ is in $\mathfrak{S}_1(L^2_\#)$.

It follows from the Hoffmann–Ostenhof inequality [138] that for all $\gamma\in\mathcal{K}$, $\rho_\gamma\ge 0$ and $\sqrt{\rho_\gamma}\in H^1_\#(\Omega)\hookrightarrow L^6_\#(\Omega)$. The real number $\mathrm{Tr}\,(\mathcal{H}\,\gamma)$ can therefore be defined for all $\gamma\in\mathcal{K}$ as

$$\mathrm{Tr}\,(\mathcal{H}\,\gamma) := \frac{1}{2}\mathrm{Tr}\,(-\Delta\gamma) + \int_\Omega \rho_\gamma V.$$

It is known in addition (see e.g. [46]) that, under Assumption 8.2.1, there exist $0<c\le C<\infty$ such that

$$c(1-\Delta) \le |\mathcal{H}-\epsilon_\mathrm{F}| \le C(1-\Delta), \tag{8.2.14}$$

where $|\mathcal{H}-\epsilon_\mathrm{F}| = -\gamma_0(\mathcal{H}-\epsilon_F)\gamma_0 + (1-\gamma_0)(\mathcal{H}-\epsilon_F)(1-\gamma_0)$ is defined by functional calculus for self-adjoint operators, and

$$\forall\gamma\in\mathcal{K},\quad \mathrm{Tr}\,(\mathcal{H}\,\gamma) - \mathrm{Tr}\,(\mathcal{H}\,\gamma_0) = \||\mathcal{H}-\epsilon_F|^{1/2}(\gamma-\gamma_0)\|^2_{\mathfrak{S}_2(L^2_\#)}. \tag{8.2.15}$$

We deduce from (8.2.14) and (8.2.15) that there exist two constants $0<c\le C<\infty$ such that

$$\forall\gamma\in\mathcal{K},\quad c\|(1-\Delta)^{1/2}(\gamma-\gamma_0)\|^2_{\mathfrak{S}_2(L^2_\#)} \le E(\gamma)-\mathcal{E}_0 \le C\|(1-\Delta)^{1/2}(\gamma-\gamma_0)\|^2_{\mathfrak{S}_2(L^2_\#)}. \tag{8.2.16}$$

This implies in particular that γ_0 is the unique minimizer of (8.2.8) and (8.2.9).

Note that for all $\gamma\in\Upsilon$, as $\gamma^2=\gamma$ and by the cyclicity of the trace, there also holds $\mathrm{Tr}\,(\mathcal{H}\,\gamma) = \mathrm{Tr}\,(\gamma\mathcal{H}\,) = \mathrm{Tr}\,(\gamma\mathcal{H}\,\gamma)$.

8.2.3 Discretization

In order to solve problem (8.2.3) numerically, we use a plane-wave discretization. For each $\mathbf{k} \in \mathcal{R}^*$, the kinetic energy of the plane-wave $e_{\mathbf{k}}$ is given by $\frac{1}{2}|\mathbf{k}|^2$, where $|\cdot|$ denotes the Euclidean norm. To construct a discretization space, we introduce some energy cut-off $E_{\rm c} > 0$ and consider all plane-waves whose kinetic energy is smaller than $E_{\rm c}$, i.e.$|\mathbf{k}| \leq \sqrt{2E_{\rm c}}$. For each cut-off energy $E_{\rm c}$, we set $N_{\rm c} = \sqrt{\frac{E_{\rm c}}{2}}\frac{L}{\pi}$ and

$$X_{N_{\rm c}} := \left\{ \sum_{\mathbf{k}\in\mathcal{R}^*, |\mathbf{k}|\leq \frac{2\pi}{L}N_{\rm c}} \widehat{v}_{\mathbf{k}}\, e_{\mathbf{k}} \;\middle|\; \widehat{v}_{-\mathbf{k}} = \overline{\widehat{v}_{\mathbf{k}}},\ \forall \mathbf{k} \right\} \subset \bigcap_{s\in\mathbb{R}} H^s_{\#}(\Omega).$$

For all $s \in \mathbb{R}$, for all $r \leq s$, and for each $v \in H^s_{\#}(\Omega)$, the best approximation of v in $X_{N_{\rm c}}$ in any $H^r_{\#}$-norm is

$$\Pi_{N_{\rm c}} v = \sum_{\mathbf{k}\in\mathcal{R}^*, |\mathbf{k}|\leq \frac{2\pi}{L}N_{\rm c}} \widehat{v}_{\mathbf{k}} e_{\mathbf{k}}.$$

We denote by $\Pi^{\perp}_{N_{\rm c}} = (1 - \Pi_{N_{\rm c}})$ the orthogonal projector on $X^{\perp}_{N_{\rm c}}$, the orthogonal of $X_{N_{\rm c}}$. The variational approximation to the ground-state energy in $X_{N_{\rm c}}$ is defined as

$$\mathcal{E}_{0,N_{\rm c}} = \inf\left\{ \mathcal{E}(\Psi_{N_{\rm c}})\ ,\ \Psi_{N_{\rm c}} \in \mathcal{M} \cap [X_{N_{\rm c}}]^N \right\}, \tag{8.2.17}$$

with $\mathcal{E}$ and $\mathcal{M}$ defined in (8.2.1) and (8.2.2). Let $\lambda_{1,N_{\rm c}} \leq \lambda_{2,N_{\rm c}} \leq \cdots \leq \lambda_{\dim(X_{N_{\rm c}}),N_{\rm c}}$ be the $\dim(X_{N_{\rm c}})$ eigenvalues (counting multiplicites) of the Hermitian linear operator $\mathcal{H}_{N_{\rm c},\rm proj} : X_{N_{\rm c}} \to X_{N_{\rm c}}$ defined as

$$\mathcal{H}_{N_{\rm c},\rm proj} = \Pi_{N_{\rm c}} \mathcal{H}\, \Pi_{N_{\rm c}} = -\frac{1}{2}\Pi_{N_{\rm c}}\Delta\Pi_{N_{\rm c}} + \Pi_{N_{\rm c}}\mathcal{V}\Pi_{N_{\rm c}}. \tag{8.2.18}$$

Let $(\phi_{1,N_{\rm c}}, \ldots, \phi_{N,N_{\rm c}})$ be an orthonormal family of eigenvectors of $\mathcal{H}_{N_{\rm c},\rm proj}$ associated with the eigenvalues $\lambda_{1,N_{\rm c}} \leq \cdots \leq \lambda_{N,N_{\rm c}}$:

$$\mathcal{H}_{N_{\rm c},\rm proj}\phi_{i,N_{\rm c}} = \lambda_{i,N_{\rm c}}\phi_{i,N_{\rm c}}, \qquad \int_{\Omega} \phi_{i,N_{\rm c}}\phi_{j,N_{\rm c}} = \delta_{ij}, \qquad \forall 1 \leq i,j \leq N,$$

and let $\Phi_{N_{\rm c}} := (\phi_{1,N_{\rm c}}, \ldots, \phi_{N,N_{\rm c}})^T$. Then $\Phi_{N_{\rm c}}$ is a minimizer of (8.2.17). Denoting by

$$\boxed{\gamma_{0,N_{\rm c}} = \sum_{i=1}^{N} |\phi_{i,N_{\rm c}}\rangle\langle\phi_{i,N_{\rm c}}|} \tag{8.2.19}$$

the associated density matrix, we have

$$\boxed{\mathcal{E}_{0,N_{\rm c}} = \mathrm{Tr}\left(\mathcal{H}\,\gamma_{0,N_{\rm c}}\right) = \sum_{j=1}^{N} \lambda_{j,N_{\rm c}}.} \tag{8.2.20}$$

8.2.4 *A priori* results on the density matrices

From now on, we make the following technical assumption:

Assumption 8.2.2. *$\mathcal{V}$ is a $\mathcal{R}-$periodic potential such that $\mathcal{V} \in L^{\infty}_{\#}(\Omega)$ and $\nabla\mathcal{V} \in L^3_{\#}(\Omega)$.*

The *a priori* error estimates established in [44] for the nonlinear Kohn–Sham model also hold true for the linear subproblem. In order to use these results in the present setting, it is convenient to reformulate them in terms of density matrices. As in [44], we introduce

$$\mathcal{M}^{\Phi^0} := \left\{ \Psi \in \mathcal{M} \;\middle|\; \|\Psi - \Phi^0\|_{L^2_\#} = \min_{U \in O(N)} \|U\Psi - \Phi^0\|_{L^2_\#} \right\},$$

where $\Phi^0 = (\phi_1^0, \dots, \phi_N^0)^T$, $(\phi_1^0, \dots, \phi_N^0)$ being a family of orthonormal eigenvectors of $\mathcal{H}$ associated with the eigenvalues $\lambda_1^0 \leq \cdots \leq \lambda_N^0$ fixed once and for all.

Proceeding as in [44], it can be shown that, for N_c large enough, (8.2.17) has a unique minimizer $\Phi^0_{N_\mathrm{c}} = (\phi^0_{1,N_\mathrm{c}}, \dots, \phi^0_{N,N_\mathrm{c}})^T$ belonging to $\mathcal{M}^{\Phi^0}$, that the set of minimizers of (8.2.17) is $O(N)\Phi^0_{N_\mathrm{c}}$, and that, consequently, all the minimizers of (8.2.17) share the same density matrix. In particular

$$\gamma_{0,N_\mathrm{c}} = \sum_{i=1}^{N} |\phi^0_{i,N_\mathrm{c}}\rangle\langle\phi^0_{i,N_\mathrm{c}}|.$$

We denote by

$$\Lambda^0_{N_\mathrm{c}} = (\lambda^0_{ij,N_\mathrm{c}})_{1\leq i,j\leq N} := (\langle\phi^0_{i,N_\mathrm{c}}|\mathcal{H}|\phi^0_{j,N_\mathrm{c}}\rangle)_{1\leq i,j\leq N} \in \mathbb{R}^{N\times N} \tag{8.2.21}$$

the Lagrange multiplier matrix of the orthonormality constraints. Note that the matrix $\Lambda^0_{N_\mathrm{c}}$ is not diagonal in general, but that we have

$$\mathcal{E}_{0,N_\mathrm{c}} = \mathrm{Tr}\,(\Lambda^0_{N_\mathrm{c}}).$$

The following lemma allows one to translate the *a priori* results of [44, Theorem 4.2] in terms of density matrices.

Lemma 8.2.3. *Under Assumption 8.2.1, there exist* $0 < c \leq C < \infty$ *and* $N_\mathrm{c}^0 \in \mathbb{N}$, *such that for all* $N_\mathrm{c} \geq N_\mathrm{c}^0$,

$$\|\Phi^0_{N_\mathrm{c}} - \Phi^0\|_{L^2_\#} \;\leq\; \|\gamma_{0,N_\mathrm{c}} - \gamma_0\|_{\mathfrak{S}_2(L^2_\#)} \;\leq\; \sqrt{2}\,\|\Phi^0_{N_\mathrm{c}} - \Phi^0\|_{L^2_\#}, \tag{8.2.22}$$

$$c\|(1-\Delta)^{1/2}(\Phi^0_{N_\mathrm{c}} - \Phi^0)\|_{L^2_\#} \;\leq\; \|(1-\Delta)^{1/2}(\gamma_{0,N_\mathrm{c}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)} \;\leq\; C\|(1-\Delta)^{1/2}(\Phi^0_{N_\mathrm{c}} - \Phi^0)\|_{L^2_\#}. \tag{8.2.23}$$

The proof is given in the Appendix.

We then immediately infer from [44, Theorem 4.2] that under Assumptions 8.2.1 and 8.2.2, there exist $C \in \mathbb{R}_+$ and $N_\mathrm{c}^0 \in \mathbb{N}$ such that for all $N_\mathrm{c} \geq N_\mathrm{c}^0$,

$$\|\gamma_0 - \gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq CN_\mathrm{c}^{-2}, \tag{8.2.24}$$

$$\|\gamma_0 - \gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq CN_\mathrm{c}^{-1}\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)}, \tag{8.2.25}$$

and

$$\|\Lambda^0 - \Lambda^0_{N_\mathrm{c}}\|_\mathrm{F} \leq C\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|^2_{\mathfrak{S}_2(L^2_\#)}, \tag{8.2.26}$$

where $\|\,.\,\|_\mathrm{F}$ denotes the Frobenius norm.

8.3 Post-processing of the plane-wave approximation

8.3.1 A key remark

Let us introduce the Hamiltonian on $L^2_{\#}(\Omega)$ with domain $H^2_{\#}(\Omega)$ defined by

$$\forall u \in H^2_{\#}(\Omega), \quad \mathcal{H}_{N_\mathrm{c}} u = -\frac{1}{2}\Delta u + \Pi_{N_\mathrm{c}} \mathcal{V} \Pi_{N_\mathrm{c}} u.$$

Since X_{N_c} and $X^\perp_{N_\mathrm{c}}$ are invariant subspaces of $\mathcal{H}_{N_\mathrm{c}}$, the Hamiltonian $\mathcal{H}_{N_\mathrm{c}}$ can be represented in term of $\mathcal{H}_{N_\mathrm{c},\mathrm{proj}}$ as follows:

$$\mathcal{H}_{N_\mathrm{c}} = \begin{array}{c} \left(\begin{array}{cc} \boxed{\mathcal{H}_{N_\mathrm{c},\mathrm{proj}}} & 0 \\ 0 & \boxed{-\frac{1}{2}\Delta} \end{array} \right) \begin{array}{l} \Big\} X_{N_\mathrm{c}} \\ \Big\} X^\perp_{N_\mathrm{c}} \end{array} \\ \underbrace{\qquad\qquad}_{X_{N_\mathrm{c}}} \quad \underbrace{\qquad\qquad}_{X^\perp_{N_\mathrm{c}}} \end{array} \tag{8.3.1}$$

The eigenvalues of the Laplace operator, which is diagonal in plane-wave bases, are explicitly known and its smallest eigenvalue on the invariant subspace $X^\perp_{N_\mathrm{c}}$ is $\frac{1}{2}\left(\frac{LN_\mathrm{c}}{\pi}\right)^2$. Therefore, as soon as

$$\lambda_{N,N_\mathrm{c}} < \frac{1}{2}\left(\frac{LN_\mathrm{c}}{\pi}\right)^2, \tag{8.3.2}$$

where we recall that $\lambda_{1,N_\mathrm{c}} \leq \cdots \leq \lambda_{N,N_\mathrm{c}}$ are the lowest N eigenvalues (counting multiplicities) of the operator $\mathcal{H}_{N_\mathrm{c},\mathrm{proj}}$ defined in (8.2.18), we have

$$\forall j = 1, \ldots, N, \quad \mathcal{H}_{N_\mathrm{c}} \phi_{j,N_\mathrm{c}} = \lambda_{j,N_\mathrm{c}}\, \phi_{j,N_\mathrm{c}}, \qquad \text{and} \qquad \forall i,j = 1,\ldots,N, \quad \int_\Omega \phi_{i,N_\mathrm{c}}\, \phi_{j,N_\mathrm{c}} = \delta_{ij}, \tag{8.3.3}$$

and $\lambda_{1,N_\mathrm{c}} \leq \cdots \leq \lambda_{N,N_\mathrm{c}}$ are also the lowest N eigenvalues (counting multiplicities) of the operator $\mathcal{H}_{N_\mathrm{c}}$. A key observation is that that the lowest energy eigenmodes of $\mathcal{H}$ satisfy

$$\forall j = 1, \ldots, N, \quad (\mathcal{H}_{N_\mathrm{c}} + \mathcal{V}^\perp_{N_\mathrm{c}})\, \phi^0_j = \lambda^0_j\, \phi^0_j, \qquad \text{and} \qquad \forall i,j = 1,\ldots,N, \quad \int_\Omega \phi^0_i\, \phi^0_j = \delta_{ij}, \tag{8.3.4}$$

where

$$\mathcal{V}^\perp_{N_\mathrm{c}} = \mathcal{V} - \Pi_{N_\mathrm{c}} \mathcal{V} \Pi_{N_\mathrm{c}}. \tag{8.3.5}$$

We can therefore apply the Rayleigh–Schrödinger perturbation method [146] using $(\phi_{j,N_\mathrm{c}}, \lambda_{j,N_\mathrm{c}})_{j=1,\ldots,N}$ as the reference solution and $(\phi^0_j, \lambda^0_j)_{j=1,\ldots,N}$ as the perturbed solution, in order to build improved approximations of the orbitals and eigenvalues respectively denoted by $(\widetilde{\phi_{j,N_\mathrm{c}}})_{j=1,\ldots,N}$ and $(\widetilde{\lambda_{j,N_\mathrm{c}}})_{j=1,\ldots,N}$, as well as an improved density matrix $\widetilde{\gamma_{N_\mathrm{c}}}$ and improved energy $\widetilde{\mathcal{E}_{0,N_\mathrm{c}}}$.

8.3.2 Corrections computation

More precisely, we apply first-order perturbation to the analytic family of operators $\mathcal{H}(\beta) = \mathcal{H}_{N_\mathrm{c}} + \beta \mathcal{V}^\perp_{N_\mathrm{c}}$, where $\beta \in \mathbb{R}$ is a parameter, which amounts to considering $\mathcal{H}(0) = \mathcal{H}_{N_\mathrm{c}}$ and

(8.3.3) as the unperturbed eigenvalue problem, and $\mathcal{H}(1) = \mathcal{H}$ and (8.3.4) as the perturbed eigenvalue problem. Assuming that the eigenvalues are not degenerate, we obtain at first order for the eigenfunctions, and at second order for the eigenvalues,

$$\forall\, j = 1, \dots, N, \qquad \phi_j^0 \simeq \phi_{j,N_c}^0 + \phi_{j,N_c}^{(1)}, \qquad \lambda_j^0 \simeq \lambda_{j,N_c} + \lambda_{j,N_c}^{(2)},$$

where

$$\phi_{j,N_c}^{(1)} = -\left(-\frac{1}{2}\Delta - \lambda_{j,N_c}\right)^{-1} r_j \in X_{N_c}^{\perp}, \tag{8.3.6}$$

with r_j being the residual

$$r_j = \left(-\frac{1}{2}\Delta + \mathcal{V} - \lambda_{j,N_c}\right)\phi_{j,N_c} = \left(\mathcal{H}_{N_c} + \mathcal{V}_{N_c}^{\perp} - \lambda_{j,N_c}\right)\phi_{j,N_c} = \mathcal{V}_{N_c}^{\perp}\phi_{j,N_c}, \tag{8.3.7}$$

and

$$\lambda_{j,N_c}^{(2)} = \langle \phi_{j,N_c}^{(1)} | r_j \rangle = -\langle r_j | \left(-\frac{1}{2}\Delta - \lambda_{j,N_c}\right)^{-1} | r_j \rangle. \tag{8.3.8}$$

We observe that the corrections on the eigenfunctions given in (8.3.6) are well-defined even if λ_{j,N_c} is degenerate. We therefore define the perturbed eigenvectors, density matrix, and energy for the general case as follows.

Definition 8.3.1 (Perturbed eigenvectors, eigenvalues, density matrix, and energy)**.** *For all $N_c \geq N_c^0$ and all $j = 1, \dots, N$, the perturbed eigenvectors are defined as*

$$\widetilde{\phi_{j,N_c}} = \phi_{j,N_c} + \phi_{j,N_c}^{(1)},$$

the perturbed eigenvalues as

$$\widetilde{\lambda_{j,N_c}} = \lambda_{j,N_c} + \lambda_{j,N_c}^{(2)},$$

the perturbed density matrix as

$$\boxed{\widetilde{\gamma_{N_c}} = \gamma_{0,N_c} + \gamma_{N_c}^{(1)},}$$

with

$$\gamma_{N_c}^{(1)} = \sum_{j=1}^{N} |\phi_{j,N_c}^{(1)}\rangle\langle\phi_{j,N_c}| + \sum_{j=1}^{N} |\phi_{j,N_c}\rangle\langle\phi_{j,N_c}^{(1)}|, \tag{8.3.9}$$

and the perturbed energy as

$$\boxed{\widetilde{\mathcal{E}_{0,N_c}} = \sum_{j=1}^{N} \widetilde{\lambda_{j,N_c}} = \mathrm{Tr}\,(\gamma_{0,N_c}\mathcal{H}\,\widetilde{\gamma_{N_c}}).} \tag{8.3.10}$$

Remark 8.3.2. *Note that even if we call $\widetilde{\gamma_{N_c}}$ a density matrix, $\widetilde{\gamma_{N_c}} \notin \mathcal{K}$ in general. Indeed, $\widetilde{\gamma_{N_c}} = \widetilde{\gamma_{N_c}}^*$ and $\mathrm{Tr}\,(\widetilde{\gamma_{N_c}}) = N$, but we do not have in general $0 \leq \widetilde{\gamma_{N_c}} \leq 1$. Hence, the perturbed energy, which is defined as the sum of the perturbed eigenvalues, is not equal to the energy of the perturbed density matrix, i.e. $\widetilde{\mathcal{E}_{0,N_c}} \neq \mathrm{Tr}\,(\mathcal{H}\,\widetilde{\gamma_{N_c}})$.*

Remark 8.3.3. *Note that the quantities $\phi_{j,N_c}^{(1)}$ are easily computable. Indeed, the operator $\left(-\frac{1}{2}\Delta - \lambda_{j,N_c}\right)$ is diagonal in plane-wave bases, hence very easy to invert. Moreover, only two FFT's are needed to compute the residual or $\mathcal{V}_{N_c}^{\perp}\phi_{j,N_c}$ on a larger grid, via a product in the physical space.*

8.4 Convergence improvement on the density matrix and the energy

8.4.1 Main results

The main results of this article are collected in the following theorem.

Theorem 8.4.1. *Under Assumptions 8.2.1–8.2.2, there exist $C \in \mathbb{R}_+$ and $N_{\mathrm{c}}^0 \in \mathbb{N}$, such that for all $N_{\mathrm{c}} \geq N_{\mathrm{c}}^0$,*

$$\boxed{\|(1-\Delta)^{1/2}(\widetilde{\gamma_{N_{\mathrm{c}}}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)} \leq C N_{\mathrm{c}}^{-2} \|(1-\Delta)^{1/2}(\gamma_{0,N_{\mathrm{c}}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)},} \tag{8.4.1}$$

and

$$\boxed{\left|\widetilde{\mathcal{E}_{0,N_{\mathrm{c}}}} - \mathcal{E}_0\right| \leq C N_{\mathrm{c}}^{-2} \left|\mathcal{E}_{0,N_{\mathrm{c}}} - \mathcal{E}_0\right|.} \tag{8.4.2}$$

8.4.2 Proofs

In order to prove Theorem 8.4.1, we first provide in Section 8.4.2 a decomposition of γ_0 based on spectral projection in Lemma 8.4.3, relying on Lemma 8.4.2 for a rigorous justification of the contour integral. In Section 8.4.2, we decompose the difference $\gamma_0 - \widetilde{\gamma_{N_{\mathrm{c}}}}$ into three parts in Lemma 8.4.4, and we then estimate each of these terms in three of the following Lemmas 8.4.5, 8.4.7, and 8.4.8, relying on an intermediary estimate presented in Lemma 8.4.6, in order to prove estimate (8.4.1). Finally, in Section 8.4.2, we provide a proof for estimate (8.4.2).

Exact density matrix in terms of approximate density matrix

Lemma 8.4.2. *Let Γ be the circle in the complex plane symmetric with respect to the real axis and containing the real numbers $\lambda_1^0 - 1$ and ϵ_{F}. There exists $N_{\mathrm{c}}^0 \in \mathbb{N}$ such that for all $N_{\mathrm{c}} \geq N_{\mathrm{c}}^0$, Γ encloses the lowest N eigenvalues of both the operators $\mathcal{H}$ and $\mathcal{H}_{N_{\mathrm{c}}}$, and none of the higher ones.*

Proof. As $\mathcal{H}$ and $\mathcal{H}_{N_{\mathrm{c}}}$ are self-adjoint operators, their eigenvalues noted respectively $(\lambda_i^0)_{i\in\mathbb{N}^*}$ and $(\lambda_{i,N_{\mathrm{c}}})_{i\in\mathbb{N}^*}$ (with increasing values and counting multiplicities) are real. From the gap assumption 8.2.1, and the definition of the Fermi level, we have

$$\forall i = 1, \ldots, N, \quad \lambda_i^0 < \epsilon_F, \quad \text{and} \quad \forall i > N, \quad \lambda_i^0 > \epsilon_F. \tag{8.4.3}$$

The plane-wave discretization being variational, there holds

$$\forall i = 1, \ldots, \dim(X_{N_{\mathrm{c}}}), \quad \lambda_i^0 \leq \lambda_{i,N_{\mathrm{c}}}.$$

Moreover, classical convergence results (see e.g. [69, Chapter 5]) guarantee that

$$\max_{i=1,\ldots,N} |\lambda_{i,N_{\mathrm{c}}} - \lambda_i^0| \underset{N_{\mathrm{c}}\to+\infty}{\longrightarrow} 0.$$

Therefore, there exists $N_{\mathrm{c}}^0 \in \mathbb{N}$ such that for all $N_{\mathrm{c}} \geq N_{\mathrm{c}}^0$, $\lambda_{N,N_{\mathrm{c}}} \leq \lambda_{N,N_{\mathrm{c}}^0} < \epsilon_F$, and the eigenvalues of the Laplace operator on $X_{N_{\mathrm{c}}}^\perp$ are larger than $\lambda_{N+1}^0 > \epsilon_F$, so that

$$\forall N_{\mathrm{c}} \geq N_{\mathrm{c}}^0, \qquad \forall i = 1, \ldots, N, \quad \lambda_{i,N_{\mathrm{c}}} \leq \lambda_{N,N_{\mathrm{c}}^0} < \epsilon_F, \quad \text{and} \quad \forall i > N, \quad \lambda_{i,N_{\mathrm{c}}} \geq \lambda_{N+1}^0 > \epsilon_F. \tag{8.4.4}$$

Combining (8.4.3) and (8.4.4) concludes the proof of the lemma. □

Using the Cauchy residue theorem and functional calculus for self-adjoint operators, the ground-state density matrix of $\mathcal{H}$ can be written as

$$\gamma_0 = \frac{1}{2\pi i} \oint_\Gamma (z - \mathcal{H})^{-1} \, dz = \frac{1}{2\pi i} \oint_\Gamma \left(z - \mathcal{H}_{N_c} - \mathcal{V}_{N_c}^\perp \right)^{-1} dz. \tag{8.4.5}$$

Since $\mathcal{V} \in L^\infty_\#(\Omega)$, $\mathcal{V}_{N_c}^\perp$ is $\mathcal{H}_{N_c}$-bounded, and we can perform a Dyson expansion of (8.4.5) at second order. We obtain

$$\begin{aligned} \gamma_0 = \frac{1}{2\pi i} \oint_\Gamma (z - \mathcal{H}_{N_c})^{-1} dz + \frac{1}{2\pi i} \oint_\Gamma (z - \mathcal{H}_{N_c})^{-1} \mathcal{V}_{N_c}^\perp (z - \mathcal{H}_{N_c})^{-1} dz \\ + \frac{1}{2\pi i} \oint_\Gamma (z - \mathcal{H})^{-1} \mathcal{V}_{N_c}^\perp (z - \mathcal{H}_{N_c})^{-1} \mathcal{V}_{N_c}^\perp (z - \mathcal{H}_{N_c})^{-1} dz, \end{aligned} \tag{8.4.6}$$

where each term of the right-hand side is well-defined.

Lemma 8.4.3 (Second order expansion of γ_0)**.** *There holds*

$$\gamma_0 = \gamma_{0,N_c} + \gamma_{N_c}^{(1)} + \widetilde{Q_{N_c}}, \tag{8.4.7}$$

where $\gamma_{N_c}^{(1)}$ is the finite-rank operator defined in (8.3.9) *and where*

$$\widetilde{Q_{N_c}} := \frac{1}{2i\pi} \oint_\Gamma (z - \mathcal{H})^{-1} \mathcal{V}_{N_c}^\perp (z - \mathcal{H}_{N_c})^{-1} \mathcal{V}_{N_c}^\perp (z - \mathcal{H}_{N_c})^{-1} dz. \tag{8.4.8}$$

Proof. The operator $\mathcal{H}_{N_c}$ being self-adjoint with compact resolvent, it can be diagonalized in an orthonormal basis. Hence, there exists a sequence $(\psi_k, \varepsilon_k)_{k\geq 1}$ with $(\psi_k)_{k\geq 1}$ an orthonormal basis of $L^2_\#(\Omega)$ consisting of functions of $H^2_\#(\Omega)$ and $(\varepsilon_k)_{k\geq 1}$ a non-decreasing sequence of real numbers such that

$$\forall k \geq 1, \quad \mathcal{H}_{N_c} \psi_k = \varepsilon_k \psi_k.$$

Without loss of generality, we can choose a basis such that, in addition, for $k = 1, \ldots, N$, $\psi_k = \phi_{k,N_c}$ and $\varepsilon_k = \lambda_{k,N_c}$. The operator $\mathcal{H}_{N_c}$ can then be written as

$$\mathcal{H}_{N_c} = \sum_{k\geq 1} \varepsilon_k |\psi_k\rangle\langle\psi_k|.$$

Let us show that the expansions (8.4.6) and (8.4.7) are identical. First, we have

$$\gamma_{0,N_c} = \frac{1}{2\pi i} \oint_\Gamma (z - \mathcal{H}_{N_c})^{-1} \, dz.$$

Let us now prove that the second term in the right hand side, that is

$$\gamma^{(1)} := \frac{1}{2\pi i} \oint_\Gamma (z - \mathcal{H}_{N_c})^{-1} \, \mathcal{V}_{N_c}^\perp \, (z - \mathcal{H}_{N_c})^{-1} \, dz, \tag{8.4.9}$$

is in fact equal to the operator $\gamma_{N_c}^{(1)}$ defined in (8.3.9). We have

$$\forall k, l \in \mathbb{N}^*, \quad \langle \psi_k | \gamma_{N_c}^{(1)} | \psi_l \rangle = \sum_{j=1}^N \langle \psi_k | \phi_{j,N_c}^{(1)} \rangle \langle \phi_{j,N_c} | \psi_l \rangle + \sum_{j=1}^N \langle \psi_k | \phi_{j,N_c} \rangle \langle \phi_{j,N_c}^{(1)} | \psi_l \rangle.$$

As for all $j = 1, \dots, N$, $\phi_{j,N_c} \in X_{N_c}$ and $\phi^{(1)}_{j,N_c} \in X^{\perp}_{N_c}$, we have $\langle \psi_k | \gamma^{(1)}_{N_c} | \psi_l \rangle = 0$ for all $k, l \in \mathbb{N}^*$ such that either $k, l > N$, or $k, l \leq N$. Moreover, for all $k \leq N$ and $l > N$, we have

$$\begin{aligned}\langle \psi_k | \gamma^{(1)}_{N_c} | \psi_l \rangle = \langle \phi_{k,N_c} | \gamma^{(1)}_{N_c} | \psi_l \rangle = \langle \phi^{(1)}_{k,N_c} | \psi_l \rangle = -\langle \phi_{k,N_c} | \mathcal{V}^{\perp}_{N_c} (\mathcal{H}_{N_c} - \lambda_{k,N_c})^{-1} | \psi_l \rangle \\ = \frac{1}{\lambda_{k,N_c} - \varepsilon_l} \langle \phi_{k,N_c} | \mathcal{V}^{\perp}_{N_c} | \psi_l \rangle = \frac{1}{\varepsilon_k - \varepsilon_l} \langle \psi_k | \mathcal{V}^{\perp}_{N_c} | \psi_l \rangle,\end{aligned}$$

and likewise

$$\langle \psi_l | \gamma^{(1)}_{N_c} | \psi_k \rangle = \frac{1}{\varepsilon_l - \varepsilon_k} \langle \psi_l | \mathcal{V}^{\perp}_{N_c} | \psi_k \rangle.$$

Thus, for all $k, l \in \mathbb{N}^*$, we have

$$\langle \psi_k | \gamma^{(1)}_{N_c} | \psi_l \rangle = \frac{1}{\varepsilon_k - \varepsilon_l} \left(\mathbb{1}_{k \leq N} \mathbb{1}_{l > N} - \mathbb{1}_{k > N} \mathbb{1}_{l \leq N} \right) \langle \psi_k | \mathcal{V}^{\perp}_{N_c} | \psi_l \rangle.$$

On the other hand, for all $k, l \in \mathbb{N}^*$, using the Cauchy residue theorem,

$$\begin{aligned}\langle \psi_k | \gamma^{(1)} | \psi_l \rangle &= \frac{1}{2\pi i} \oint_{\Gamma} \langle \psi_k | (z - \mathcal{H}_{N_c})^{-1} \mathcal{V}^{\perp}_{N_c} (z - \mathcal{H}_{N_c})^{-1} | \psi_l \rangle dz \\ &= \frac{1}{2\pi i} \left(\oint_{\Gamma} (z - \varepsilon_k)^{-1} (z - \varepsilon_l)^{-1} dz \right) \langle \psi_k | \mathcal{V}^{\perp}_{N_c} | \psi_l \rangle \\ &= \frac{1}{\varepsilon_k - \varepsilon_l} \left(\mathbb{1}_{k \leq N} \mathbb{1}_{l > N} - \mathbb{1}_{k > N} \mathbb{1}_{l \leq N} \right) \langle \psi_k | \mathcal{V}^{\perp}_{N_c} | \psi_l \rangle.\end{aligned}$$

Therefore, for all $k, l \in \mathbb{N}^*$, $\langle \psi_k | \gamma^{(1)} | \psi_l \rangle = \langle \psi_k | \gamma^{(1)}_{N_c} | \psi_l \rangle$. Finally $\gamma^{(1)} = \gamma^{(1)}_{N_c}$, and the definition of $\widetilde{Q_{N_c}}$ in (8.4.8) allows one to conclude the proof of the lemma. □

Proof of estimate (8.4.1)

Lemma 8.4.4. *There holds*

$$\gamma_0 - \widetilde{\gamma_{N_c}} = (\gamma_0 - \gamma_{0,N_c})^2 + \widetilde{Q_{N_c}} \gamma_{0,N_c} + \gamma_{0,N_c} \widetilde{Q_{N_c}}, \tag{8.4.10}$$

with $\widetilde{Q_{N_c}}$ *defined in* (8.4.8).

Proof. Let us first remark, from (8.4.7), and the property $\gamma^2_{0,N_c} = \gamma_{0,N_c}$, that

$$\begin{aligned}\gamma_{0,N_c} \gamma_0 &= \gamma_{0,N_c} + \gamma_{0,N_c} \gamma^{(1)}_{N_c} + \gamma_{0,N_c} \widetilde{Q_{N_c}}, \\ \gamma_0 \gamma_{0,N_c} &= \gamma_{0,N_c} + \gamma^{(1)}_{N_c} \gamma_{0,N_c} + \widetilde{Q_{N_c}} \gamma_{0,N_c}.\end{aligned}$$

Moreover, as for all $i, j = 1, 2, \dots, N$, $\phi_{i,N_c} \in X_{N_c}$ and $\phi^{(1)}_{j,N_c} \in X^{\perp}_{N_c}$, ϕ_{i,N_c} is orthogonal to

$\phi^{(1)}_{j,N_\mathrm{c}}$, and therefore

$$\begin{aligned}\gamma_{0,N_\mathrm{c}}\gamma^{(1)}_{N_\mathrm{c}} + \gamma^{(1)}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}} =& \gamma_{0,N_\mathrm{c}}\left(\sum_{i=1}^{N} |\phi^{(1)}_{i,N_\mathrm{c}}\rangle\langle\phi_{i,N_\mathrm{c}}| + |\phi_{i,N_\mathrm{c}}\rangle\langle\phi^{(1)}_{i,N_\mathrm{c}}|\right) \\ &+ \left(\sum_{i=1}^{N} |\phi^{(1)}_{i,N_\mathrm{c}}\rangle\langle\phi_{i,N_\mathrm{c}}| + |\phi_{i,N_\mathrm{c}}\rangle\langle\phi^{(1)}_{i,N_\mathrm{c}}|\right)\gamma_{0,N_\mathrm{c}} \\ =& \gamma_{0,N_\mathrm{c}}\sum_{i=1}^{N} |\phi_{i,N_\mathrm{c}}\rangle\langle\phi^{(1)}_{i,N_\mathrm{c}}| + \sum_{i=1}^{N} |\phi^{(1)}_{i,N_\mathrm{c}}\rangle\langle\phi_{i,N_\mathrm{c}}|\gamma_{0,N_\mathrm{c}} \\ =& \sum_{i=1}^{N} |\phi_{i,N_\mathrm{c}}\rangle\langle\phi^{(1)}_{i,N_\mathrm{c}}| + \sum_{i=1}^{N} |\phi^{(1)}_{i,N_\mathrm{c}}\rangle\langle\phi_{i,N_\mathrm{c}}| = \gamma^{(1)}_{N_\mathrm{c}}.\end{aligned}$$

Hence

$$\gamma_{0,N_\mathrm{c}}\gamma_0 + \gamma_0\gamma_{0,N_\mathrm{c}} = 2\gamma_{0,N_\mathrm{c}} + \gamma^{(1)}_{N_\mathrm{c}} + \widetilde{Q_{N_\mathrm{c}}}\gamma_{0,N_\mathrm{c}} + \gamma_{0,N_\mathrm{c}}\widetilde{Q_{N_\mathrm{c}}} = \gamma_{0,N_\mathrm{c}} + \widetilde{\gamma_{N_\mathrm{c}}} + \widetilde{Q_{N_\mathrm{c}}}\gamma_{0,N_\mathrm{c}} + \gamma_{0,N_\mathrm{c}}\widetilde{Q_{N_\mathrm{c}}},$$

so that

$$(\gamma_0 - \gamma_{0,N_\mathrm{c}})^2 = \gamma_0 - (\gamma_{0,N_\mathrm{c}}\gamma_0 + \gamma_0\gamma_{0,N_\mathrm{c}}) + \gamma_{0,N_\mathrm{c}} = \gamma_0 - \widetilde{\gamma_{N_\mathrm{c}}} - \widetilde{Q_{N_\mathrm{c}}}\gamma_{0,N_\mathrm{c}} - \gamma_{0,N_\mathrm{c}}\widetilde{Q_{N_\mathrm{c}}},$$

from which we deduce (8.4.10). □

Lemma 8.4.5. *There exist $C \in \mathbb{R}_+$ and $N_\mathrm{c}^0 \in \mathbb{N}$, such that for all $N_\mathrm{c} \geq N_\mathrm{c}^0$,*

$$\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})^2\|_{\mathfrak{S}_2(L^2_\#)} \leq CN_\mathrm{c}^{-2}\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)}.$$

Proof. By cyclicity of the trace, noting that $(\gamma_0 - \gamma_{0,N_\mathrm{c}})$ is of finite rank, and using (8.2.11) and (8.2.10),

$$\begin{aligned}\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})^2\|^2_{\mathfrak{S}_2(L^2_\#)} &= \mathrm{Tr}\,\left((\gamma_0 - \gamma_{0,N_\mathrm{c}})^2(1-\Delta)(\gamma_0 - \gamma_{0,N_\mathrm{c}})^2\right) \\ &= \mathrm{Tr}\,\left((\gamma_0 - \gamma_{0,N_\mathrm{c}})^2(\gamma_0 - \gamma_{0,N_\mathrm{c}})(1-\Delta)(\gamma_0 - \gamma_{0,N_\mathrm{c}})\right) \\ &\leq \|(\gamma_0 - \gamma_{0,N_\mathrm{c}})^2(\gamma_0 - \gamma_{0,N_\mathrm{c}})(1-\Delta)(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_1(L^2_\#)} \\ &\leq \|\gamma_0 - \gamma_{0,N_\mathrm{c}}\|^2\|(\gamma_0 - \gamma_{0,N_\mathrm{c}})(1-\Delta)(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_1(L^2_\#)} \\ &\leq \|\gamma_0 - \gamma_{0,N_\mathrm{c}}\|^2_{\mathfrak{S}_2(L^2_\#)}\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|^2_{\mathfrak{S}_2(L^2_\#)}.\end{aligned}$$

Using the *a priori* estimate of $\|\gamma_0 - \gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}$ given in (8.2.24) finishes the proof. □

We now provide an estimate for $\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}$ which will be useful in the proof of Lemma 8.4.7 and estimate (8.4.2).

Lemma 8.4.6. *There exist $C \in \mathbb{R}_+$ and $N_\mathrm{c}^0 \in \mathbb{N}$, such that for $N_\mathrm{c} \geq N_\mathrm{c}^0$,*

$$\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq C\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)}. \tag{8.4.11}$$

Proof. Decomposing $\mathcal{V}^\perp_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}$ as

$$\mathcal{V}^\perp_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}} \;=\; \mathcal{H}\,(\gamma_{0,N_\mathrm{c}} - \gamma_0) + \mathcal{H}\,\gamma_0 - \mathcal{H}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}},$$

we get

$$\|(1-\Delta)^{-1/2}\mathcal{V}_{N_\mathrm{c}}^{\perp}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq \|(1-\Delta)^{-1/2}\mathcal{H}\,(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)} + \|(1-\Delta)^{-1/2}(\mathcal{H}\,\gamma_0-\mathcal{H}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)}. \quad (8.4.12)$$

Since $(1-\Delta)^{-1/2}\mathcal{H}\,(1-\Delta)^{-1/2}$ is a bounded operator (see e.g. [46, Lemma 1] for a proof of this classical result), there exists $C\in\mathbb{R}_+$ such that for all $N_\mathrm{c}\in\mathbb{N}$,

$$\|(1-\Delta)^{-1/2}\mathcal{H}\,(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)} \;\leq\; C\|(1-\Delta)^{1/2}(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}. \quad (8.4.13)$$

In order to bound the second term of the right-hand side of (8.4.12), we first rewrite the operator $\mathcal{H}\,\gamma_0-\mathcal{H}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}$ as follows, denoting by $\lambda^0_{ij}=\lambda^0_i\delta_{ij}$, and using $\lambda^0_{ij,N_\mathrm{c}}$ defined in (8.2.21):

$$\begin{aligned}
\mathcal{H}\,\gamma_0-\mathcal{H}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}} &= \sum_{i=1}^N \lambda^0_i|\phi^0_i\rangle\langle\phi^0_i| - \sum_{i,j=1}^N \lambda^0_{ij,N_\mathrm{c}}|\phi^0_{i,N_\mathrm{c}}\rangle\langle\phi^0_{j,N_\mathrm{c}}| \\
&= \sum_{i=1}^N \lambda^0_i\left(|\phi^0_i\rangle\langle\phi^0_i| - |\phi^0_{i,N_\mathrm{c}}\rangle\langle\phi^0_{i,N_\mathrm{c}}|\right) + \sum_{i=1}^N \lambda^0_i|\phi^0_{i,N_\mathrm{c}}\rangle\langle\phi^0_{i,N_\mathrm{c}}| \\
&\quad - \sum_{i,j=1}^N \lambda^0_{ij,N_\mathrm{c}}|\phi^0_{i,N_\mathrm{c}}\rangle\langle\phi^0_{j,N_\mathrm{c}}| \\
&= \sum_{i=1}^N \lambda^0_i\left(|\phi^0_i\rangle\langle\phi^0_i| - |\phi^0_{i,N_\mathrm{c}}\rangle\langle\phi^0_{i,N_\mathrm{c}}|\right) + \sum_{i,j=1}^N (\lambda^0_{ij}-\lambda^0_{ij,N_\mathrm{c}})|\phi^0_{i,N_\mathrm{c}}\rangle\langle\phi^0_{j,N_\mathrm{c}}|.
\end{aligned}$$

Using the triangle and the Cauchy–Schwarz inequality, we get

$$\begin{aligned}
\|\mathcal{H}\,\gamma_0-\mathcal{H}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} &\leq \left\|\sum_{i=1}^N \lambda^0_i\left(|\phi^0_i\rangle\langle\phi^0_i| - |\phi^0_{i,N_\mathrm{c}}\rangle\langle\phi^0_{i,N_\mathrm{c}}|\right)\right\|_{\mathfrak{S}_2(L^2_\#)} \\
&\quad + \left\|\sum_{i,j=1}^N (\lambda^0_{ij}-\lambda^0_{ij,N_\mathrm{c}})|\phi^0_{i,N_\mathrm{c}}\rangle\langle\phi^0_{j,N_\mathrm{c}}|\right\|_{\mathfrak{S}_2(L^2_\#)} \\
&\leq \sum_{i=1}^N |\lambda^0_i|\,\left\|\left(|\phi^0_i\rangle\langle\phi^0_i| - |\phi^0_{i,N_\mathrm{c}}\rangle\langle\phi^0_{i,N_\mathrm{c}}|\right)\right\|_{\mathfrak{S}_2(L^2_\#)} \\
&\quad + \sum_{i,j=1}^N |\lambda^0_{ij}-\lambda^0_{ij,N_\mathrm{c}}|\,\left\||\phi^0_{i,N_\mathrm{c}}\rangle\langle\phi^0_{j,N_\mathrm{c}}|\right\|_{\mathfrak{S}_2(L^2_\#)} \\
&\leq \left(\sum_{i=1}^N |\lambda^0_i|^2\right)^{1/2}\left(\sum_{i=1}^N \left\||\phi^0_i\rangle\langle\phi^0_i| - |\phi^0_{i,N_\mathrm{c}}\rangle\langle\phi^0_{i,N_\mathrm{c}}|\right\|^2_{\mathfrak{S}_2(L^2_\#)}\right)^{1/2} \\
&\quad + \|\Lambda^0-\Lambda^0_{N_\mathrm{c}}\|_F\left(\sum_{i,j=1}^N \left\||\phi^0_{i,N_\mathrm{c}}\rangle\langle\phi^0_{j,N_\mathrm{c}}|\right\|^2_{\mathfrak{S}_2(L^2_\#)}\right)^{1/2} \\
&\leq 2\left(\sum_{i=1}^N |\lambda^0_i|^2\right)^{1/2} \|\Phi^0-\Phi^0_{N_\mathrm{c}}\|_{L^2_\#} + N\|\Lambda^0-\Lambda^0_{N_\mathrm{c}}\|_F.
\end{aligned}$$

Using (8.2.22), (8.2.25), and (8.2.26), we obtain that there exist $C \in \mathbb{R}_+$ and $N_{\rm c}^0 \in \mathbb{N}$ such that for all $N_{\rm c} \geq N_{\rm c}^0$,

$$\begin{aligned}\|\mathcal{H}\,\gamma_0 - \mathcal{H}_{N_{\rm c}}\gamma_{0,N_{\rm c}}\|_{\mathfrak{S}_2(L^2_\#)} &\leq CN_{\rm c}^{-1}\|(1-\Delta)^{1/2}(\gamma_0-\gamma_{0,N_{\rm c}})\|_{\mathfrak{S}_2(L^2_\#)} \\ &\quad + C\|(1-\Delta)^{1/2}(\gamma_0-\gamma_{0,N_{\rm c}})\|^2_{\mathfrak{S}_2(L^2_\#)}.\end{aligned}$$

Since $\|(1-\Delta)^{-1/2}\| \leq 1$, this shows in particular that

$$\|(1-\Delta)^{-1/2}(\mathcal{H}\,\gamma_0 - \mathcal{H}_{N_{\rm c}}\gamma_{0,N_{\rm c}})\|_{\mathfrak{S}_2(L^2_\#)} \leq C\|(1-\Delta)^{1/2}(\gamma_0-\gamma_{0,N_{\rm c}})\|_{\mathfrak{S}_2(L^2_\#)}. \tag{8.4.14}$$

Inserting (8.4.13) and (8.4.14) in (8.4.12) concludes the proof of the lemma. □

Lemma 8.4.7. *There exist $C \in \mathbb{R}_+$ and $N_{\rm c}^0 \in \mathbb{N}$ such that for all $N_{\rm c} \geq N_{\rm c}^0$,*

$$\|(1-\Delta)^{1/2}\widetilde{Q_{N_{\rm c}}}\gamma_{0,N_{\rm c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq CN_{\rm c}^{-2}\|(1-\Delta)^{1/2}(\gamma_0-\gamma_{0,N_{\rm c}})\|_{\mathfrak{S}_2(L^2_\#)}.$$

Proof. Using definition (8.4.8) and the fact that $(z-\mathcal{H}_{N_{\rm c}})^{-1}$ and $\gamma_{0,N_{\rm c}}$ commute, we obtain

$$(1-\Delta)^{1/2}\widetilde{Q_{N_{\rm c}}}\gamma_{0,N_{\rm c}} = \frac{1}{2i\pi}\oint_\Gamma (1-\Delta)^{1/2}(z-\mathcal{H}\,)^{-1}\mathcal{V}_{N_{\rm c}}^\perp(z-\mathcal{H}_{N_{\rm c}})^{-1}\mathcal{V}_{N_{\rm c}}^\perp\gamma_{0,N_{\rm c}}(z-\mathcal{H}_{N_{\rm c}})^{-1}dz.$$

Since $\mathrm{Ran}(\mathcal{V}_{N_{\rm c}}^\perp\gamma_{0,N_{\rm c}}) \subset X_{N_{\rm c}}^\perp$, we have $\mathcal{V}_{N_{\rm c}}^\perp\gamma_{0,N_{\rm c}} = \Pi_{N_{\rm c}}^\perp\mathcal{V}_{N_{\rm c}}^\perp\gamma_{0,N_{\rm c}}$. Observing that

$$(z-\mathcal{H}_{N_{\rm c}})^{-1}\Pi_{N_{\rm c}}^\perp = \Pi_{N_{\rm c}}^\perp\left(z+\frac{1}{2}\Delta_{|X_{N_{\rm c}}^\perp}\right)^{-1}\Pi_{N_{\rm c}}^\perp,$$

we thus obtain

$$(z-\mathcal{H}_{N_{\rm c}})^{-1}\mathcal{V}_{N_{\rm c}}^\perp\gamma_{0,N_{\rm c}} = (z-\mathcal{H}_{N_{\rm c}})^{-1}\Pi_{N_{\rm c}}^\perp\mathcal{V}_{N_{\rm c}}^\perp\gamma_{0,N_{\rm c}} = \Pi_{N_{\rm c}}^\perp\left(z+\frac{1}{2}\Delta_{|X_{N_{\rm c}}^\perp}\right)^{-1}\mathcal{V}_{N_{\rm c}}^\perp\gamma_{0,N_{\rm c}}.$$

Therefore,

$$\begin{aligned}(1-\Delta)^{1/2}\widetilde{Q_{N_{\rm c}}}\gamma_{0,N_{\rm c}} = \frac{1}{2i\pi}\oint_\Gamma &\left[(1-\Delta)^{1/2}(z-\mathcal{H}\,)^{-1}(1-\Delta)^{1/2}\right] \\ &\times\left[(1-\Delta)^{-1/2}\mathcal{V}_{N_{\rm c}}^\perp(1-\Delta)^{-1/2}\Pi_{N_{\rm c}}^\perp\right] \\ &\times\left[\Pi_{N_{\rm c}}^\perp(1-\Delta)^{1/2}(z+\frac{1}{2}\Delta)^{-1}(1-\Delta)^{1/2}\Pi_{N_{\rm c}}^\perp\right] \\ &\times\left[(1-\Delta)^{-1/2}\mathcal{V}_{N_{\rm c}}^\perp\gamma_{0,N_{\rm c}}(z-\mathcal{H}_{N_{\rm c}})^{-1}\right]dz.\end{aligned} \tag{8.4.15}$$

First,

$$\begin{aligned}\|(1-\Delta)^{-1/2}\mathcal{V}_{N_{\rm c}}^\perp(1-\Delta)^{-1/2}\Pi_{N_{\rm c}}^\perp\| &= \|(1-\Delta)^{-1/2}\mathcal{V}(1-\Delta)^{-1/2}\Pi_{N_{\rm c}}^\perp\| \\ &= \|\Pi_{N_{\rm c}}^\perp(1-\Delta)^{-1/2}\mathcal{V}(1-\Delta)^{-1/2}\| \\ &\leq \|\Pi_{N_{\rm c}}^\perp(1-\Delta)^{-1}\|\|(1-\Delta)^{1/2}\mathcal{V}(1-\Delta)^{-1/2}\|.\end{aligned}$$

Since $\|(1-\Delta)^{1/2}\mathcal{V}(1-\Delta)^{-1/2}\|$ equals the operator norm of $\mathcal{V}$, considered as a multiplicative operator from $H^1_\#(\Omega)$ to $H^1_\#(\Omega)$, it can be shown using classical Sobolev embeddings that there exists $C \in \mathbb{R}_+$ such that

$$\|(1-\Delta)^{1/2}\mathcal{V}(1-\Delta)^{-1/2}\| \leq C\left(\|\mathcal{V}\|_{L^\infty} + \|\nabla\mathcal{V}\|_{L^3}\right),$$

which is bounded under Assumption 8.2.2. Since $\|\Pi_{N_\mathrm{c}}^\perp(1-\Delta)^{-1}\| \leq (1+N_\mathrm{c}^2)^{-1}$ for all $N_\mathrm{c} \in \mathbb{N}$, there exist $C \in \mathbb{R}_+$ and $N_\mathrm{c}^0 \in \mathbb{N}$ such that for all $N_\mathrm{c} \geq N_\mathrm{c}^0$,

$$\|(1-\Delta)^{-1/2}\mathcal{V}_{N_\mathrm{c}}^\perp(1-\Delta)^{-1/2}\Pi_{N_\mathrm{c}}^\perp\| \leq CN_\mathrm{c}^{-2}. \tag{8.4.16}$$

Finally,

$$\begin{aligned}
\max_{z\in\Gamma} &\|(1-\Delta)^{-1/2}\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\|_{\mathfrak{S}_2(L^2_\#)} \\
&= \left(\sum_{i=1}^N \max_{z\in\Gamma}|z-\lambda_{i,N_\mathrm{c}}|^{-2}\|(1-\Delta)^{-1/2}\mathcal{V}_{N_\mathrm{c}}^\perp\phi_{i,N_\mathrm{c}}\|_{L^2_\#}^2\right)^{1/2} \\
&\leq \max_{\substack{z\in\Gamma,\\ i=1,\dots,N}} |z-\lambda_{i,N_\mathrm{c}}|^{-2}\left(\sum_{i=1}^N \|(1-\Delta)^{-1/2}\mathcal{V}_{N_\mathrm{c}}^\perp\phi_{i,N_\mathrm{c}}\|_{L^2_\#}^2\right)^{1/2} \\
&= \max_{\substack{z\in\Gamma,\\ i=1,\dots,N}} |z-\lambda_{i,N_\mathrm{c}}|^{-2}\|(1-\Delta)^{-1/2}\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}.
\end{aligned}$$

From the definition of the contour Γ, $\max\limits_{\substack{z\in\Gamma,\\ i=1,\dots,N}} |z-\lambda_{i,N_\mathrm{c}}|^{-2}$ is bounded uniformly in N_c for N_c large enough. Hence, combining the above inequality with (8.4.11), we obtain that there exist $C \in \mathbb{R}_+$ and $N_\mathrm{c}^0 \in \mathbb{N}$ such that for all $N_\mathrm{c} \geq N_\mathrm{c}^0$,

$$\max_{z\in\Gamma}\|(1-\Delta)^{-1/2}\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\|_{\mathfrak{S}_2(L^2_\#)} \leq\ C\|(1-\Delta)^{1/2}(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}. \tag{8.4.17}$$

We are now in position to estimate $\|(1-\Delta)^{1/2}\widetilde{Q_{N_\mathrm{c}}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}$. We start from (8.4.15). It is classical that $\max\limits_{z\in\Gamma}\|(1-\Delta)^{1/2}(z-\mathcal{H})^{-1}(1-\Delta)^{1/2}\|$ is bounded (see e.g. [46, Lemma 1]). Moreover, $\max\limits_{z\in\Gamma}\|\Pi_{N_\mathrm{c}}^\perp(1-\Delta)^{1/2}(z+\frac{1}{2}\Delta)^{-1}(1-\Delta)^{1/2}\Pi_{N_\mathrm{c}}^\perp\|$ is also bounded. Using estimates (8.4.16) and (8.4.17) allows one to conclude the proof of the lemma. □

Lemma 8.4.8. *There exist $C \in \mathbb{R}_+$ and $N_\mathrm{c}^0 \in \mathbb{N}$ such that for all $N_\mathrm{c} \geq N_\mathrm{c}^0$,*

$$\|(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\widetilde{Q_{N_\mathrm{c}}}\|_{\mathfrak{S}_2(L^2_\#)} \leq CN_\mathrm{c}^{-2}\|(1-\Delta)^{1/2}(\gamma_0-\gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)}.$$

Proof. Noting that $\gamma_{0,N_\mathrm{c}}^2 = \gamma_{0,N_\mathrm{c}}$, using (8.2.13) and the cyclicity of the trace, we obtain

$$\begin{aligned}
\|(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\widetilde{Q_{N_\mathrm{c}}}\|_{\mathfrak{S}_2(L^2_\#)} &\leq \|(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}\|\gamma_{0,N_\mathrm{c}}\widetilde{Q_{N_\mathrm{c}}}\|_{\mathfrak{S}_2(L^2_\#)} \\
&= \|(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}\|\widetilde{Q_{N_\mathrm{c}}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)},
\end{aligned}$$

since γ_{0,N_c} is a finite-rank orthogonal projector. Moreover, as the orbitals $(\phi_{i,N_\mathrm{c}})_{i=1,\cdots,N}$ are bounded in $H^1_\#(\Omega)$ uniformly in N_c, $\|(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}$ is also bounded uniformly in N_c. On top of that, noting that $\|(1-\Delta)^{-1/2}\| \leq 1$, we have

$$\|\widetilde{Q_{N_\mathrm{c}}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq \|(1-\Delta)^{-1/2}\|\|(1-\Delta)^{1/2}\widetilde{Q_{N_\mathrm{c}}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq \|(1-\Delta)^{1/2}\widetilde{Q_{N_\mathrm{c}}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}.$$

Therefore, we can use the estimate of Lemma 8.4.7 to conclude. □

From Lemma 8.4.4, and using the estimates of Lemmas 8.4.5, 8.4.7 and 8.4.8, we easily get estimate (8.4.1).

Proof of estimate (8.4.2)

If the perturbed density matrix were satisfying $\widetilde{\gamma_{N_c}} \in \mathcal{K}$, we could deduce from (8.2.16) that the error $\widetilde{\mathcal{E}_{0,N_c}} - \mathcal{E}_0$ would be non-negative and converge to zero as $\|(1-\Delta)^{1/2}(\widetilde{\gamma_{N_c}} - \gamma_0)\|^2_{\mathfrak{S}_2(L^2_\#)}$ when N_c goes to infinity, yielding an improvement factor for the energy of order N_c^{-4}. However, as pointed out in Remark 8.3.2, $\widetilde{\gamma_{N_c}}$ does not belong to $\mathcal{K}$ in general. We are going to show that the improvement factor for the energy is in fact of order N_c^{-2}.

We have $\widetilde{\mathcal{E}_{0,N_c}} - \mathcal{E}_0 = \mathrm{Tr}\,(\gamma_{0,N_c}\mathcal{H}\widetilde{\gamma_{N_c}}) - \mathrm{Tr}\,(\gamma_0\mathcal{H}\gamma_0)$. As $\mathrm{Tr}\,\left((\gamma_0)^2\right) = N$ and $\mathrm{Tr}\,(\gamma_{0,N_c}\widetilde{\gamma_{N_c}}) = \mathrm{Tr}\,(\gamma_{0,N_c}^2) = N$, the energy difference can be written as follows

$$\begin{aligned}
\widetilde{\mathcal{E}_{0,N_c}} - \mathcal{E}_0 &= \mathrm{Tr}\,(\gamma_{0,N_c}(\mathcal{H} - \epsilon_F)\widetilde{\gamma_{N_c}}) - \mathrm{Tr}\,(\gamma_0(\mathcal{H} - \epsilon_F)\gamma_0) \\
&= \mathrm{Tr}\,((\gamma_{0,N_c} - \gamma_0)(\mathcal{H} - \epsilon_F)(\widetilde{\gamma_{N_c}} - \gamma_0)) + \mathrm{Tr}\,(\gamma_0(\mathcal{H} - \epsilon_F)(\widetilde{\gamma_{N_c}} + \gamma_{0,N_c} - 2\gamma_0)) \\
&= \mathrm{Tr}\,((\gamma_{0,N_c} - \gamma_0)(\mathcal{H} - \epsilon_F)(\widetilde{\gamma_{N_c}} - \gamma_0)) \\
&\quad + 2\mathrm{Tr}\,(\gamma_0(\mathcal{H} - \epsilon_F)(\gamma_{0,N_c} - \gamma_0)) + \mathrm{Tr}\,\left(\gamma_0(\mathcal{H} - \epsilon_F)\gamma_{N_c}^{(1)}\right). \qquad (8.4.18)
\end{aligned}$$

We now estimate each of these three terms. First, noting that $(\mathcal{H} - \epsilon_F) = -\gamma_0|\mathcal{H} - \epsilon_F| + (1 - \gamma_0)|\mathcal{H} - \epsilon_F|$, using the triangle and the Cauchy–Schwarz inequalities, the fact that $|\mathcal{H} - \epsilon_F|, \gamma_0$ and $(1 - \gamma_0)$ commute, and (8.2.13), we get

$$\begin{aligned}
|\mathrm{Tr}\,((\gamma_{0,N_c} - \gamma_0)(\mathcal{H} - \epsilon_F)(\widetilde{\gamma_{N_c}} - \gamma_0))\,| &= \Big|\mathrm{Tr}\,((\gamma_{0,N_c} - \gamma_0)(1 - \gamma_0)|\mathcal{H} - \epsilon_F|(\widetilde{\gamma_{N_c}} - \gamma_0)) \\
&\quad - \mathrm{Tr}\,((\gamma_{0,N_c} - \gamma_0)\gamma_0|\mathcal{H} - \epsilon_F|(\widetilde{\gamma_{N_c}} - \gamma_0))\Big| \\
&\le \Big(\||\mathcal{H} - \epsilon_F|^{1/2}(1 - \gamma_0)(\gamma_{0,N_c} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)} \\
&\quad + \||\mathcal{H} - \epsilon_F|^{1/2}\gamma_0(\gamma_{0,N_c} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}\Big) \\
&\quad \times \|\mathcal{H} - \epsilon_F|^{1/2}(\widetilde{\gamma_{N_c}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)} \\
&\le 2\||\mathcal{H} - \epsilon_F|^{1/2}(\gamma_{0,N_c} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)} \\
&\quad \times \||\mathcal{H} - \epsilon_F|^{1/2}(\widetilde{\gamma_{N_c}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}.
\end{aligned}$$

From (8.2.14) and (8.4.1), there exist $C \in \mathbb{R}_+$ and $N_c^0 \in \mathbb{N}$, such that for all $N_c \ge N_c^0$,

$$|\mathrm{Tr}\,((\gamma_{0,N_c} - \gamma_0)(\mathcal{H} - \epsilon_F)(\widetilde{\gamma_{N_c}} - \gamma_0))\,| \le CN_c^{-2}\|(1 - \Delta)^{1/2}(\gamma_{0,N_c} - \gamma_0)\|^2_{\mathfrak{S}_2(L^2_\#)}. \qquad (8.4.19)$$

Second, noting that for all $i = 1, \dots, N$, $\langle\phi_i^0|\gamma_0 - \gamma_{0,N_c}|\phi_i^0\rangle \ge 0$, we get

$$\begin{aligned}
|\mathrm{Tr}\,(\gamma_0(\mathcal{H} - \epsilon_F)(\gamma_{0,N_c} - \gamma_0)| &= \left|\sum_{i=1}^N (\lambda_i^0 - \epsilon_F)\langle\phi_i^0|\gamma_{0,N_c} - \gamma_0|\phi_i^0\rangle\right| \\
&\le \max_{i=1,\dots,N}|\lambda_i^0 - \epsilon_F| \sum_{i=1}^N \langle\phi_i^0|\gamma_0 - \gamma_{0,N_c}|\phi_i^0\rangle \\
&= \max_{i=1,\dots,N}|\lambda_i^0 - \epsilon_F| \left(N - \mathrm{Tr}\,(\gamma_{0,N_c}\gamma_0)\right) \\
&= \frac{1}{2}\max_{i=1,\dots,N}|\lambda_i^0 - \epsilon_F|\,\|\gamma_0 - \gamma_{0,N_c}\|^2_{\mathfrak{S}_2(L^2_\#)}.
\end{aligned}$$

From (8.2.25), there exist $C \in \mathbb{R}_+$ and $N_c^0 \in \mathbb{N}$, such that for all $N_c \ge N_c^0$,

$$|\mathrm{Tr}\,((\gamma_{0,N_c} - \gamma_0)(\mathcal{H} - \epsilon_F)(\widetilde{\gamma_{N_c}} - \gamma_0))\,| \le CN_c^{-2}\|(1 - \Delta)^{1/2}(\gamma_0 - \gamma_{0,N_c})\|^2_{\mathfrak{S}_2(L^2_\#)}. \qquad (8.4.20)$$

Third, noting that for $i,j=1,\dots,N$, $\langle\phi^{(1)}_{j,N_\mathrm{c}}|\phi^0_{i,N_\mathrm{c}}\rangle=0$, $\|\phi^0_i\|_{L^2_\#}=1$, $\|\phi_{j,N_\mathrm{c}}\|_{L^2_\#}=1$, and using (8.3.9) and the Cauchy–Schwarz inequality, we get

$$\begin{aligned}
\left|\mathrm{Tr}\left(\gamma_0(\mathcal{H}-\epsilon_F)\gamma^{(1)}_{N_\mathrm{c}}\right)\right| &= \left|\sum_{i=1}^N(\lambda^0_i-\epsilon_F)\langle\phi^0_i|\gamma^{(1)}_{N_\mathrm{c}}|\phi^0_i\rangle\right| \\
&= 2\left|\sum_{i=1}^N\sum_{j=1}^N(\lambda^0_i-\epsilon_F)\langle\phi^0_i|\phi_{j,N_\mathrm{c}}\rangle\langle\phi^{(1)}_{j,N_\mathrm{c}}|\phi^0_i\rangle\right| \\
&\le 2\max_{i=1,\dots,N}|\lambda^0_i-\epsilon_F|\sum_{i=1}^N\sum_{j=1}^N\left|\langle\phi^0_i|\phi_{j,N_\mathrm{c}}\rangle\langle\phi^{(1)}_{j,N_\mathrm{c}}|\phi^0_i-\phi^0_{i,N_\mathrm{c}}\rangle\right| \\
&\le 2\max_{i=1,\dots,N}|\lambda^0_i-\epsilon_F|\sum_{i=1}^N\|\phi^0_i-\phi^0_{i,N_\mathrm{c}}\|_{L^2_\#}\sum_{j=1}^N\|\phi^{(1)}_{j,N_\mathrm{c}}\|_{L^2_\#} \\
&\le 2N\max_{i=1,\dots,N}|\lambda^0_i-\epsilon_F|\|\Phi^0-\Phi^0_{N_\mathrm{c}}\|_{L^2_\#}\left(\sum_{j=1}^N\|\phi^{(1)}_{j,N_\mathrm{c}}\|^2_{L^2_\#}\right)^{1/2}.
\end{aligned}$$

Let us now estimate $\sum_{j=1}^N\|\phi^{(1)}_{j,N_\mathrm{c}}\|^2_{L^2_\#}$. Using (8.3.6)–(8.3.7) and noting that $\Pi^\perp_{N_\mathrm{c}}$ and $(1-\Delta)^{-1/2}$ commute, we get

$$\begin{aligned}
\sum_{i=1}^N\|\phi^{(1)}_{i,N_\mathrm{c}}\|^2_{L^2_\#} &= \sum_{i=1}^N\|(-\frac{1}{2}\Delta-\lambda_{i,N_\mathrm{c}})^{-1}\Pi^\perp_{N_\mathrm{c}}\mathcal{V}^\perp_{N_\mathrm{c}}\phi_{i,N_\mathrm{c}}\|^2_{L^2_\#} \\
&\le \max_{i=1,\cdots,N}\|(-\frac{1}{2}\Delta-\lambda_{i,N_\mathrm{c}})^{-1}(1-\Delta)^{1/2}\Pi^\perp_{N_\mathrm{c}}\|^2\sum_{i=1}^N\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}\phi_{i,N_\mathrm{c}}\|^2_{L^2_\#} \\
&\le (1+N_\mathrm{c}^2)^{-1}\max_{i=1,\cdots,N}\|(-\frac{1}{2}\Delta-\lambda_{i,N_\mathrm{c}})^{-1}(1-\Delta)\|^2\sum_{i=1}^N\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}\phi_{i,N_\mathrm{c}}\|^2_{L^2_\#} \\
&= (1+N_\mathrm{c}^2)^{-1}\max_{i=1,\cdots,N}\|(-\frac{1}{2}\Delta-\lambda_{i,N_\mathrm{c}})^{-1}(1-\Delta)\|^2\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}\|^2_{\mathfrak{S}_2(L^2_\#)}.
\end{aligned}$$

Therefore, since $\max_{i=1,\cdots,N}\|(-\frac{1}{2}\Delta-\lambda_{i,N_\mathrm{c}})^{-1}(1-\Delta)\|^2$ is bounded uniformly in N_c, we deduce from (8.4.11) that there exist $C\in\mathbb{R}_+$ and $N_\mathrm{c}^0\in\mathbb{N}$, such that for all $N_\mathrm{c}\ge N_\mathrm{c}^0$,

$$\sum_{i=1}^N\|\phi^{(1)}_{i,N_\mathrm{c}}\|^2_{L^2_\#} \le CN_\mathrm{c}^{-2}\|(1-\Delta)^{1/2}(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|^2_{\mathfrak{S}_2(L^2_\#)}. \tag{8.4.21}$$

From (8.2.22), (8.2.25) and (8.4.21), we obtain that there exist $C\in\mathbb{R}_+$ and $N_\mathrm{c}^0\in\mathbb{N}$, such that for all $N_\mathrm{c}\ge N_\mathrm{c}^0$,

$$\left|\mathrm{Tr}\left(\gamma_0(\mathcal{H}-\epsilon_F)\gamma^{(1)}_{N_\mathrm{c}}\right)\right|\le CN_\mathrm{c}^{-2}\|(1-\Delta)^{1/2}(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|^2_{\mathfrak{S}_2(L^2_\#)}. \tag{8.4.22}$$

Putting together (8.4.18), (8.4.19), (8.4.20), and (8.4.22), we obtain estimate (8.4.2).

8.5 Numerical results

We present in this section some results to illustrate the statements of Theorem 8.4.1 for several eigenvalue clusters and potentials with different regularities. We focus mainly on the conver-

gence rate improvement, and not in the low computational computational cost of the method, which has been demonstrated for the nonlinear problem of Kohn–Sham equations in [50].

In all what follows, we consider a domain $\Omega = [0, 10]^3$ in atomic units (a.u.). The computed solutions are compared to a reference solution, which is a solution computed on a very fine grid with a kinetic energy cutoff $E_{\rm ref} = 800$ a.u. depends, which corresponds to a discretization parameter $N_{\rm ref} \simeq 58.5$, and 382323 Fourier coefficients per orbital. In each case, we denote the reference energy by $\mathcal{E}_0$ and the reference density matrix by γ_0.

The coarse solutions are computed on a grid with cutoff E_c, and corresponding N_c, and have energy $\mathcal{E}_{0,N_{\rm c}}$ and density matrix $\gamma_{0,N_{\rm c}}$. In order to avoid errors coming from the size of the finite grid used for the computation of the corrections, we compute the post-processed solutions in the same grid as the reference solution, *i.e.* on a grid with energy cutoff $E_{\rm ref}$. Note that the components of the orbitals on the coarse grid are not modified by the post-processing. One only needs to compute the coefficients corresponding to basis functions with wave-numbers larger than E_c.

The implementation is based on KSSOLV [252], a Matlab library for solving Kohn–Sham equations. We use the linear eigenvalue resolution routine for solving (8.3.3). However, the routine has been modified in order to take into account all necessary Fourier coefficients of the potential $\mathcal{V}$ when solving the problem on coarse grids with energy cutoffs E_c.

The tested potentials denoted by $\mathcal{V}^s$ are defined by their Fourier coefficients as

$$\hat{\mathcal{V}}^s_{\mathbf{0}} = 0, \quad \text{and} \quad \forall \mathbf{k} \in \mathcal{R}^* \backslash \{\mathbf{0}\}, \quad \hat{\mathcal{V}}^s_{\mathbf{k}} = -\frac{c_s}{|\mathbf{k}|^{2s}},$$

where s is a regularity parameter which varies between 1 and 2, and c_s a multiplicative constant. For $s > 5/4$, the potential $\mathcal{V}^s$ is smooth enough to verify Assumption 8.2.2, hence we should observe the improvement in the convergence rate given in Theorem 8.4.1. For $s \leq 5/4$, the potential does not verify Assumption 8.2.2, but we can compute post-processed eigenfunctions and eigenvalues. It actually still yields an improvement on the energy and the density matrix. Note that for $s = 1$, this potential has the same regularity as the Coulomb potential.

For all the tested potentials, the lowest eigenvalue of the Hamiltonian $-\frac{1}{2}\Delta + \mathcal{V}$ is simple. There are gaps between the $5^{\rm th}$ and $6^{\rm th}$ eigenvalues and the $10^{\rm th}$ and $11^{\rm th}$ eigenvalues. Therefore, in the following, we present the results of the post-processing method for clusters including one, five, and ten eigenvalues. This guarantees that the gap Assumption 8.2.1 is satisfied. Let us remind that we consider the lowest eigenvalues of the Hamiltonian.

In Subsection 8.5.1, we show how the post-processing procedure decreases both the energy error $\mathcal{E}_{0,N_{\rm c}} - \mathcal{E}_0$ and the Hilbert–Schmidt norm of the density matrix error $\|(1-\Delta)^{1/2}(\gamma_{0,N_{\rm c}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}$ in the case of a potential with regularity coefficient $s = 2$ and a cluster of five eigenvalues. In Subsection 8.5.2, we study the convergence rate improvement of both the energy and the density matrix for different clusters of eigenvalues, still in the case of a potential with regularity coefficient $s = 2$. Finally, we study in Subsection 8.5.3 the influence of the potential regularity on the convergence rate improvement for the energy and the density matrix, in the case of the cluster composed of the five lowest eigenvalues.

8.5.1 Convergence of the density matrix and the energy

We consider the potential $\mathcal{V}^2$ with Fourier coefficients

$$\hat{\mathcal{V}}^2_{\mathbf{0}} = 0, \quad \text{and} \quad \forall \mathbf{k} \in \mathcal{R}^* \backslash \{\mathbf{0}\}, \quad \hat{\mathcal{V}}^2_{\mathbf{k}} = -\frac{0.01}{|\mathbf{k}|^4}.$$

For all energy cutoffs E_c between 10 and 200 a.u. by step of 10, we compute the five lowest eigenvalues and eigenvectors of the discrete Hamiltonian. We build the discrete density matrix γ_{0,N_c} as in (8.2.19), and compute the discrete energy $\mathcal{E}_{0,N_c}$ (8.2.20). Then, we apply the post-processing as described in Section 8.3 and we compute the post-processed density matrix $\widetilde{\gamma_{N_c}}$ as well as the perturbed energy $\widetilde{\mathcal{E}_{0,N_c}}$. For the potential, we choose such a small multiplicative constant to better observe the asymptotic regime numerically within the range of tested cutoffs E_c.

As we can see on the left part of Figure 8.1, the energy error between the post-processed energy and the reference energy $\widetilde{\mathcal{E}_{0,N_c}} - \mathcal{E}_0$ is 5 to 50 times smaller than the energy error between the coarse energy and the reference energy $\mathcal{E}_{0,N_c} - \mathcal{E}_0$. More precisely, the energy error is reduced by a factor of about 5 for small values of N_c and up to 50 for large values of N_c.

We observe a similar behavior for the density matrix error on the right part of Figure 8.1. Indeed, the Hilbert–Schmidt norm of the difference between the reference and the coarse density matrices $\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_c})\|_{\mathfrak{S}_2(L^2_\#)}$ is 5 to 50 times larger than the error between the reference and the post-processed density matrix $\|(1-\Delta)^{1/2}(\gamma_0 - \widetilde{\gamma_{N_c}})\|_{\mathfrak{S}_2(L^2_\#)}$.

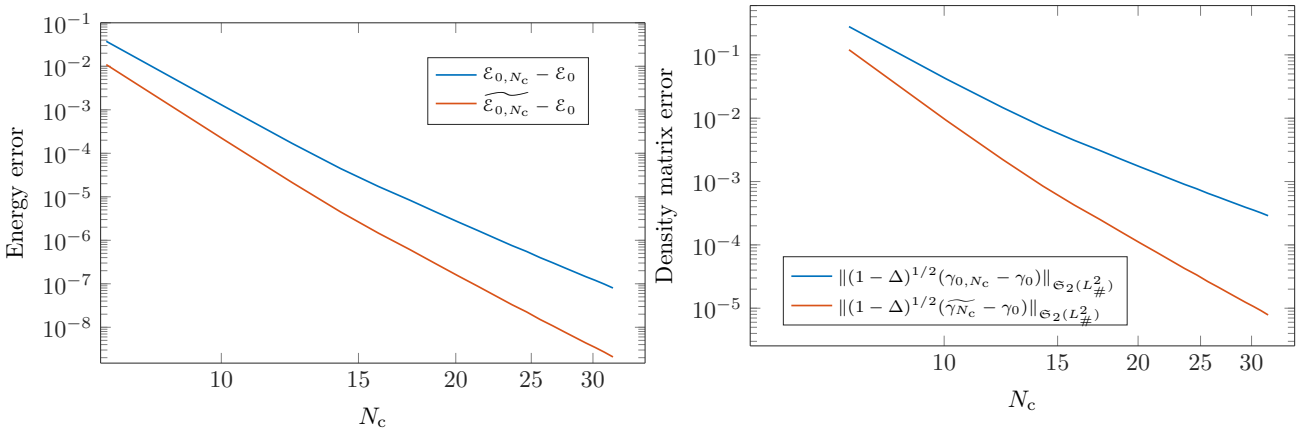

Figure 8.1 – Left: plot of the energy errors $\widetilde{\mathcal{E}_{0,N_c}} - \mathcal{E}_0$ and $\mathcal{E}_{0,N_c} - \mathcal{E}_0$ for energy cutoffs E_c between 10 and 200 a.u. Right: plot the density matrix error $\|(1-\Delta)^{1/2}(\gamma_0 - \widetilde{\gamma_{N_c}})\|_{\mathfrak{S}_2(L^2_\#)}$ and $\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_c})\|_{\mathfrak{S}_2(L^2_\#)}$
for energy cutoffs between 10 and 200 a.u. This corresponds to values of N_c between 7 and 31.

8.5.2 Comparison between different eigenvalue clusters

We now consider three different eigenvalue clusters composed of one, four and ten eigenvalues, with the same potential $\mathcal{V}^2$. The three corresponding gaps are respectively equal to $8.84.10^{-1}$, $1.80.10^{-1}$ and $3.42.10^{-1}$.

For these three clusters, we compute a reference solution, and then we compute discrete solutions within cutoffs E_c varying between 10 and 200 a.u. On the left of Figure 8.2, we plot the ratio between the energy error with post-processing and without post-processing $\frac{\widetilde{\mathcal{E}_{0,N_c}} - \mathcal{E}_0}{\mathcal{E}_{0,N_c} - \mathcal{E}_0}$ for

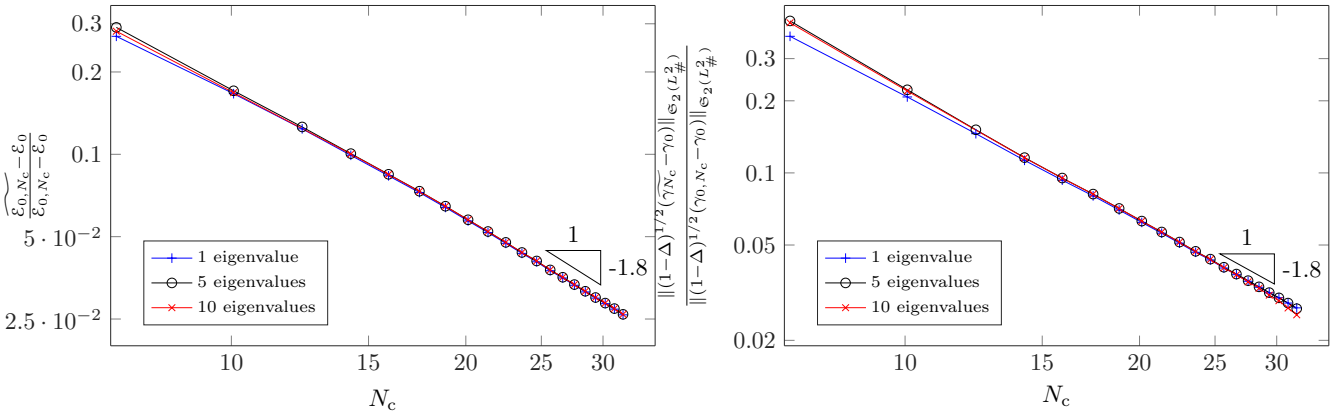

Figure 8.2 – Plots of the energy error ratio (left) and the density matrix error ratio (right) for three different clusters of eigenvalues (1, 5 and 10 eigenvalues) with a potential with regularity coefficient $s = 2.2$.

the three different cases. According to Theorem 8.4.1, this ratio should at least decrease as N_c^{-2} in the asymptotic regime of large N_c's. Numerically, the ratio decreases about as $N_c^{-1.8}$ when N_c is large, similarly in the three test cases. The ratio of the Hilbert–Schmidt norms of the error on the density matrix $\frac{\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_c})\|_{\mathfrak{S}_2(L^2_\#)}}{\|(1-\Delta)^{1/2}(\gamma_0 - \widetilde{\gamma_{N_c}})\|_{\mathfrak{S}_2(L^2_\#)}}$ behaves also like $N_c^{-1.8}$ in the asymptotic regime, as shown on the left of Figure 8.2.

8.5.3 Comparison of different regularities

Lastly, we compute the post-processing in the case of the cluster composed of the five lowest eigenvalues with four potentials having different regularity coefficients. More precisely, we consider potentials $\mathcal{V}^1, \mathcal{V}^{1.25}, \mathcal{V}^{1.5}$, and $\mathcal{V}^2$,, with constants c_s equal to 0.01. In theory, the potentials $\mathcal{V}^2$ and $\mathcal{V}^{1.5}$ verify Assumption 8.2.2, the potential $\mathcal{V}^{1.25}$ is just at the limit, and $\mathcal{V}^1$ does not verify this assumption. However, for each of these potentials, we can compute the post-processed energy and density matrix. Numerically, we observe the same improved rate of convergence both for the energy error ratio (on the top of Figure 8.3) and for the density matrix error ratio (on the bottom of Figure 8.3) for the potentials $\mathcal{V}^2$ and $\mathcal{V}^{1.5}$ close to N_c^{-2}, in fact about $N_c^{-1.8}$. For the potentials $\mathcal{V}^{1.25}$ and $\mathcal{V}^1$, the improvement is lower in rate, but we still observe an improvement, which is about $N_c^{-1.5}$ for the energy for both potentials, and about $N_c^{-1.6}$ for the density matrix for $\mathcal{V}^{1.25}$, and $N_c^{-1.3}$ for $\mathcal{V}^1$. Note that for these two cases, the tested E_c are far from convergence, so the asymptotic regime is clearly not reached. However, testing larger values of N_c is numerically costly. Thus, with a low regularity, we observe an improved convergence both for the energy and the density matrix already in the preasymptotic regime.

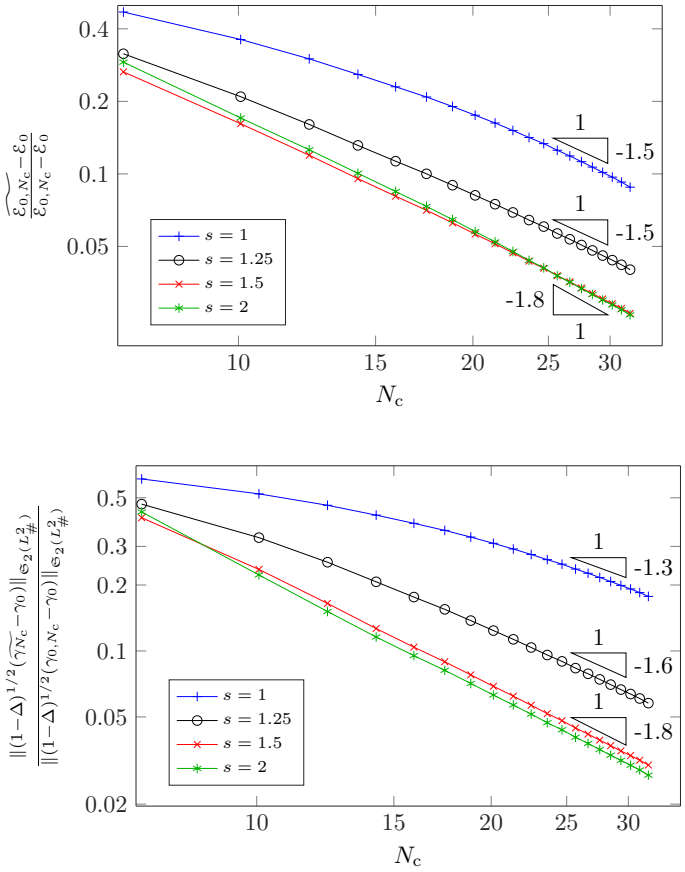

Figure 8.3 – Plots of the energy error improvement (top) and the density matrix error improvement (bottom) for four different regularities for the potential: $s = 1, 1.25, 1.5, 2$.

Appendix: proof of Lemma 8.2.3

We start by proving (8.2.22). Denoting by M_{N_c} the $N \times N$ overlap matrix with entries $(M_{N_c})_{i,j} = \langle \phi^0_{i,N_c} | \phi^0_j \rangle$, we have

$$\|\gamma_{0,N_c} - \gamma_0\|^2_{\mathfrak{S}_2(L^2_\#)} = 2\left(N - \sum_{i,j=1}^{N} |\langle \phi^0_{j,N_c} | \phi^0_i \rangle|^2\right) = 2\left(N - \mathrm{Tr}\left(M_{N_c} M_{N_c}^T\right)\right)$$

and

$$\|\Phi^0_{N_c} - \Phi^0\|^2_{L^2_\#} = 2\left(N - \sum_{i=1}^{N} \langle \phi^0_{i,N_c} | \phi^0_i \rangle\right) = 2\left(N - \mathrm{Tr}\left(M_{N_c}\right)\right).$$

We therefore have to show that

$$2N - 2\mathrm{Tr}\left(M_{N_c}\right) \leq 2N - 2\mathrm{Tr}\left(M_{N_c} M_{N_c}^T\right) \leq 2(2N - 2\mathrm{Tr}\left(M_{N_c}\right)).$$

Since $\Phi^0_{N_c}$ belongs to $\mathcal{M}^{\Phi^0}$, we have from [44, Lemma 4.3]

$$M_{N_c} = M_{N_c}^T = (M_{N_c} M_{N_c}^T)^{1/2} \quad \text{and} \quad 0 \leq M_{N_c} \leq 1.$$

Hence,

$$\mathrm{Tr}\,(M_{N_\mathrm{c}} M_{N_\mathrm{c}}^T) = \mathrm{Tr}\,(M_{N_\mathrm{c}}^2) \leq \mathrm{Tr}\,(M_{N_\mathrm{c}}),$$

from which we deduce the left inequality in (8.2.22). The right inequality in (8.2.22) holds since

$$2N - 4\mathrm{Tr}\,(M_{N_\mathrm{c}}) + 2\mathrm{Tr}\,(M_{N_\mathrm{c}} M_{N_\mathrm{c}}^T) = 2N - 4\mathrm{Tr}\,(M_{N_\mathrm{c}}) + 2\mathrm{Tr}\,(M_{N_\mathrm{c}}^2) = 2\mathrm{Tr}\,\left((M_{N_\mathrm{c}} - I_N)^2\right) \geq 0.$$

Let us now show (8.2.23). From [44, Theorem 4.2], there exist two constants $0 < c \leq C < \infty$ and $N_\mathrm{c}^0 \in \mathbb{N}$ such that for all $N_\mathrm{c} \geq N_\mathrm{c}^0$,

$$c\|(1-\Delta)^{1/2}(\Phi_{N_\mathrm{c}}^0 - \Phi^0)\|_{L^2_\#}^2 \leq \mathcal{E}_{0,N_\mathrm{c}} - \mathcal{E}_0 \leq C\|(1-\Delta)^{1/2}(\Phi_{N_\mathrm{c}}^0 - \Phi^0)\|_{L^2_\#}^2 .$$

Hence, using (8.2.16) finishes the proof.

Acknowledgements

This work was partially undertaken in the framework of CALSIMLAB, supported by the public grant ANR-11-LABX- 0037-01 overseen by the French National Research Agency (ANR) as part of the Investissements d'avenir program (reference: ANR-11-IDEX-0004-02). Financial support from the ANR grant BECASIM (reference ANR-12-MONU-0007-02) is also acknowledged.

Chapter 9

Post-processing of the plane-wave approximation of nonlinear Schrödinger equations

We expose in this chapter the results of [97].

Abstract

In the first part of this article [48], we have presented *a priori* estimates for the perturbation-based post-processing of the plane-wave approximation of *linear* Schrödinger equations. In this article, we extend the proofs of such estimates in the nonlinear case of Kohn–Sham LDA models with pseudopotentials. As in [50], where these *a priori* results were announced and tested numerically, we use a periodic setting, and the problem is discretized with planewaves (Fourier series). This post-processing method consists of performing a full computation in a coarse planewave basis, and then to compute corrections based on first-order perturbation theory in a fine basis, which numerically only requires the computation of the residuals of the ground-state orbitals in the fine basis. We show that this procedure asymptotically improves the accuracy of two quantities of interest: the ground-state density matrix, *i.e.* the orthogonal projector on the lowest N eigenvectors, and the ground-state energy.

9.1 Introduction

To determine the electronic ground-state of a system within the Born–Oppenheimer approximation [31], DFT Kohn–Sham models [152] are among the state-of-the-art methods, especially for their good trade-off between accuracy and computational cost. In the context of condensed matter physics and materials science, most simulations of the Kohn–Sham models are performed with periodic boundary conditions, for which a planewave (Fourier) discretization method is particularly suited (see the introduction of [48] for more detail on the physical context). Nevertheless, this method scales cubically with respect to the number of electrons in the system, and becomes expensive for large systems.

In previous works [49, 50, 48], we have proposed a post-processing method to provide cheaper and still accurate results for this problem. This two-grid method consists of computing first a rough approximation of the solution to the Kohn–Sham problem in a coarse planewave basis. This solution is then corrected in a fine basis, based on first-order Rayleigh–Schrödinger perturbation theory, considering the exact Kohn–Sham ground-state as a perturbation of the approximate ground-state computed in the coarse basis. As shown in [50], this method leads numerically to a substantial improvement for the ground-state energy, the improvement factor varying between 10 and 100 for small size systems such as the alanine molecule. Besides, the computational extra-cost did not exceed about 3-5% of the total computations, depending on the size of the chosen fine basis.

In this article, we focus on the theoretical improvement of this post-processing method for the Kohn–Sham problem. We provide the proofs of theoretical estimates presented in [50], which partly rely on the proofs for the linear subproblem of the Kohn–Sham model presented in the first part of this contribution [48]. Compared to the procedure proposed in [48], we construct here two different post-processed sets of orbitals from the ground-state orbitals of the discrete Kohn–Sham problem in the coarse basis. The first one is derived directly from first-order Rayleigh–Schrödinger perturbation theory, but is not *a priori* orthonormal; the second one is orthonormal. From these two sets of orbitals, we define in Lemma 9.4.1 two corresponding density matrices, which are orthogonal projectors on the space spanned by the post-processed orbitals, and two post-processed energies. Note that, since the problem is nonlinear, the corrections given by the perturbative expansion at first-order cannot be computed exactly. However, we derive that the neglected uncomputable contributions are *a priori* small.

The main result of this article is provided in Theorem 9.5.1. We show that, as in the linear case, the convergence rates of both the post-processed ground-state density matrices and the post-processed ground-state energies are improved within the asymptotic regime where the discretization space is large enough. On top of that, we show that the two versions of the post-processing lead to the same improvement on the density matrix error, but to different improvements on the energy. Indeed, only the post-processed energy computed from the orthonormal post-processed orbitals presents a convergence doubling compared to the density matrix error. These results are valid under the assumption that there is gap between the highest occupied orbital and the lowest unoccupied orbital, which corresponds to considering insulators. All other assumptions come from the *a priori* analysis for the Kohn–Sham problem and do not differ from [44]. Also, our post-processing method crucially relies on the fact that the Laplace operator, which is the leading part in the Hamiltonian, is diagonal in a planewave basis, so that it commutes with the orthogonal projector on the discretization space.

This article is organized as follows. In Section 9.2.1, we present the Kohn–Sham model in the periodic setting, and define the main quantities of interest: the ground-state orbitals

$(\phi_1^0, \ldots, \phi_N^0)$, the density matrix γ_0 and the energy $\mathcal{E}_0$. In Section 9.2.2, we briefly recall the functional setting used in the following sections. In Section 9.3.1, we present the planewave discretization of this Kohn–Sham problem. In Section 9.3.2, we detail the smoothness assumptions on the potentials in the Hamiltonian and we recall *a priori* estimates derived in [44]. We also translate these results in terms of density matrix formalism. In Section 9.4, we describe the post-processing method based on Rayleigh–Schrödinger perturbation theory, and in particular define the corrections. In Section 9.5.1, we present the main results of this paper, *i.e.* an improved convergence rate on the post-processed ground-state density matrices and energies. The proofs are given in Section 9.5.2.

9.2 Periodic Kohn–Sham models with pseudopotentials

9.2.1 Problem setting

In this article, we adopt the system of atomic units, for which $\hbar = 1$, $m_e = 1$, $e = 1$, $4\pi\epsilon_0 = 1$. Thus, the electric charge of the electron is -1, and the charges of the nuclei are positive integers. We consider a periodic setting, therefore the nuclear configuration is supposed to be $\mathcal{R}$-periodic, $\mathcal{R}$ being a periodic lattice with corresponding supercell Ω. To simplify the notation, we consider a cubic lattice $\mathcal{R} = L\mathbb{Z}^3$ ($L > 0$), which corresponds to a cubic supercell $\Omega = [0, L)^3$. But our arguments also apply in the more general case of a lattice with lower or no point symmetry. For $1 \leq p \leq \infty$ and $s \in \mathbb{R}_+$, we denote by

$$L^p_\#(\Omega) := \left\{ u \in L^p_{\mathrm{loc}}(\mathbb{R}^3, \mathbb{R}) \mid u \text{ is } \mathcal{R}\text{-periodic} \right\},$$
$$H^s_\#(\Omega) := \left\{ u \in H^s_{\mathrm{loc}}(\mathbb{R}^3, \mathbb{R}) \mid u \text{ is } \mathcal{R}\text{-periodic} \right\},$$

the spaces of real-valued $\mathcal{R}$-periodic L^p and H^s functions.

We consider a spin-restricted LDA Kohn–Sham model [152] with pseudopotentials. This method is typically used for computing condensed phase properties, when the number of atoms in the simulation cell is limited. A detailed presentation of this model employing the same notations can be found in [50, Section 2], see also [44]. We recall here only the main features of the model. Given a system with N valence electron pairs, we are considering the following energy functional

$$\mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\Psi) = \sum_{i=1}^{N} \int_\Omega |\nabla \psi_i|^2 + \int_\Omega V_{\mathrm{local}}\, \rho_{[\Psi]} + 2\sum_{i=1}^{N} \langle \psi_i | V_{\mathrm{nl}} | \psi_i \rangle + \frac{1}{2} D_\Omega(\rho_{[\Psi]}, \rho_{[\Psi]}) + E^c_{\mathrm{xc},\Omega}(\rho_{[\Psi]}), \quad (9.2.1)$$

where the different terms of the energy are described below. The set of admissible states is

$$\mathcal{M} = \left\{ \Psi = (\psi_1, \ldots, \psi_N)^T \in \left[H^1_\#(\Omega)\right]^N \;\middle|\; \int_\Omega \psi_i \psi_j = \delta_{ij} \right\}. \quad (9.2.2)$$

The electronic density reads

$$\rho_{[\Psi]}(\mathbf{r}) = 2\sum_{i=1}^{N} |\psi_i(\mathbf{r})|^2. \quad (9.2.3)$$

The Coulomb energy is defined as

$$D_\Omega(\rho, \rho') = \int_\Omega \int_\Omega G_\Omega(\mathbf{r}, \mathbf{r}') \rho(\mathbf{r}) \rho'(\mathbf{r}') d\mathbf{r} d\mathbf{r}' = \int_\Omega \rho(\mathbf{r}') [V_{\mathrm{coul}}(\rho')](\mathbf{r}') d\mathbf{r}',$$

where the Green's function G_Ω and the periodic Coulomb potential $V_{\rm coul}(\rho')$ are respectively solutions to the following problems

$$\left\{\begin{array}{l} -\Delta G_\Omega = 4\pi\left(\displaystyle\sum_{\mathbf{k}\in\mathcal{R}}\delta_{\mathbf{k}} - \frac{1}{|\Omega|}\right) \quad \text{in} \quad \mathbb{R}^3, \\ G_\Omega \quad \mathcal{R}-\text{periodic}, \\ \displaystyle\int_\Omega G_\Omega = 0, \end{array}\right. \quad \text{and} \quad \left\{\begin{array}{l} -\Delta V_{\rm coul}(\rho') = 4\pi\left(\rho' - \frac{1}{|\Omega|}\displaystyle\int_\Omega \rho'\right) \text{ in } \mathbb{R}^3, \\ V_{\rm coul}(\rho') \quad \mathcal{R}-\text{periodic}, \\ \displaystyle\int_\Omega V_{\rm coul}(\rho') = 0. \end{array}\right.$$

The pseudopotential, modeling the effect of the nuclei and the core electrons (and some relativistic effects for heavy atoms) consists of two terms: a local component $V_{\rm local}$ (whose associated operator is the multiplication by the $\mathcal{R}$-periodic function $V_{\rm local}$) and a nonlocal component $V_{\rm nl}$ given by

$$V_{\rm nl}\psi = \sum_{j=1}^{J}\left(\int_\Omega \xi_j(\mathbf{r})\psi(\mathbf{r})\,d\mathbf{r}\right)\xi_j,$$

where ξ_j are regular enough $\mathcal{R}$-periodic functions and J is an integer depending on the chemical nature of the ions in the unit cell. The exchange-correlation functional based on a local density approximation is given in this periodic setting with pseudopotentials by

$$E^c_{{\rm xc},\Omega}(\rho_{[\Psi]}) = \int_\Omega e^{\rm LDA}_{\rm xc}(\rho_c(\mathbf{r}) + \rho_{[\Psi]}(\mathbf{r}))\,d\mathbf{r},$$

where $\rho_c \geq 0$ is a nonlinear core correction, and $e^{\rm LDA}_{\rm xc}(\overline{\rho})$ is an approximation of the exchange-correlation energy per unit volume in a homogeneous electron gas with density $\overline{\rho}$.

The ground-state energy is then the solution of the following minimization problem:

$$I_0^{KS} = \inf\left\{\mathcal{E}^{\rm KS}_{0,\Omega}(\Psi),\ \Psi\in\mathcal{M}\right\}. \tag{9.2.4}$$

Under some assumptions on $V_{\rm nl}, V_{\rm local}$, and $E^c_{{\rm xc},\Omega}$ presented in [44] and recalled in Section 9.3.2, (9.2.4) has a minimizer $\Phi^0 = (\phi^0_1,\ldots,\phi^0_N)\in\mathcal{M}$. Noting that the energy is invariant under a unitary transformation of the orbitals, *i.e.*

$$\forall\Psi\in\mathcal{M},\quad \forall U\in\mathcal{U}(N),\quad U\Psi\in\mathcal{M},\quad \rho_{[U\Psi]}=\rho_{[\Psi]}\quad\text{and}\quad \mathcal{E}^{\rm KS}_{0,\Omega}(U\Psi)=\mathcal{E}^{\rm KS}_{0,\Omega}(\Psi), \tag{9.2.5}$$

where $\mathcal{U}(N)$ is the group of orthogonal matrices:

$$\mathcal{U}(N) = \left\{U\in\mathbb{R}^{N\times N}\mid U^TU = 1_N\right\}, \tag{9.2.6}$$

1_N denoting the identity matrix of rank N, any unitary transform of the Kohn–Sham orbitals Φ^0 in the sense of (9.2.5) is also a minimizer of the Kohn–Sham energy, and (9.2.4) has an infinity of minimizers. It is therefore possible to diagonalize the matrix of the Lagrange multipliers in the first-order optimality conditions relative to (9.2.4), and to show the existence of a minimizer (still denoted by Φ^0), such that

$$\forall i=1,\ldots,N,\quad \mathcal{H}\,\phi^0_i = \lambda^0_i\phi^0_i,\qquad\text{and}\qquad \forall i,j=1,\ldots,N,\quad \langle\phi^0_i|\phi^0_j\rangle = \delta_{ij},$$

for some $\lambda^0_1\leq\lambda^0_2\leq\cdots\leq\lambda^0_N$, where the Hamiltonian $\mathcal{H}$ is the self-adjoint operator on $L^2_\#(\Omega)$ with domain $H^2_\#(\Omega)$ defined by

$$\forall u\in H^2_\#(\Omega),\quad \mathcal{H}u = -\frac{1}{2}\Delta u + V_{\rm ion}u + V_{\rm coul}(\rho^0)u + V_{\rm xc}(\rho^0)u,$$

with $\rho^0 = \rho_{[\Phi^0]}$, $V_{\rm ion} = V_{\rm local} + V_{\rm nl}$, and where

$$V_{\rm xc}(\rho)(\mathbf{r}) = \frac{de_{\rm xc}^{\rm LDA}}{d\rho}(\rho_c(\mathbf{r}) + \rho(\mathbf{r})).$$

Let us also define the Kohn–Sham operator for a given density ρ as

$$\mathcal{H}_{[\rho]} = -\frac{1}{2}\Delta + V_{\rm ion} + V_{\rm coul}(\rho) + V_{\rm xc}(\rho), \tag{9.2.7}$$

so that $\mathcal{H} = \mathcal{H}_{[\rho^0]}$. The potentials $V_{\rm local}$, $V_{\rm coul}(\rho)$, and $V_{\rm xc}(\rho)$ being multiplicative, we use the same notations for the potentials as functions on Ω, and for the corresponding multiplicative operators.

We will suppose in the following that the system under consideration satisfies the *Aufbau* principle, so that $\lambda_1^0 \leq \lambda_2^0 \leq \cdots \leq \lambda_N^0$ are the lowest N eigenvalues of the Kohn–Sham Hamiltonian $\mathcal{H}^0$. Note that, although this property seems to hold in practice for most systems, it has not been proved in general, except for the extended Kohn–Sham model (see [44] for details).

Also, as in the linear case [48], we will make the following assumption:

Assumption 9.2.1. *There is a gap between the $N^{\rm th}$ and the $(N+1)^{\rm st}$ eigenvalues of $\mathcal{H}$, i.e.*

$$g := \lambda_{N+1}^0 - \lambda_N^0 > 0.$$

In this setting, the Fermi level $\epsilon_{\rm F}$ could be defined as any real number in the range $(\lambda_N^0, \lambda_{N+1}^0)$. We define it as $\epsilon_{\rm F} := \frac{\lambda_N^0 + \lambda_{N+1}^0}{2}$.

The purpose of this problem is to compute two quantities of interest:

1. the ground-state density matrix γ_0 based on the orbitals $\Phi^0 = (\phi_1^0, \ldots, \phi_N^0)^T$, defined as

$$\boxed{\gamma_0 := \mathbb{1}_{(-\infty, \epsilon_{\rm F}]}(\mathcal{H}^0) = \sum_{i=1}^{N} |\phi_i^0\rangle\langle\phi_i^0|,} \tag{9.2.8}$$

   which belongs to the Grassmann manifold

$$\Upsilon = \left\{\gamma \in \mathcal{L}(L^2_\#) \;\middle|\; \gamma^* = \gamma,\ \gamma^2 = \gamma,\ {\rm Tr}\,(\gamma) = N,\ {\rm Tr}\,(-\Delta\gamma) < \infty\right\};$$

2. the ground-state energy defined as

$$\boxed{\mathcal{E}_0 := \mathcal{E}^{\rm KS}_{0,\Omega}(\Phi^0).}$$

We refer to [48] for the definition of the operator trace ${\rm Tr}\,$.

9.2.2 Functional setting

In this article, the functional setting is similar to [48]. We denote by $\|\cdot\|$ the operator norm on $\mathcal{L}(L^2_\#)$, the space of bounded linear operators on $L^2_\#(\Omega)$. We also denote by $\mathfrak{S}_1(L^2_\#)$ the Banach space of trace-class operators on $L^2_\#(\Omega)$ endowed with the norm defined by $\|A\|_{\mathfrak{S}_1(L^2_\#)} :=$

$\mathrm{Tr}\,(|A|) = \mathrm{Tr}\,(\sqrt{A^*A})$. Also, let the Hilbert space of Hilbert–Schmidt operators $\mathfrak{S}_2(L^2_\#)$ on $L^2_\#(\Omega)$ be endowed with the inner product defined by $(A,B)_{\mathfrak{S}_2(L^2_\#)} := \mathrm{Tr}\,(A^*B)$. Moreover, let us define for any operator A on $L^2_\#(\Omega)$ with domain $D(A)$,

$$\forall \Psi \in [D(A)]^N, \quad \|A\Psi\|_{L^2_\#} := \left(\sum_{i=1}^N \|A\psi_i\|^2_{L^2_\#}\right)^{1/2},$$

which corresponds to $\|\Psi\|_{L^2_\#}$ when A is the identity operator and $\|\Psi\|_{H^1_\#}$ when $A = (1-\Delta)^{1/2}$.

9.3 Discretization and resolution of the Kohn–Sham model

9.3.1 Planewave discretization

In the context of periodic boundary conditions, we discretize the Kohn–Sham problem (9.2.4) in Fourier modes, also called planewaves. We denote by $\mathcal{R}^* = \frac{2\pi}{L}\mathbb{Z}^3$ the dual lattice of the periodic lattice $\mathcal{R} = L\mathbb{Z}^3$. For $\mathbf{k} \in \mathcal{R}^*$, we denote by $e_\mathbf{k}$ the planewave with wavevector $\mathbf{k}$ and kinetic energy $\frac{1}{2}|\mathbf{k}|^2$, with $|\cdot|$ the Euclidean norm, defined by

$$\begin{aligned} e_\mathbf{k}: \quad \mathbb{R}^3 &\to \mathbb{C} \\ \mathbf{x} &\mapsto |\Omega|^{-1/2} e^{i\mathbf{k}\cdot\mathbf{x}}, \end{aligned}$$

where $|\Omega| = L^3$. The family $(e_\mathbf{k})_{\mathbf{k}\in\mathcal{R}^*}$ forms an orthonormal basis of $L^2_\#(\Omega,\mathbb{C})$ endowed with the scalar product

$$\forall u, v \in L^2_\#(\Omega,\mathbb{C}), \quad \langle u|v\rangle = \int_\Omega \overline{u(\mathbf{r})}\, v(\mathbf{r})\, d\mathbf{r},$$

where $\overline{u(\mathbf{r})}$ denotes the complex conjugate of $u(\mathbf{r})$, and for all $v \in L^2_\#(\Omega,\mathbb{C})$,

$$v(\mathbf{r}) = \sum_{\mathbf{k}\in\mathcal{R}^*} \widehat{v}_\mathbf{k}\, e_\mathbf{k}(\mathbf{r}) \qquad \text{with} \qquad \widehat{v}_\mathbf{k} = \langle e_\mathbf{k}|v\rangle = |\Omega|^{-1/2}\int_\Omega v(\mathbf{r}) e^{-i\mathbf{k}\cdot\mathbf{r}}\, d\mathbf{r}.$$

To discretize the variational set $\mathcal{M}$, we introduce some energy cutoff $E_\mathrm{c} > 0$ and consider all basis functions with kinetic energy smaller than E_c, *i.e.* $|\mathbf{k}| \le \sqrt{2E_\mathrm{c}}$. That is, for each cutoff E_c, we set $N_\mathrm{c} = \sqrt{\frac{E_\mathrm{c}}{2}}\frac{L}{\pi}$ and consider the finite-dimensional discretization space

$$X_{N_\mathrm{c}} := \left\{ \sum_{\mathbf{k}\in\mathcal{R}^*, |\mathbf{k}|\le \frac{2\pi}{L}N_\mathrm{c}} \widehat{v}_\mathbf{k}\, e_\mathbf{k} \;\middle|\; \forall \mathbf{k}, \quad \widehat{v}_{-\mathbf{k}} = \widehat{v}^*_\mathbf{k} \right\} \subset \bigcap_{s\in\mathbb{R}} H^s_\#(\Omega).$$

We denote by Π_{N_c} the orthogonal projector on X_{N_c} for any $H^s_\#(\Omega)$, $s \in \mathbb{R}$, defined as

$$\Pi_{N_\mathrm{c}} v = \sum_{\mathbf{k}\in\mathcal{R}^*, |\mathbf{k}|\le \frac{2\pi}{L}N_\mathrm{c}} \widehat{v}_\mathbf{k} e_\mathbf{k},$$

and by $\Pi^\perp_{N_\mathrm{c}} = (1 - \Pi_{N_\mathrm{c}})$ the orthogonal projector on $X^\perp_{N_\mathrm{c}}$, the orthogonal complement to X_{N_c}. Finally, the variational approximation to the ground-state energy in X_{N_c} is defined as

$$I^{\mathrm{KS}}_{0,N_\mathrm{c}} = \inf\left\{\mathcal{E}^{\mathrm{KS}}_0(\Psi_{N_\mathrm{c}})\ ,\ \Psi_{N_\mathrm{c}} \in \mathcal{M} \cap [X_{N_\mathrm{c}}]^N\right\}. \tag{9.3.1}$$

Using again the invariance property (9.2.5), the Euler equations of this minimization problem can be diagonalized and reduced to find the pairs $(\phi_{j,N_c}, \lambda_{j,N_c})_{j=1,\dots,N}$ satisfying

$$\forall j = 1, \dots, N, \quad \mathcal{H}_{N_c,\text{proj}}\, \phi_{j,N_c} = \lambda_{j,N_c}\, \phi_{j,N_c}, \quad \text{and} \quad \forall i, j = 1, \dots, N, \quad \langle \phi_{i,N_c} | \phi_{j,N_c} \rangle = \delta_{ij}, \tag{9.3.2}$$

$\lambda_{1,N_c} \leq \lambda_{2,N_c} \leq \ldots \leq \lambda_{N,N_c}$, where $\mathcal{H}_{N_c,\text{proj}} : X_{N_c} \to X_{N_c}$ is defined as

$$\mathcal{H}_{N_c,\text{proj}} = \Pi_{N_c} \mathcal{H}_{[\rho_{N_c}]} \Pi_{N_c} = -\frac{1}{2} \Pi_{N_c} \Delta \Pi_{N_c} + \Pi_{N_c} \Big[V_{\text{ion}} + V_{\text{coul}}(\rho_{N_c}) + V_{\text{xc}}(\rho_{N_c}) \Big] \Pi_{N_c}, \tag{9.3.3}$$

with $\rho_{N_c} = \rho_{[\Phi_{N_c}]}$, $\Phi_{N_c} = (\phi_{1,N_c}, \dots, \phi_{N,N_c})^T$ and where $\mathcal{H}_{[\rho_{N_c}]}$ is defined by (9.2.7) for the approximate ground-state density ρ_{N_c}. The corresponding density matrix, which is independent of the chosen basis for $\text{Span}(\phi_{1,N_c}, \phi_{2,N_c}, \dots, \phi_{N,N_c})$, is denoted by $\gamma_{0,N_c} \in \Upsilon$, and defined as

$$\boxed{\gamma_{0,N_c} = \sum_{i=1}^{N} |\phi_{i,N_c}\rangle\langle\phi_{i,N_c}|.}$$

Finally, the ground-state energy is defined as

$$\boxed{I^{\text{KS}}_{0,N_c} = \mathcal{E}^{\text{KS}}_0(\Phi_{N_c}).} \tag{9.3.4}$$

In order to solve the nonlinear eigenvalue problem (9.3.2), a Self-Consistent Field (SCF) procedure is employed [252]. It consists of solving a linear eigenvalue problem at each step, at which the Hamiltonian is computed from the density found at the previous step. The details of the algorithm in this setting can be found in [50] and the references therein.

9.3.2 Smoothness assumptions and *a priori* results

Smoothness assumptions

In order to guarantee the existence of minimizers of problem (9.2.4) and to study the convergence of the solutions to the discretized problem (9.3.1) to those of the continuous problem (9.2.4), some assumptions are needed. They are described in [44] but we recall here the main hypotheses, under which the proofs of Theorem 9.5.1 will hold.

We assume for the local potential V_{local} that,

$$\exists m > 3,\ C \geq 0 \text{ s.t. } \forall \mathbf{k} \in \mathcal{R}^*,\ |(\widehat{V_{\text{local}}})_{\mathbf{k}}| \leq C|\mathbf{k}|^{-m}, \tag{9.3.5}$$

and for the functions defining the non-local potential V_{nl} that

$$\forall 1 \leq j \leq J, \quad \forall \epsilon > 0, \quad \xi_j \in H^{m-3/2-\epsilon}_{\#}(\Omega). \tag{9.3.6}$$

For example, Troullier-Martins pseudopotentials [235] have Fourier coefficients $(\widehat{V_{\text{local}}})_{\mathbf{k}}$ decreasing as $|\mathbf{k}|^{-m}$ with $m = 5$. Moreover, for the exchange-correlation function, we assume that

$$\text{the function } \ \rho \mapsto e^{\text{LDA}}_{\text{xc}}(\rho) \text{ is in } C^1([0, +\infty)) \cap C^3((0, +\infty)), \tag{9.3.7}$$

$$e^{\text{LDA}}_{\text{xc}}(0) = 0, \quad \frac{de^{\text{LDA}}_{\text{xc}}}{d\rho}(0) = 0, \tag{9.3.8}$$

and there exists $0 < \alpha < 1$ and $C \in \mathbb{R}^+$ such that

$$\forall \rho \in \mathbb{R}^+ \backslash \{0\}, \quad \left| \frac{d^2 e_{\rm xc}^{\rm LDA}}{d\rho^2}(\rho) \right| + \left| \rho \frac{d^3 e_{\rm xc}^{\rm LDA}}{d\rho^3}(\rho) \right| \leq C(1 + \rho^{\alpha - 1}). \tag{9.3.9}$$

The assumptions (9.3.7), (9.3.8) and (9.3.9) are for example satisfied by the Xα exchange-correlation functional with $\alpha = 1/3$ ($e_{\rm xc}^{\rm X\alpha}(\rho) = -C_X \rho^{4/3}$, where $C_X > 0$ is a given constant). The exact exchange-correlation functional also verifies these assumptions [93]. Also, let us assume that

$$\rho_c \in H_\#^{m-3/2-\epsilon}(\Omega). \tag{9.3.10}$$

Under assumptions (9.3.5)-(9.3.10), (9.2.4) has a minimizer $\Phi^0 = (\phi_1^0, \phi_2^0, \ldots, \phi_N^0)^T$ [44]. For the *a priori* results of [44] to be valid, some additional assumptions on the exchange-correlation function $e_{\rm xc}^{\rm LDA}$ are needed. More precisely, we assume that

$$\begin{aligned} &e_{\rm xc}^{\rm LDA} \in C^{n_m, \alpha_m}((0, +\infty)) \\ &\qquad \text{where} \\ &\qquad\qquad \begin{vmatrix} n_m = [m] + 1 \text{ and } \alpha_m = m - [m] + 1/2 \text{ if } 0 \leq m - [m] \leq 1/2, \\ n_m = [m] + 2 \text{ and } \alpha_m = m - [m] - 1/2 \text{ if } 1/2 < m - [m] \leq 1, \end{vmatrix} \end{aligned} \tag{9.3.11}$$

(where $[m]$ denotes the integer part of m) and

$$e_{\rm xc}^{\rm LDA} \in C^{n_m, \alpha_m}([0, +\infty)) \quad \text{or} \quad \rho_c + \rho^0 > 0 \text{ in } \mathbb{R}^3. \tag{9.3.12}$$

A priori results on the density matrices

In order to use the *a priori* results of [44] in the proofs of our estimates, we first show that similar *a priori* results hold in the density matrix formalism. To start with, let us define the solution to the discrete problem lying in the space

$$\mathcal{M}^{\Phi^0} := \left\{ \Psi \in \mathcal{M} \;\middle|\; \|\Psi - \Phi^0\|_{L^2_\#} = \min_{U \in \mathcal{U}(N)} \|U\Psi - \Phi^0\|_{L^2_\#} \right\},$$

where $\mathcal{U}(N)$ is defined in (9.2.6) and $\mathcal{M}$ in (9.2.2). Therefore we define $\Phi_{N_{\rm c}}^0 \in \mathcal{M}^{\Phi^0}$ such that

$$\|\Phi_{N_{\rm c}}^0 - \Phi^0\|_{L^2_\#} = \min_{U \in U(N)} \|U\Phi_{N_{\rm c}} - \Phi^0\|_{L^2_\#}, \tag{9.3.13}$$

where $\Phi_{N_{\rm c}}$ is a solution to (9.3.2). From the following lemma, we can write the *a priori* results presented in [44, Theorem 4.2] in the density matrix formalism.

Lemma 9.3.1 ($L^2_\#$ and $H^1_\#$ norm equivalences)**.** *There exist $c, C > 0$ such that for all $\Psi^0 = (\psi_1^0, \ldots, \psi_N^0) \in \mathcal{M}^{\Phi^0}$, with corresponding density matrix $\gamma_\Psi^0 = \sum_{i=1}^N |\psi_i^0\rangle\langle\psi_i^0|$,*

$$\|\Psi^0 - \Phi^0\|_{L^2_\#} \quad \leq \quad \|\gamma_\Psi^0 - \gamma_0\|_{\mathfrak{S}_2(L^2_\#)} \quad \leq \quad \sqrt{2}\|\Psi^0 - \Phi^0\|_{L^2_\#}, \tag{9.3.14}$$

$$c\|(1-\Delta)^{1/2}(\Psi^0 - \Phi^0)\|_{L^2_\#} \quad \leq \quad \|(1-\Delta)^{1/2}(\gamma_\Psi^0 - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)} \quad \leq \quad C\|(1-\Delta)^{1/2}(\Psi^0 - \Phi^0)\|_{L^2_\#}. \tag{9.3.15}$$

The proof is given in the Appendix. Note that this lemma is more general than the similar result provided in the linear case [48, Lemma 2.3], as γ^0_Ψ can be any density matrix and not only the discrete density matrix γ_{0,N_c}.

Based on Lemma 9.3.1, it is possible to express the results of [44, Theorem 4.2] in terms of density matrices, and in the following, we will use some of these results stated for $-m+3/2 < s < m+1/2$ in the particular cases where $s=-1,0,1$. Thus, under assumptions (9.3.5)-(9.3.12), there exist $c, C>0$, and $N_\mathrm{c}^0 \in \mathbb{N}$ such that for $N_\mathrm{c} \geq N_\mathrm{c}^0$,

$$c\|(1-\Delta)^{1/2}(\gamma_0-\gamma_{0,N_\mathrm{c}})\|^2_{\mathfrak{S}_2(L^2_\#)} \leq I^{\mathrm{KS}}_{0,N_\mathrm{c}} - I_0^{KS} \leq C\|(1-\Delta)^{1/2}(\gamma_0-\gamma_{0,N_\mathrm{c}})\|^2_{\mathfrak{S}_2(L^2_\#)}, \quad (9.3.16)$$

$$\|(1-\Delta)^{-1/2}(\Phi^0_{N_\mathrm{c}}-\Phi^0)\|_{L^2_\#} \leq CN_\mathrm{c}^{-2}\|(1-\Delta)^{1/2}(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}, \quad (9.3.17)$$

$$\|\gamma_0-\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq CN_\mathrm{c}^{-2}, \quad (9.3.18)$$

$$\|(1-\Delta)^{1/2}(\gamma_0-\gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)} \leq CN_\mathrm{c}^{-1}, \quad (9.3.19)$$

and

$$\|\Lambda^0-\Lambda^0_{N_\mathrm{c}}\|_\mathrm{F} \leq C\|(1-\Delta)^{1/2}(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|^2_{\mathfrak{S}_2(L^2_\#)}, \quad (9.3.20)$$

where $\|\,.\,\|_\mathrm{F}$ denotes the Frobenius norm.

To show that the uncomputable terms in the perturbative development, which will be defined in (9.4.6), can be neglected, some properties on the Coulomb potential, as well as on the total potential and the exchange-correlation functional are needed. From [44, (3.19)], there holds for the Coulomb multiplicative potential

$$\forall s \in \mathbb{R}, \quad \forall \rho_1,\rho_2 \in H^s_\#(\Omega), \quad \|V_\mathrm{coul}(\rho_1)-V_\mathrm{coul}(\rho_2)\|_{H^{s+2}_\#} \leq C\|\rho_1-\rho_2\|_{H^s_\#}, \quad (9.3.21)$$

from which we deduce in particular with $s=0$ using Sobolev embeddings that

$$\|V_\mathrm{coul}(\rho^0)-V_\mathrm{coul}(\rho_{N_\mathrm{c}})\|_{L^\infty_\#} + \|\nabla\left(V_\mathrm{coul}(\rho^0)-V_\mathrm{coul}(\rho_{N_\mathrm{c}})\right)\|_{L^3_\#} \leq C\|\rho^0-\rho_{N_\mathrm{c}}\|_{L^2_\#}, \quad (9.3.22)$$

which is in particular bounded by a constant independent of N_c. Moreover, under Assumptions (9.3.7)-(9.3.12), $V_\mathrm{local}+V_\mathrm{coul}(\rho_{N_\mathrm{c}})+V_\mathrm{xc}(\rho_{N_\mathrm{c}}) \in H^{3/2+\epsilon}_\#(\Omega)$, for $\epsilon>0$, therefore there exist $C \in \mathbb{R}^+$ and $N_\mathrm{c}^0 \in \mathbb{N}$ such that for $N_\mathrm{c} \geq N_\mathrm{c}^0$,

$$\|V_\mathrm{local}+V_\mathrm{coul}(\rho_{N_\mathrm{c}})+V_\mathrm{xc}(\rho_{N_\mathrm{c}})\|_{L^\infty_\#} + \|\nabla\left(V_\mathrm{local}+V_\mathrm{coul}(\rho_{N_\mathrm{c}})+V_\mathrm{xc}(\rho_{N_\mathrm{c}})\right)\|_{L^3_\#} \leq C, \quad (9.3.23)$$

and

$$\|V_\mathrm{xc}(\rho^0)-V_\mathrm{xc}(\rho_{N_\mathrm{c}})\|_{L^\infty_\#} + \|\nabla\left(V_\mathrm{xc}(\rho^0)-V_\mathrm{xc}(\rho_{N_\mathrm{c}})\right)\|_{L^3_\#} \leq C. \quad (9.3.24)$$

Using a Taylor formula with integral remainder, there holds

$$\begin{aligned}\|V_\mathrm{xc}(\rho^0)-V_\mathrm{xc}(\rho_{N_\mathrm{c}})\|_{H^{-1}_\#} &= \left\|\int_0^1 \frac{d^2e^{\mathrm{LDA}}_\mathrm{xc}}{d\rho^2}\left(\rho_c+s\rho^0+(1-s)\rho_{N_\mathrm{c}}\right)(\rho^0-\rho_{N_\mathrm{c}})ds\right\|_{H^{-1}_\#} \\ &\leq \int_0^1 \left\|\frac{d^2e^{\mathrm{LDA}}_\mathrm{xc}}{d\rho^2}\left(\rho_c+s\rho^0+(1-s)\rho_{N_\mathrm{c}}\right)(\rho^0-\rho_{N_\mathrm{c}})\right\|_{H^{-1}_\#} ds.\end{aligned}$$

From [44, (4.25)] and the definition of the density (9.2.3), $\rho_c+s\rho^0+(1-s)\rho_{N_\mathrm{c}}$ is uniformly bounded in $H^\sigma_\#(\Omega)$ for some $\sigma>3/2$ uniformly in N_c and s. As for all $s\in[0,1]$, $\rho_c+s\rho^0+(1-s)\rho_{N_\mathrm{c}}$ is bounded away from zero uniformly in N_c, and from (9.3.9) and (9.3.12), the quantity

$\frac{d^2 e_{\rm xc}^{\rm LDA}}{d\rho^2}\left(\rho_c + s\rho^0 + (1-s)\rho_{N_{\rm c}}\right)$ is also uniformly bounded in $H^\sigma_\#(\Omega)$. Note that for all $\sigma > 3/2$ and for all $0 \le r \le \sigma$, $H^\sigma_\#(\Omega)$ is an ideal of $H^r_\#(\Omega)$, which implies, in particular that for all $r \ge 0$ and all $\sigma > 3/2$,

$$\forall f \in H^{\max(r,\sigma)}_\#(\Omega), \quad \forall g \in H^{-r}_\#(\Omega), \quad fg \in H^{-r}_\#(\Omega),$$

and

$$\|fg\|_{H^{-r}_\#(\Omega)} \le C_{r,\sigma}\|f\|_{H^{\max(r,\sigma)}_\#(\Omega)}\|g\|_{H^{-r}_\#(\Omega)}, \tag{9.3.25}$$

for some constant $C_{r,\sigma} \ge 0$ independent of f and g. From this, we deduce that there exist $C > 0$ and $N_{\rm c}^0 \in \mathbb{N}$ such that for $N_{\rm c} > N_{\rm c}^0$,

$$\|V_{\rm xc}(\rho^0) - V_{\rm xc}(\rho_{N_{\rm c}})\|_{H^{-1}_\#} \le C\|\rho^0 - \rho_{N_{\rm c}}\|_{H^{-1}_\#}. \tag{9.3.26}$$

9.4 Post-processing of the planewave approximation

Our post-processing method strongly relies on the fact that the Laplace operator is diagonal in planewaves, so that its eigenvalues are explicitly known. Indeed, the smallest eigenvalue on $X^\perp_{N_{\rm c}}$ being strictly larger than $4\left(\frac{\pi N_{\rm c}}{L}\right)^2$, if the $N^{\rm th}$ eigenvalue of the operator $\mathcal{H}_{N_{\rm c},\rm proj}$ defined in (9.3.3) verifies $\lambda_{N,N_{\rm c}} < 2\left(\frac{\pi N_{\rm c}}{L}\right)^2$, which holds for $N_{\rm c}$ large enough, the discrete solution $\Phi_{N_{\rm c}}$ is also the ground-state of the following Kohn-Sham problem

$$\forall j = 1,\dots,N, \quad \mathcal{H}_{N_{\rm c}}\,\phi_{j,N_{\rm c}} = \lambda_{j,N_{\rm c}}\,\phi_{j,N_{\rm c}}, \qquad \text{and} \qquad \forall i,j = 1,\dots,N, \quad \langle\phi_{i,N_{\rm c}}|\phi_{j,N_{\rm c}}\rangle = \delta_{ij}, \tag{9.4.1}$$

$\lambda_{1,N_{\rm c}} \le \lambda_{2,N_{\rm c}} \le \ldots \le \lambda_{N,N_{\rm c}}$, where

$$\mathcal{H}_{N_{\rm c}} = -\frac{1}{2}\Delta + \Pi_{N_{\rm c}}\Big[V_{\rm ion} + V_{\rm coul}(\rho_{N_{\rm c}}) + V_{\rm xc}(\rho_{N_{\rm c}})\Big]\Pi_{N_{\rm c}}.$$

Replacing $\mathcal{H}_{N_{\rm c},\rm proj}$ by $\mathcal{H}_{N_{\rm c}}$ in the equations satisfied by $\Phi_{N_{\rm c}}$ will be crucial in our analysis. Conversely, the exact solution $(\phi_j^0, \lambda_j^0)_{j=1,\dots,N}$ satisfies

$$(\mathcal{H}_{N_{\rm c}} + \mathcal{V}^\perp_{N_{\rm c}} + \mathcal{W}_{N_{\rm c}})\,\phi_j^0 = \lambda_j^0\,\phi_j^0, \qquad \int_\Omega \phi_i^0\,\phi_j^0 = \delta_{ij}, \tag{9.4.2}$$

where

$$\mathcal{V}^\perp_{N_{\rm c}} = [V_{\rm ion} + V_{\rm coul}(\rho_{N_{\rm c}}) + V_{\rm xc}(\rho_{N_{\rm c}})] - \Pi_{N_{\rm c}}\Big[V_{\rm ion} + V_{\rm coul}(\rho_{N_{\rm c}}) + V_{\rm xc}(\rho_{N_{\rm c}})\Big]\Pi_{N_{\rm c}}, \tag{9.4.3}$$

and

$$\mathcal{W}_{N_{\rm c}} = [V_{\rm coul}(\rho^0) + V_{\rm xc}(\rho^0)] - [V_{\rm coul}(\rho_{N_{\rm c}}) + V_{\rm xc}(\rho_{N_{\rm c}})]. \tag{9.4.4}$$

As described in [50, Section 4], we rely on Rayleigh–Schrödinger perturbation method [146] to define improved orbitals $(\widetilde{\phi_{j,N_{\rm c}}}, \widetilde{\lambda_{j,N_{\rm c}}})_{j=1,\dots,N}$, taking (9.4.2) as the perturbed equation, and (9.4.1) as the unperturbed one. For non-degenerate eigenvalues, the corrections arising from first-order perturbation theory are

$$\forall\, j = 1,\dots,N, \quad \phi_j^0 \simeq \phi^0_{j,N_{\rm c}} + \phi^{(1,1)}_{j,N_{\rm c}} + \phi^{(2)}_{j,N_{\rm c}},$$

where

$$\phi^{(1,1)}_{j,N_{\rm c}} = -\left(-\frac{1}{2}\Delta - \lambda_{j,N_{\rm c}}\right)^{-1} r_j \in X^\perp_{N_{\rm c}}, \tag{9.4.5}$$

is computable, $r_j \in X_{N_\mathrm{c}}^\perp$ being the residual

$$\begin{aligned} r_j &= \left(-\frac{1}{2}\Delta + V_{\mathrm{ion}} + V_{\mathrm{coul}}(\rho_{N_\mathrm{c}}) + V_{\mathrm{xc}}(\rho_{N_\mathrm{c}}) - \lambda_{j,N_\mathrm{c}}\right)\phi_{j,N_\mathrm{c}} \\ &= \left(\mathcal{H}_{N_\mathrm{c}} + \mathcal{V}_{N_\mathrm{c}}^\perp - \lambda_{j,N_\mathrm{c}}\right)\phi_{j,N_\mathrm{c}} = \mathcal{V}_{N_\mathrm{c}}^\perp \phi_{j,N_\mathrm{c}}, \end{aligned}$$

and $\phi_{j,N_\mathrm{c}}^{(2)}$ defined by

$$\phi_{j,N_\mathrm{c}}^{(2)} = -\left(\mathcal{H}_{N_\mathrm{c}} - \lambda_{j,N_\mathrm{c}}\right)^{-1}_{|_{(\phi_{j,N_\mathrm{c}})^\perp}} \mathcal{W}_{N_\mathrm{c}}\phi_{j,N_\mathrm{c}}, \tag{9.4.6}$$

is not computable as $\mathcal{W}_{N_\mathrm{c}}$ depends on the exact density ρ^0. Note that the definition of $\phi_{j,N_\mathrm{c}}^{(1,1)}$ is only consistent because the residuals r_j belong to $X_{N_\mathrm{c}}^\perp$ for all $j = 1, \ldots, N$.

Compared to the linear case [48], the main difference is the presence of the uncomputable potential $\mathcal{W}_{N_\mathrm{c}}$, which leads at first order to uncomputable corrections (9.4.6). However, we will derive that these uncomputable terms are *a priori* small, and define the post-processed orbitals only from the computable corrections defined in (9.4.5), which are also well-defined for degenerate eigenvalues. We therefore define the perturbed orbitals, as well as density matrix and energy as follows.

Definition 9.4.1 (Perturbed eigenvectors, density matrix and energy)**.** *For all $j = 1, \ldots, N$, we define the perturbed eigenvectors as*

$$\widetilde{\phi_{j,N_\mathrm{c}}} = \phi_{j,N_\mathrm{c}} + \phi_{j,N_\mathrm{c}}^{(1,1)}. \tag{9.4.7}$$

We also define orthonormal perturbed eigenvectors as an orthonormalization of $(\widetilde{\phi_{j,N_\mathrm{c}}})_{j=1,\ldots,N}$. More precisely, for all $j = 1, \ldots, N$, define

$$\widetilde{\widetilde{\Phi_{N_\mathrm{c}}}} = {S_{N_\mathrm{c}}}^{-1/2}\widetilde{\Phi_{N_\mathrm{c}}}, \tag{9.4.8}$$

where S_{N_c}, the $N \times N$ overlap matrix of $\widetilde{\Phi_{N_\mathrm{c}}} = (\widetilde{\phi_{1,N_\mathrm{c}}}, \ldots, \widetilde{\phi_{N,N_\mathrm{c}}})$, is defined as

$$\forall i, j = 1, \ldots, N, \quad (S_{N_\mathrm{c}})_{i,j} = \langle\widetilde{\phi_{i,N_\mathrm{c}}}|\widetilde{\phi_{j,N_\mathrm{c}}}\rangle. \tag{9.4.9}$$

We define the perturbed density matrix as

$$\boxed{\widetilde{\gamma_{N_\mathrm{c}}} = \sum_{i=1}^N |\widetilde{\phi_{j,N_\mathrm{c}}}\rangle\langle\widetilde{\phi_{j,N_\mathrm{c}}}| = \gamma_{0,N_\mathrm{c}} + \gamma_{N_\mathrm{c}}^{(1)} + \sum_{i=1}^N |\phi_{j,N_\mathrm{c}}^{(1,1)}\rangle\langle\phi_{j,N_\mathrm{c}}^{(1,1)}|,} \tag{9.4.10}$$

where

$$\gamma_{N_\mathrm{c}}^{(1)} = \sum_{i=1}^N |\phi_{j,N_\mathrm{c}}^{(1,1)}\rangle\langle\phi_{j,N_\mathrm{c}}| + \sum_{i=1}^N |\phi_{j,N_\mathrm{c}}\rangle\langle\phi_{j,N_\mathrm{c}}^{(1,1)}|. \tag{9.4.11}$$

We also define an orthonormalized perturbed density matrix as

$$\boxed{\widetilde{\widetilde{\gamma_{N_\mathrm{c}}}} = \sum_{i=1}^N |\widetilde{\widetilde{\phi_{i,N_\mathrm{c}}}}\rangle\langle\widetilde{\widetilde{\phi_{i,N_\mathrm{c}}}}|.} \tag{9.4.12}$$

We define the perturbed energy as the energy of the perturbed eigenvectors computed with (9.2.1)

$$\boxed{\widetilde{\mathcal{E}_{0,N_\mathrm{c}}} = \mathcal{E}_{0,\Omega}^{\mathrm{KS}}(\widetilde{\Phi_{N_\mathrm{c}}}),}$$

and the orthonormalized perturbed energy as

$$\boxed{\widetilde{\widetilde{\mathcal{E}_{N_\mathrm{c}}}} = \mathcal{E}_{0,\Omega}^{\mathrm{KS}}(\widetilde{\widetilde{\Phi_{N_\mathrm{c}}}}).} \tag{9.4.13}$$

Remark 9.4.2. *Since the post-processed orbitals* (9.4.7) *are not orthonormal,* $\widetilde{\gamma_{N_\mathrm{c}}} \in \Upsilon$ *does not hold in general, although* $\widetilde{\gamma_{N_\mathrm{c}}} = \widetilde{\gamma_{N_\mathrm{c}}}^*$. *Indeed, a priori,* $\widetilde{\gamma_{N_\mathrm{c}}}^2 \neq \widetilde{\gamma_{N_\mathrm{c}}}$ *and* $\mathrm{Tr}\,(\widetilde{\gamma_{N_\mathrm{c}}}) \neq N$. *On the other hand, the post-processed orbitals* (9.4.8) *being orthonormal, there holds* $\widetilde{\widetilde{\gamma_{N_\mathrm{c}}}} \in \Upsilon$.

Remark 9.4.3. *Note that the computational cost of the corrections is limited, and similar to the linear case [48]. For all* $j = 1, \ldots, N$, *the operator* $\left(-\frac{1}{2}\Delta - \lambda_{j,N_\mathrm{c}}\right)$ *is diagonal in a planewave basis, hence trivial to invert. Each residual* $\mathcal{V}_{N_\mathrm{c}}^{\perp}\phi_{j,N_\mathrm{c}}$ *can be computed with only two FFT's. On top of that, to compute the orthonormalized density matrix* (9.4.12) *and energy* (9.4.13), *one needs to orthonormalize the post-processed orbitals. We refer to [50, Section 5] for numerical results illustrating the low computational cost of the post-processing.*

9.5 Convergence improvement on the density matrix and the energy

9.5.1 Theorem

The improvement results on the post-processed density matrices and the energies are collected in the following theorem. Compared to the linear case [48], the results are similar except for the post-processed energy (9.4.13) based on the orthonormal version of the post-processed density matrix, for which we derive a convergence doubling compared to the density matrix improvement factor of N_c^{-2}.

Theorem 9.5.1 (Improved convergence for the density matrix and the energy)**.** *There exist* $C \in \mathbb{R}^+$ *and* $N_\mathrm{c}^0 \in \mathbb{N}$ *such that for* $N_\mathrm{c} \geq N_\mathrm{c}^0$,

$$\boxed{\|(1-\Delta)^{1/2}(\gamma_0 - \widetilde{\gamma_{N_\mathrm{c}}})\|_{\mathfrak{S}_2(L^2_\#)} \leq C N_\mathrm{c}^{-2} \|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)},} \tag{9.5.1}$$

$$\boxed{\|(1-\Delta)^{1/2}(\gamma_0 - \widetilde{\widetilde{\gamma_{N_\mathrm{c}}}})\|_{\mathfrak{S}_2(L^2_\#)} \leq C N_\mathrm{c}^{-2} \|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)},} \tag{9.5.2}$$

$$\boxed{|\widetilde{\mathcal{E}_{0,N_\mathrm{c}}} - \mathcal{E}_0| \leq C N_\mathrm{c}^{-2} |\mathcal{E}_{0,N_\mathrm{c}} - \mathcal{E}_0|,} \tag{9.5.3}$$

and

$$\boxed{|\widetilde{\widetilde{\mathcal{E}_{N_\mathrm{c}}}} - \mathcal{E}_0| \leq C N_\mathrm{c}^{-4} |\mathcal{E}_{0,N_\mathrm{c}} - \mathcal{E}_0|.} \tag{9.5.4}$$

9.5.2 Proof

In order to prove Theorem 9.5.1, we first provide in Section 9.5.2 a decomposition of γ_0 based on spectral projection in Lemma 9.5.2, and then, in Section 9.5.2, we provide three preliminary lemmas. In Section 9.5.2, we decompose the difference $\gamma_0 - \widetilde{\gamma_{N_\mathrm{c}}}$ into six parts in Lemma 9.5.6, and we then estimate each of these terms in the following lemmas 9.5.7, 9.5.8, 9.5.9, 9.5.10, 9.5.11, and 9.5.12 in order to prove estimate (9.5.1). Lemma 9.5.13 then allows to extend the proof to estimate (9.5.2). Finally, in Section 9.5.2, we provide a proof for estimates (9.5.3) and (9.5.4).

Exact density matrix in terms of approximate density matrix

From [48, Lemma 4.2], whose proof is identical in the nonlinear case, relying on the gap assumption 9.2.1, there exists a contour Γ and $N_{\mathrm{c}}^0 \in \mathbb{N}$, such that for all $N_{\mathrm{c}} \geq N_{\mathrm{c}}^0$, Γ contains the lowest N eigenvalues of both operators $\mathcal{H}$ and $\mathcal{H}_{N_{\mathrm{c}}}$ and none of the higher ones. Taking such a contour Γ, writing $\mathcal{H} = \mathcal{H}_{N_{\mathrm{c}}} + \mathcal{V}_{N_{\mathrm{c}}}^{\perp} + \mathcal{W}_{N_{\mathrm{c}}}$, and using the definition of spectral projection, the density matrix defined in (9.2.8) can be decomposed as

$$\gamma_0 \quad = \quad \frac{1}{2\pi i} \oint_{\Gamma} (z - \mathcal{H})^{-1} \, dz = \frac{1}{2\pi i} \oint_{\Gamma} \left(z - \mathcal{H}_{N_{\mathrm{c}}} - \mathcal{V}_{N_{\mathrm{c}}}^{\perp} - \mathcal{W}_{N_{\mathrm{c}}} \right)^{-1} dz.$$

Then using the Dyson equation twice [100, 231] to decompose the operator $\left(z - \mathcal{H}_{N_{\mathrm{c}}} - \mathcal{V}_{N_{\mathrm{c}}}^{\perp} - \mathcal{W}_{N_{\mathrm{c}}} \right)^{-1}$, we obtain

$$\begin{aligned} \gamma_0 = \frac{1}{2\pi i} \oint_{\Gamma} (z - \mathcal{H}_{N_{\mathrm{c}}})^{-1} \, dz + \frac{1}{2\pi i} \oint_{\Gamma} (z - \mathcal{H}_{N_{\mathrm{c}}})^{-1} \left(\mathcal{V}_{N_{\mathrm{c}}}^{\perp} + \mathcal{W}_{N_{\mathrm{c}}} \right) (z - \mathcal{H}_{N_{\mathrm{c}}})^{-1} \, dz \\ + \frac{1}{2\pi i} \oint_{\Gamma} (z - \mathcal{H})^{-1} (\mathcal{V}_{N_{\mathrm{c}}}^{\perp} + \mathcal{W}_{N_{\mathrm{c}}}) \, (z - \mathcal{H}_{N_{\mathrm{c}}})^{-1} \left(\mathcal{V}_{N_{\mathrm{c}}}^{\perp} + \mathcal{W}_{N_{\mathrm{c}}} \right) (z - \mathcal{H}_{N_{\mathrm{c}}})^{-1} \, dz. \end{aligned} \tag{9.5.5}$$

Lemma 9.5.2 (Decomposition of γ_0)**.** *There holds*

$$\gamma_0 = \gamma_{0,N_{\mathrm{c}}} + \gamma_{N_{\mathrm{c}}}^{(1)} + \gamma_{N_{\mathrm{c}}}^{(2)} + \widetilde{Q_{N_{\mathrm{c}}}}, \tag{9.5.6}$$

where $\gamma_{N_{\mathrm{c}}}^{(1)}$ *is defined in* (9.4.11)*,*

$$\gamma_{N_{\mathrm{c}}}^{(2)} = \frac{1}{2\pi i} \oint_{\Gamma} (z - \mathcal{H}_{N_{\mathrm{c}}})^{-1} \, \mathcal{W}_{N_{\mathrm{c}}} \, (z - \mathcal{H}_{N_{\mathrm{c}}})^{-1} \, dz, \tag{9.5.7}$$

and

$$\widetilde{Q_{N_{\mathrm{c}}}} = \frac{1}{2\pi i} \oint_{\Gamma} (z - \mathcal{H})^{-1} (\mathcal{V}_{N_{\mathrm{c}}}^{\perp} + \mathcal{W}_{N_{\mathrm{c}}}) \, (z - \mathcal{H}_{N_{\mathrm{c}}})^{-1} \left(\mathcal{V}_{N_{\mathrm{c}}}^{\perp} + \mathcal{W}_{N_{\mathrm{c}}} \right) (z - \mathcal{H}_{N_{\mathrm{c}}})^{-1} \, dz. \tag{9.5.8}$$

Proof. We start from (9.5.5). By definition of the spectral projection, there holds

$$\gamma_{0,N_{\mathrm{c}}} = \frac{1}{2\pi i} \oint_{\Gamma} (z - \mathcal{H}_{N_{\mathrm{c}}})^{-1} \, dz,$$

i.e. the first term of the right hand side of (9.5.5). Following the proof of [48, Lemma 4.3], one can show that

$$\gamma_{N_{\mathrm{c}}}^{(1)} = \frac{1}{2\pi i} \oint_{\Gamma} (z - \mathcal{H}_{N_{\mathrm{c}}})^{-1} \, \mathcal{V}_{N_{\mathrm{c}}}^{\perp} \, (z - \mathcal{H}_{N_{\mathrm{c}}})^{-1} \, .$$

From the definition of $\gamma_{N_{\mathrm{c}}}^{(2)}$ in (9.5.7), we get that $\gamma_{N_{\mathrm{c}}}^{(1)} + \gamma_{N_{\mathrm{c}}}^{(2)}$ corresponds to the second term of the right hand side in (9.5.5). Finally, the definition of $\widetilde{Q_{N_{\mathrm{c}}}}$ in (9.5.8) allows to conclude the proof of the lemma. □

Preliminary lemmas

Lemma 9.5.3 (Estimation of $(1 - \Delta)^{-1/2} \mathcal{W}_{N_{\mathrm{c}}}$)**.** *There exist* $C \in \mathbb{R}^+$ *and* $N_{\mathrm{c}}^0 \in \mathbb{N}$ *such that for* $N_{\mathrm{c}} \geq N_{\mathrm{c}}^0$*,*

$$\| (1 - \Delta)^{-1/2} \mathcal{W}_{N_{\mathrm{c}}} \gamma_0 \|_{\mathfrak{S}_2(L^2_{\#})} \leq C N_{\mathrm{c}}^{-2} \| (1 - \Delta)^{1/2} (\gamma_0 - \gamma_{0,N_{\mathrm{c}}}) \|_{\mathfrak{S}_2(L^2_{\#})}, \tag{9.5.9}$$

and

$$\| (1 - \Delta)^{-1/2} \mathcal{W}_{N_{\mathrm{c}}} \gamma_{0,N_{\mathrm{c}}} \|_{\mathfrak{S}_2(L^2_{\#})} \leq C N_{\mathrm{c}}^{-2} \| (1 - \Delta)^{1/2} (\gamma_0 - \gamma_{0,N_{\mathrm{c}}}) \|_{\mathfrak{S}_2(L^2_{\#})}. \tag{9.5.10}$$

Proof. By definition of the Hilbert–Schmidt norm,

$$\begin{aligned}\|(1-\Delta)^{-1/2}\mathcal{W}_{N_\mathrm{c}}\gamma_0\|^2_{\mathfrak{S}_2(L^2_\#)} &= \sum_{i=1}^{N}\|(1-\Delta)^{-1/2}\mathcal{W}_{N_\mathrm{c}}\phi_i^0\|^2_{L^2_\#}\\ &= \sum_{i=1}^{N}\|\mathcal{W}_{N_\mathrm{c}}\phi_i^0\|^2_{H^{-1}_\#}.\end{aligned}$$

As for all $1 \le i \le N$, $\phi_i^0 \in H^2_\#(\Omega)$, and for all $N_\mathrm{c} \in \mathbb{N}^*$, $\mathcal{W}_{N_\mathrm{c}} \in H^{-1}_\#(\Omega)$, we show using (9.3.25) that there exists $C \in \mathbb{R}^+$ such that for all $N_\mathrm{c} \in \mathbb{N}^*$,

$$\|\mathcal{W}_{N_\mathrm{c}}\phi_i^0\|_{H^{-1}_\#} \le C\|\mathcal{W}_{N_\mathrm{c}}\|_{H^{-1}_\#}\|\phi_i^0\|_{H^2_\#}.$$

Therefore, as $\sum_{i=1}^N \|\phi_i^0\|^2_{H^2_\#(\Omega)}$ is bounded, there exists $C \in \mathbb{R}^+$ such that for all $N_\mathrm{c} \in \mathbb{N}^*$,

$$\|(1-\Delta)^{-1/2}\mathcal{W}_{N_\mathrm{c}}\gamma_0\|^2_{\mathfrak{S}_2(L^2_\#)} \le C\|\mathcal{W}_{N_\mathrm{c}}\|^2_{H^{-1}_\#}\sum_{i=1}^{N}\|\phi_i^0\|^2_{H^2_\#} \le C\|\mathcal{W}_{N_\mathrm{c}}\|^2_{H^{-1}_\#}.$$

From (9.3.21) with $s = -1$ and (9.3.26), we derive

$$\|(1-\Delta)^{-1/2}\mathcal{W}_{N_\mathrm{c}}\gamma_0\|_{\mathfrak{S}_2(L^2_\#)} \le C\|\rho^0 - \rho_{N_\mathrm{c}}\|_{H^{-1}_\#}. \tag{9.5.11}$$

Moreover, as shown in [44, (4.85)], there holds

$$\|\rho^0 - \rho_{N_\mathrm{c}}\|_{H^{-1}_\#} \le C\|(1-\Delta)^{-1/2}(\Phi^0 - \Phi^0_{N_\mathrm{c}})\|_{L^2_\#}. \tag{9.5.12}$$

Finally, using (9.3.17) in (9.5.12) together with (9.5.11) allows to prove (9.5.9).

Since for all $1 \le i \le N$, ϕ_{i,N_c} is bounded in $H^2_\#(\Omega)$ independently of N_c, the proof can be easily adapted to show (9.5.10). □

Lemma 9.5.4 (Estimation of $(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}$)**.** *There exist $C \in \mathbb{R}^+$ and $N_\mathrm{c}^0 \in \mathbb{N}$ such that for $N_\mathrm{c} \ge N_\mathrm{c}^0$,*

$$\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} \le C\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)}. \tag{9.5.13}$$

Proof. The proof is similar to the proof of estimate (51) in [48, Lemma 4.6]. First,

$$\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} = \left\|(1-\Delta)^{-1/2}(\mathcal{H}_{[\rho_{N_\mathrm{c}}]} - \mathcal{H}_{N_\mathrm{c}})\gamma_{0,N_\mathrm{c}}\right\|_{\mathfrak{S}_2(L^2_\#)}.$$

The difference $(\mathcal{H}_{[\rho_{N_\mathrm{c}}]} - \mathcal{H}_{N_\mathrm{c}})\gamma_{0,N_\mathrm{c}}$ can be decomposed as

$$\begin{aligned}(\mathcal{H}_{[\rho_{N_\mathrm{c}}]} - \mathcal{H}_{N_\mathrm{c}})\gamma_{0,N_\mathrm{c}} &= \mathcal{H}_{[\rho_{N_\mathrm{c}}]}(\gamma_{0,N_\mathrm{c}} - \gamma_0) + \mathcal{H}_{[\rho_{N_\mathrm{c}}]}\gamma_0 - \mathcal{H}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}\\ &= \mathcal{H}_{[\rho_{N_\mathrm{c}}]}(\gamma_{0,N_\mathrm{c}} - \gamma_0) + \mathcal{H}\gamma_0 - \mathcal{H}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}} - \mathcal{W}_{N_\mathrm{c}}\gamma_0.\end{aligned}$$

The Hilbert–Schmidt norms of $(1-\Delta)^{-1/2}\mathcal{H}_{[\rho_{N_\mathrm{c}}]}(\gamma_{0,N_\mathrm{c}} - \gamma_0)$ and $(1-\Delta)^{-1/2}(\mathcal{H}\gamma_0 - \mathcal{H}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}})$ can be estimated exactly as in the proof of [48, Lemma 4.6] using in particular the *a priori* estimate (9.3.20). Moreover, $(1-\Delta)^{-1/2}\mathcal{W}_{N_\mathrm{c}}\gamma_0$ can be estimated using (9.5.9). This concludes the proof of the lemma. □

Lemma 9.5.5 (Estimation of $\|S_{N_\mathrm{c}} - 1_N\|_\mathrm{F}$.)**.** *There exist $C \in \mathbb{R}^+$ and $N_\mathrm{c}^0 \in \mathbb{N}$ such that for $N_\mathrm{c} \geq N_\mathrm{c}^0$,*

$$\|S_{N_\mathrm{c}} - 1_N\|_\mathrm{F} \leq C N_\mathrm{c}^{-2} \|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|^2_{\mathfrak{S}_2(L^2_\#)}. \tag{9.5.14}$$

Proof. Given the definition of S_{N_c} in (9.4.9), noting that for all $1 \leq i \leq N$, $\phi^{(1,1)}_{i,N_\mathrm{c}} \in X^\perp_{N_\mathrm{c}}$, and for all $1 \leq i,j \leq N$, $\langle \phi_{i,N_\mathrm{c}} | \phi_{j,N_\mathrm{c}} \rangle = \delta_{ij}$, and using the Cauchy–Schwarz inequality, there holds

$$\begin{aligned}
\|S_{N_\mathrm{c}} - 1_N\|^2_\mathrm{F} &= \sum_{i,j=1}^N |\langle \widetilde{\phi_{i,N_\mathrm{c}}} | \widetilde{\phi_{j,N_\mathrm{c}}} \rangle - \delta_{ij}|^2 \\
&= \sum_{i,j=1}^N |\langle \phi_{i,N_\mathrm{c}} | \phi_{j,N_\mathrm{c}} \rangle + \langle \phi^{(1,1)}_{i,N_\mathrm{c}} | \phi^{(1,1)}_{j,N_\mathrm{c}} \rangle - \delta_{ij}|^2 \quad \text{(by orthogonality)} \\
&= \sum_{i,j=1}^N |\langle \phi^{(1,1)}_{i,N_\mathrm{c}} | \phi^{(1,1)}_{j,N_\mathrm{c}} \rangle|^2 \\
&\leq \sum_{i,j=1}^N \|\phi^{(1,1)}_{i,N_\mathrm{c}}\|^2_{L^2_\#} \|\phi^{(1,1)}_{j,N_\mathrm{c}}\|^2_{L^2_\#} \\
&= \left(\sum_{j=1}^N \|\phi^{(1,1)}_{j,N_\mathrm{c}}\|^2_{L^2_\#} \right)^2 .
\end{aligned} \tag{9.5.15}$$

Using definition (9.4.5), and noting that for all $j = 1, \ldots, N$, the operator $\left(-\frac{1}{2}\Delta - \lambda_{j,N_\mathrm{c}}\right)^{-1}$ is diagonal and commute with $\Pi^\perp_{N_\mathrm{c}}$, we obtain

$$\begin{aligned}
\|S_{N_\mathrm{c}} - 1_N\|_\mathrm{F} =& \sum_{j=1}^N \|(-\frac{1}{2}\Delta - \lambda_{j,N_\mathrm{c}})^{-1} \Pi^\perp_{N_\mathrm{c}} \mathcal{V}^\perp_{N_\mathrm{c}} \phi_{j,N_\mathrm{c}}\|^2_{L^2_\#} \\
\leq& \|\Pi^\perp_{N_\mathrm{c}} (1-\Delta)^{-1/2}\|^2 \max_{i=1,\ldots,N} \|(1-\Delta)^{1/2}(-\frac{1}{2}\Delta - \lambda_{i,N_\mathrm{c}})^{-1}(1-\Delta)^{1/2}\|^2 \\
&\times \sum_{j=1}^N \|(1-\Delta)^{-1/2} \mathcal{V}^\perp_{N_\mathrm{c}} \phi_{j,N_\mathrm{c}}\|^2_{L^2_\#} .
\end{aligned}$$

There exists $N_\mathrm{c}^0 \in \mathbb{N}$ such that for $N_\mathrm{c} \geq N_\mathrm{c}^0$, the operator $(1-\Delta)^{1/2}(-\frac{1}{2}\Delta - \lambda_{i,N_\mathrm{c}})^{-1}(1-\Delta)^{1/2}$ is bounded in $\mathcal{L}(L^2_\#)$ for all $i = 1, \ldots, N$ independently of N_c. Hence, there exist $C \in \mathbb{R}^+$ and $N_\mathrm{c}^0 \in \mathbb{N}$ such that for $N_\mathrm{c} \geq N_\mathrm{c}^0$,

$$\begin{aligned}
\|S_{N_\mathrm{c}} - 1_N\|_\mathrm{F} \leq& C N_\mathrm{c}^{-2} \|(1-\Delta)^{-1/2} \mathcal{V}^\perp_{N_\mathrm{c}} \gamma_{0,N_\mathrm{c}}\|^2_{\mathfrak{S}_2(L^2_\#)} \\
\leq& C N_\mathrm{c}^{-2} \|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|^2_{\mathfrak{S}_2(L^2_\#)},
\end{aligned}$$

from (9.5.13), which concludes the proof of the lemma. □

Proof of estimates (9.5.1) **and** (9.5.2)

Lemma 9.5.6. *The density matrix difference $\gamma_0 - \widetilde{\gamma_{N_\mathrm{c}}}$ can be decomposed as*

$$\gamma_0 - \widetilde{\gamma_{N_\mathrm{c}}} = (\gamma_0 - \gamma_{0,N_\mathrm{c}})^2 + \gamma_{0,N_\mathrm{c}} \gamma^{(2)}_{N_\mathrm{c}} + \gamma^{(2)}_{N_\mathrm{c}} \gamma_{0,N_\mathrm{c}} + \widetilde{Q_{N_\mathrm{c}}} \gamma_{0,N_\mathrm{c}} + \gamma_{0,N_\mathrm{c}} \widetilde{Q_{N_\mathrm{c}}} - \sum_{j=1}^N |\phi^{(1,1)}_{j,N_\mathrm{c}}\rangle\langle \phi^{(1,1)}_{j,N_\mathrm{c}}|.$$

Proof. Let first remark, using (9.5.6), and $\gamma_{0,N_c}^2 = \gamma_{0,N_c}$, that

$$\begin{aligned}\gamma_{0,N_c}\gamma_0 &= \gamma_{0,N_c} + \gamma_{0,N_c}\gamma_{N_c}^{(1)} + \gamma_{0,N_c}\gamma_{N_c}^{(2)} + \gamma_{0,N_c}\widetilde{Q_{N_c}},\\ \gamma_0\gamma_{0,N_c} &= \gamma_{0,N_c} + \gamma_{N_c}^{(1)}\gamma_{0,N_c} + \gamma_{N_c}^{(2)}\gamma_{0,N_c} + \widetilde{Q_{N_c}}\gamma_{0,N_c}.\end{aligned}$$

Moreover, since for all $i,j = 1,2,\cdots,N$, ϕ_{i,N_c} is orthogonal to $\phi_{j,N_c}^{(1)}$, one can show as in the proof of [48, Lemma 4.4] that

$$\gamma_{0,N_c}\gamma_{N_c}^{(1)} + \gamma_{N_c}^{(1)}\gamma_{0,N_c} = \gamma_{N_c}^{(1)}. \tag{9.5.16}$$

Hence

$$\gamma_{0,N_c}\gamma_0 + \gamma_0\gamma_{0,N_c} = 2\gamma_{0,N_c} + \gamma_{N_c}^{(1)} + \gamma_{0,N_c}\gamma_{N_c}^{(2)} + \gamma_{N_c}^{(2)}\gamma_{0,N_c} + \gamma_{0,N_c}\widetilde{Q_{N_c}} + \widetilde{Q_{N_c}}\gamma_{0,N_c},$$

and thus

$$(\gamma_0 - \gamma_{0,N_c})^2 = \gamma_0 - \gamma_{0,N_c} - \gamma_{N_c}^{(1)} - \gamma_{0,N_c}\gamma_{N_c}^{(2)} - \gamma_{N_c}^{(2)}\gamma_{0,N_c} - \widetilde{Q_{N_c}}\gamma_{0,N_c} - \gamma_{0,N_c}\widetilde{Q_{N_c}},$$

from which we deduce using (9.4.10) that

$$\gamma_0 - \widetilde{\gamma_{N_c}} = (\gamma_0 - \gamma_{0,N_c})^2 + \gamma_{0,N_c}\gamma_{N_c}^{(2)} + \gamma_{N_c}^{(2)}\gamma_{0,N_c} + \widetilde{Q_{N_c}}\gamma_{0,N_c} + \gamma_{0,N_c}\widetilde{Q_{N_c}} - \sum_{j=1}^{N} |\phi_{j,N_c}^{(1,1)}\rangle\langle\phi_{j,N_c}^{(1,1)}|.$$

□

Lemma 9.5.7 (Estimation of $(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_c})^2$)**.** *There exist $C \in \mathbb{R}^+$ and $N_c^0 \in \mathbb{N}$, such that for $N_c \geq N_c^0$,*

$$\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_c})^2\|_{\mathfrak{S}_2(L^2_\#)} \leq CN_c^{-2}\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_c})\|_{\mathfrak{S}_2(L^2_\#)}. \tag{9.5.17}$$

Proof. The proof is identical to [48, proof of Lemma 4.5], given the *a priori* estimate (9.3.18). □

Lemma 9.5.8 (Estimation of $(1-\Delta)^{1/2}\widetilde{Q_{N_c}}\gamma_{0,N_c}$)**.** *There exist $C \in \mathbb{R}^+$ and $N_c^0 \in \mathbb{N}$ such that for $N_c \geq N_c^0$,*

$$\|(1-\Delta)^{1/2}\widetilde{Q_{N_c}}\gamma_{0,N_c}\|_{\mathfrak{S}_2(L^2_\#)} \leq CN_c^{-2}\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_c})\|_{\mathfrak{S}_2(L^2_\#)}. \tag{9.5.18}$$

Proof. Using the definition of $\widetilde{Q_{N_c}}$ given in (9.5.8), we have

$$(1-\Delta)^{1/2}\widetilde{Q_{N_c}}\gamma_{0,N_c} = \frac{1}{2\pi i}\oint_\Gamma (1-\Delta)^{1/2}(z-\mathcal{H})^{-1}(\mathcal{V}_{N_c}^\perp + \mathcal{W}_{N_c})\,(z-\mathcal{H}_{N_c})^{-1}\,(\mathcal{V}_{N_c}^\perp + \mathcal{W}_{N_c})\,(z-\mathcal{H}_{N_c})^{-1}\,\gamma_{0,N_c}dz.$$

Therefore, using [48, (12)] two times, we show that there exists $C \in \mathbb{R}^+$ such that

$$\begin{aligned}\|(1-\Delta)^{1/2}\widetilde{Q_{N_c}}\gamma_{0,N_c}\|_{\mathfrak{S}_2(L^2_\#)} \leq C &\max_{z\in\Gamma}\|(1-\Delta)^{1/2}(z-\mathcal{H})^{-1}(1-\Delta)^{1/2}\|\\ &\times \|(1-\Delta)^{-1/2}(\mathcal{V}_{N_c}^\perp + \mathcal{W}_{N_c})(1-\Delta)^{1/2}\|\\ &\times \max_{z\in\Gamma}\|(1-\Delta)^{-1/2}(z-\mathcal{H}_{N_c})^{-1}(\mathcal{V}_{N_c}^\perp + \mathcal{W}_{N_c})\gamma_{0,N_c}\,(z-\mathcal{H}_{N_c})^{-1}\,\|_{\mathfrak{S}_2(L^2_\#)}.\end{aligned}$$

First, $\max_{z\in\Gamma}\|(1-\Delta)^{1/2}(z-\mathcal{H})^{-1}(1-\Delta)^{1/2}\|$ is bounded, which is a classical result (see e.g. [46, Lemma 1] for a proof). Second, there exists $C\in\mathbb{R}^+$ such that

$$\begin{aligned}
\|(1-\Delta)^{-1/2}(\mathcal{V}_{N_\mathrm{c}}^\perp+\mathcal{W}_{N_\mathrm{c}})(1-\Delta)^{1/2}\| &= \|\left[(1-\Delta)^{-1/2}(\mathcal{V}_{N_\mathrm{c}}^\perp+\mathcal{W}_{N_\mathrm{c}})(1-\Delta)^{1/2}\right]^*\| \\
&= \|(1-\Delta)^{1/2}(\mathcal{V}_{N_\mathrm{c}}^\perp+\mathcal{W}_{N_\mathrm{c}})(1-\Delta)^{-1/2}\| \\
&\le C\Big(\|(1-\Delta)^{1/2}\mathcal{V}_{N_\mathrm{c}}^\perp(1-\Delta)^{-1/2}\| \\
&\quad +\|\mathcal{W}_{N_\mathrm{c}}\|_{L^\infty}+\|\nabla\mathcal{W}_{N_\mathrm{c}}\|_{L^3}\Big) \\
&\le C\Big(\|(1-\Delta)^{1/2}\big(V_\mathrm{local}+V_\mathrm{coul}(\rho_{N_\mathrm{c}})+V_\mathrm{xc}(\rho_{N_\mathrm{c}})\big)(1-\Delta)^{-1/2}\| \\
&\quad +\|(1-\Delta)^{1/2}V_\mathrm{nl}(1-\Delta)^{-1/2}\| \\
&\quad +\|\mathcal{W}_{N_\mathrm{c}}\|_{L^\infty}+\|\nabla\mathcal{W}_{N_\mathrm{c}}\|_{L^3}\Big) \\
&\le C\Big(\|V_\mathrm{local}+V_\mathrm{coul}(\rho_{N_\mathrm{c}})+V_\mathrm{xc}(\rho_{N_\mathrm{c}})\|_{L^\infty_\#} \\
&\quad +\|\nabla\left(V_\mathrm{local}+V_\mathrm{coul}(\rho_{N_\mathrm{c}})+V_\mathrm{xc}(\rho_{N_\mathrm{c}})\right)\|_{L^3_\#} \\
&\quad +\|(1-\Delta)^{1/2}V_\mathrm{nl}(1-\Delta)^{-1/2}\| \\
&\quad +\|\mathcal{W}_{N_\mathrm{c}}\|_{L^\infty}+\|\nabla\mathcal{W}_{N_\mathrm{c}}\|_{L^3}\Big),
\end{aligned}$$

from [59, Lemma 17], which is bounded from (9.3.23), (9.3.22), (9.3.24), and the inequality $\|(1-\Delta)^{1/2}V_\mathrm{nl}(1-\Delta)^{-1/2}\|\le\sum_{j=1}^J\|\xi_j\|_{L^2_\#}\|\xi_j\|_{H^1_\#}$. Thus, there exists $C\in\mathbb{R}^+$ such that

$$\|(1-\Delta)^{1/2}\widetilde{Q_{N_\mathrm{c}}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}\le C\max_{z\in\Gamma}\|(1-\Delta)^{-1/2}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}(\mathcal{V}_{N_\mathrm{c}}^\perp+\mathcal{W}_{N_\mathrm{c}})\gamma_{0,N_\mathrm{c}}\,(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\,\|_{\mathfrak{S}_2(L^2_\#)}.$$

Moreover, $\mathrm{Ran}(\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}})\subset X_{N_\mathrm{c}}^\perp$, and the Laplace operator and the projection $\Pi_{N_\mathrm{c}}^\perp$ commute. Therefore, noting that $(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}}=(z+\frac{1}{2}\Delta)^{-1}\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}}$, we obtain

$$\begin{aligned}
\|(1-\Delta)^{1/2}\widetilde{Q_{N_\mathrm{c}}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}\le C\Big[&\|(1-\Delta)^{-1}\Pi_{N_\mathrm{c}}^\perp\|\max_{z\in\Gamma}\|(1-\Delta)^{1/2}(z+\frac{1}{2}\Delta)^{-1}(1-\Delta)^{1/2}\| \\
&\times\max_{z\in\Gamma}\|(1-\Delta)^{-1/2}\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\|_{\mathfrak{S}_2(L^2_\#)} \\
&+\|(1-\Delta)^{-1}\|\max_{z\in\Gamma}\|(1-\Delta)^{1/2}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}(1-\Delta)^{1/2}\| \\
&\times\|(1-\Delta)^{-1/2}\mathcal{W}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\|_{\mathfrak{S}_2(L^2_\#)}\Big].
\end{aligned}$$

As $\|(1-\Delta)^{-1}\|$ and $\max_{z\in\Gamma}\|(1-\Delta)^{1/2}(z+\frac{1}{2}\Delta)^{-1}(1-\Delta)^{1/2}\|$ are bounded, noting that $\|(1-\Delta)^{-1}\Pi_{N_\mathrm{c}}^\perp\|\le CN_\mathrm{c}^{-2}$, and $\max_{z\in\Gamma}\|(1-\Delta)^{1/2}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}(1-\Delta)^{1/2}\|$ is bounded independently of N_c, we can proceed as in [48, (49)] and show that there exist $C\in\mathbb{R}^+$ and $N_\mathrm{c}^0\in\mathbb{N}$ such that for $N_\mathrm{c}\ge N_\mathrm{c}^0$,

$$\begin{aligned}
\|(1-\Delta)^{1/2}\widetilde{Q_{N_\mathrm{c}}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} &\le C\Big(N_\mathrm{c}^{-2}\max_{z\in\Gamma}\|(1-\Delta)^{-1/2}\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\|_{\mathfrak{S}_2(L^2_\#)} \\
&\quad+\max_{z\in\Gamma}\|(1-\Delta)^{-1/2}\mathcal{W}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\|_{\mathfrak{S}_2(L^2_\#)}\Big) \\
&\le C\Big(N_\mathrm{c}^{-2}\|(1-\Delta)^{-1/2}\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} \\
&\quad+\|(1-\Delta)^{-1/2}\mathcal{W}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}\Big).
\end{aligned}$$

We conclude by using (9.5.10) and (9.5.13). □

Lemma 9.5.9 (Estimation of $(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\widetilde{Q_{N_\mathrm{c}}}$)**.** *There exist $C \in \mathbb{R}^+$ and $N_\mathrm{c}^0 \in \mathbb{N}$, such that for $N_\mathrm{c} \geq N_\mathrm{c}^0$,*

$$\|(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\widetilde{Q_{N_\mathrm{c}}}\|_{\mathfrak{S}_2(L^2_\#)} \leq CN_\mathrm{c}^{-2}\|(1-\Delta)^{1/2}(\gamma_0-\gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)}.$$

Proof. The proof is similar to the proof of [48, Lemma 4.7], relying here on Lemma 9.5.8. □

Lemma 9.5.10 (Estimation of $(1-\Delta)^{1/2}\gamma_{N_\mathrm{c}}^{(2)}\gamma_{0,N_\mathrm{c}}$)**.** *There exist $C \in \mathbb{R}^+$ and $N_\mathrm{c}^0 \in \mathbb{N}$, such that for $N_\mathrm{c} \geq N_\mathrm{c}^0$,*

$$\|(1-\Delta)^{1/2}\gamma_{N_\mathrm{c}}^{(2)}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq CN_\mathrm{c}^{-2}\|(1-\Delta)^{1/2}(\gamma_0-\gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)}. \tag{9.5.19}$$

Proof. First, by definition of $\gamma_{N_\mathrm{c}}^{(2)}$ in (9.5.7),

$$(1-\Delta)^{1/2}\gamma_{N_\mathrm{c}}^{(2)}\gamma_{0,N_\mathrm{c}} = \frac{1}{2i\pi}\oint_\Gamma (1-\Delta)^{1/2}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\mathcal{W}_{N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\gamma_{0,N_\mathrm{c}}dz.$$

Hence, using that γ_{0,N_c} and $(z-\mathcal{H}_{N_\mathrm{c}})^{-1}$ commute, that $\max_{z\in\Gamma}\|(1-\Delta)^{1/2}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}(1-\Delta)^{1/2}\|$ is bounded, and proceeding as in [48, (48)], we obtain that there exist $C \in \mathbb{R}^+$ and $N_\mathrm{c}^0 \in \mathbb{N}$ such that for $N_\mathrm{c} \geq N_\mathrm{c}^0$,

$$\begin{aligned}
\|(1-\Delta)^{1/2}\gamma_{N_\mathrm{c}}^{(2)}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} &\leq C\max_{z\in\Gamma}\|(1-\Delta)^{1/2}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}(1-\Delta)^{1/2}\| \\
&\quad\times\max_{z\in\Gamma}\|(1-\Delta)^{-1/2}\mathcal{W}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\|_{\mathfrak{S}_2(L^2_\#)} \\
&\leq C\max_{z\in\Gamma}\|(1-\Delta)^{-1/2}\mathcal{W}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\|_{\mathfrak{S}_2(L^2_\#)} \\
&\leq C\|(1-\Delta)^{-1/2}\mathcal{W}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} \\
&\leq CN_\mathrm{c}^{-2}\|(1-\Delta)^{1/2}(\gamma_0-\gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)},
\end{aligned}$$

where we have used (9.5.10) for this last inequality. This concludes the proof of the lemma. □

Lemma 9.5.11 (Estimation of $(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\gamma_{N_\mathrm{c}}^{(2)}$)**.** *There exist $C \in \mathbb{R}^+$ and $N_\mathrm{c}^0 \in \mathbb{N}$, such that for $N_\mathrm{c} \geq N_\mathrm{c}^0$,*

$$\|(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\gamma_{N_\mathrm{c}}^{(2)}\|_{\mathfrak{S}_2(L^2_\#)} \leq CN_\mathrm{c}^{-2}\|(1-\Delta)^{1/2}(\gamma_0-\gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)}. \tag{9.5.20}$$

Proof. Noting that $\gamma_{0,N_\mathrm{c}}^2 = \gamma_{0,N_\mathrm{c}}$, and from [48, (10)] and [48, (12)], we obtain

$$\begin{aligned}
\|(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\gamma_{N_\mathrm{c}}^{(2)}\|_{\mathfrak{S}_2(L^2_\#)} &\leq \|(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}\|\gamma_{0,N_\mathrm{c}}\gamma_{N_\mathrm{c}}^{(2)}\|_{\mathfrak{S}_2(L^2_\#)} \\
&= \|(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}\|\gamma_{N_\mathrm{c}}^{(2)}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)},
\end{aligned}$$

since γ_{0,N_c} is an orthogonal projector of finite rank. Moreover, as the orbitals $(\phi_{i,N_\mathrm{c}})_{i=1,\cdots,N}$ are bounded in $H^1_\#(\Omega)$ independently of N_c, $\|(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}$ is bounded. On top of that, using [48, (12)],

$$\|\gamma_{N_\mathrm{c}}^{(2)}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq \|(1-\Delta)^{-1/2}\|\|(1-\Delta)^{1/2}\gamma_{N_\mathrm{c}}^{(2)}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq \|(1-\Delta)^{1/2}\gamma_{N_\mathrm{c}}^{(2)}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}.$$

Therefore, we can use the estimate of Lemma 9.5.10 to conclude. □

Lemma 9.5.12 (Estimation of $(1-\Delta)^{1/2}\sum_{j=1}^{N}|\phi_{j,N_\mathrm{c}}^{(1,1)}\rangle\langle\phi_{j,N_\mathrm{c}}^{(1,1)}|$)**.** *There exist $C\in\mathbb{R}^+$ and $N_\mathrm{c}^0\in\mathbb{N}$, such that for $N_\mathrm{c}\geq N_\mathrm{c}^0$,*

$$\|(1-\Delta)^{1/2}\sum_{j=1}^{N}|\phi_{j,N_\mathrm{c}}^{(1,1)}\rangle\langle\phi_{j,N_\mathrm{c}}^{(1,1)}|\|_{\mathfrak{S}_2(L^2_\#)}\leq CN_\mathrm{c}^{-2}\|(1-\Delta)^{1/2}(\gamma_0-\gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)}. \tag{9.5.21}$$

Proof. Expanding $\|(1-\Delta)^{1/2}\sum_{j=1}^{N}|\phi_{j,N_\mathrm{c}}^{(1,1)}\rangle\langle\phi_{j,N_\mathrm{c}}^{(1,1)}|\|^2_{\mathfrak{S}_2(L^2_\#)}$, and using the Cauchy–Schwarz inequality twice, we obtain

$$\begin{aligned}
\|(1-\Delta)^{1/2}\sum_{j=1}^{N}|\phi_{j,N_\mathrm{c}}^{(1,1)}\rangle\langle\phi_{j,N_\mathrm{c}}^{(1,1)}|\|^2_{\mathfrak{S}_2(L^2_\#)} =&\mathrm{Tr}\left(\sum_{i,j=1}^{N}|\phi_{i,N_\mathrm{c}}^{(1,1)}\rangle\langle\phi_{i,N_\mathrm{c}}^{(1,1)}|(1-\Delta)|\phi_{j,N_\mathrm{c}}^{(1,1)}\rangle\langle\phi_{j,N_\mathrm{c}}^{(1,1)}|\right)\\
=&\sum_{i,j=1}^{N}\langle\phi_{i,N_\mathrm{c}}^{(1,1)}|(1-\Delta)|\phi_{j,N_\mathrm{c}}^{(1,1)}\rangle\langle\phi_{j,N_\mathrm{c}}^{(1,1)}|\phi_{i,N_\mathrm{c}}^{(1,1)}\rangle\\
\leq&\sum_{i,j=1}^{N}\|(1-\Delta)^{1/2}\phi_{i,N_\mathrm{c}}^{(1,1)}\|_{L^2_\#}\|(1-\Delta)^{1/2}\phi_{j,N_\mathrm{c}}^{(1,1)}\|_{L^2_\#}\\
&\times\|\phi_{i,N_\mathrm{c}}^{(1,1)}\|_{L^2_\#}\|\phi_{j,N_\mathrm{c}}^{(1,1)}\|_{L^2_\#}\\
\leq&\left(\sum_{i=1}^{N}\|(1-\Delta)^{1/2}\phi_{i,N_\mathrm{c}}^{(1,1)}\|_{L^2_\#}\|\phi_{i,N_\mathrm{c}}^{(1,1)}\|_{L^2_\#}\right)^2.
\end{aligned}$$

Noting that for all $j=1,\dots,N$, $\phi_{j,N_\mathrm{c}}^{(1,1)}\in X_{N_\mathrm{c}}^{\perp}$, so that $\|\phi_{j,N_\mathrm{c}}^{(1,1)}\|_{L^2_\#}\leq CN_\mathrm{c}^{-1}\|\phi_{j,N_\mathrm{c}}^{(1,1)}\|_{H^1_\#}$, with $C=\frac{2\sqrt{2}\pi}{L}$, there holds

$$\|(1-\Delta)^{1/2}\sum_{j=1}^{N}|\phi_{j,N_\mathrm{c}}^{(1,1)}\rangle\langle\phi_{j,N_\mathrm{c}}^{(1,1)}|\|_{\mathfrak{S}_2(L^2_\#)}\leq CN_\mathrm{c}^{-1}\sum_{j=1}^{N}\|(1-\Delta)^{1/2}\phi_{j,N_\mathrm{c}}^{(1,1)}\|^2_{L^2_\#} \tag{9.5.22}$$

$$\leq CN_\mathrm{c}^{-1}\sum_{j=1}^{N}\|(1-\Delta)^{1/2}(-\frac{1}{2}\Delta-\lambda_{j,N_\mathrm{c}})^{-1}(1-\Delta)^{1/2}\|^2 \tag{9.5.23}$$

$$\times\|(1-\Delta)^{-1/2}\mathcal{V}_{N_\mathrm{c}}^{\perp}\phi_{j,N_\mathrm{c}}\|^2_{L^2_\#}. \tag{9.5.24}$$

Moreover, for all $j=1,\dots,N$, $(1-\Delta)^{1/2}(-\frac{1}{2}\Delta-\lambda_{j,N_\mathrm{c}})^{-1}(1-\Delta)^{1/2}$ is a bounded operator. Then using (9.5.13) and (9.3.19), we show that there exist $C\in\mathbb{R}^+$ and $N_\mathrm{c}^0\in\mathbb{N}$ such that for $N_\mathrm{c}\geq N_\mathrm{c}^0$,

$$\begin{aligned}
\|(1-\Delta)^{1/2}\sum_{j=1}^{N}|\phi_{j,N_\mathrm{c}}^{(1,1)}\rangle\langle\phi_{j,N_\mathrm{c}}^{(1,1)}|\|_{\mathfrak{S}_2(L^2_\#)} &\leq CN_\mathrm{c}^{-1}\sum_{j=1}^{N}\|(1-\Delta)^{-1/2}\mathcal{V}_{N_\mathrm{c}}^{\perp}\phi_{j,N_\mathrm{c}}\|^2_{L^2_\#} &(9.5.25)\\
&\leq CN_\mathrm{c}^{-1}\|(1-\Delta)^{-1/2}\mathcal{V}_{N_\mathrm{c}}^{\perp}\gamma_{0,N_\mathrm{c}}\|^2_{\mathfrak{S}_2(L^2_\#)} &(9.5.26)\\
&\leq CN_\mathrm{c}^{-1}\|(1-\Delta)^{1/2}(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|^2_{\mathfrak{S}_2(L^2_\#)} &(9.5.27)\\
&\leq CN_\mathrm{c}^{-2}\|(1-\Delta)^{1/2}(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)} &(9.5.28)
\end{aligned}$$

This concludes the proof. □

Combining the estimations given in Lemmas 9.5.7, 9.5.8, 9.5.9, 9.5.10, 9.5.11, and 9.5.12 in the density matrix difference decomposition of Lemma 9.5.6, we easily obtain (9.5.1).

Lemma 9.5.13 (Estimation of $(1-\Delta)^{1/2}(\widetilde{\gamma_{N_\mathrm{c}}}-\widetilde{\widetilde{\gamma_{N_\mathrm{c}}}})$)**.** *There exist $C\in\mathbb{R}^+$ and $N_\mathrm{c}^0\in\mathbb{N}$, such that for $N_\mathrm{c}\geq N_\mathrm{c}^0$,*

$$\|(1-\Delta)^{1/2}(\widetilde{\gamma_{N_\mathrm{c}}}-\widetilde{\widetilde{\gamma_{N_\mathrm{c}}}})\|_{\mathfrak{S}_2(L^2_\#)}\leq CN_\mathrm{c}^{-2}\|(1-\Delta)^{1/2}(\gamma_0-\gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)}. \tag{9.5.29}$$

Proof. First,

$$\begin{aligned}
\widetilde{\gamma_{N_\mathrm{c}}}-\widetilde{\widetilde{\gamma_{N_\mathrm{c}}}} &= \sum_{i=1}^N|\widetilde{\phi_{i,N_\mathrm{c}}}\rangle\langle\widetilde{\phi_{i,N_\mathrm{c}}}|-\sum_{i=1}^N|\widetilde{\widetilde{\phi_{i,N_\mathrm{c}}}}\rangle\langle\widetilde{\widetilde{\phi_{i,N_\mathrm{c}}}}|\\
&= \sum_{i=1}^N|\widetilde{\phi_{i,N_\mathrm{c}}}\rangle\langle\widetilde{\phi_{i,N_\mathrm{c}}}|-\sum_{i=1}^N|{S_{N_\mathrm{c}}}^{-1/2}\widetilde{\phi_{i,N_\mathrm{c}}}\rangle\langle{S_{N_\mathrm{c}}}^{-1/2}\widetilde{\phi_{i,N_\mathrm{c}}}|\\
&= \sum_{i=1}^N|\widetilde{\phi_{i,N_\mathrm{c}}}\rangle\langle\widetilde{\phi_{i,N_\mathrm{c}}}|-\sum_{i=1}^N\sum_{k,l=1}^N({S_{N_\mathrm{c}}}^{-1/2})_{i,k}|\widetilde{\phi_{k,N_\mathrm{c}}}\rangle\langle\widetilde{\phi_{l,N_\mathrm{c}}}|({S_{N_\mathrm{c}}}^{-1/2})_{l,i}\\
&= \sum_{i=1}^N|\widetilde{\phi_{i,N_\mathrm{c}}}\rangle\langle\widetilde{\phi_{i,N_\mathrm{c}}}|-\sum_{k,l=1}^N\left(\sum_{i=1}^N({S_{N_\mathrm{c}}}^{-1/2})_{i,k}({S_{N_\mathrm{c}}}^{-1/2})_{l,i}\right)|\widetilde{\phi_{k,N_\mathrm{c}}}\rangle\langle\widetilde{\phi_{l,N_\mathrm{c}}}|\\
&= \sum_{i=1}^N|\widetilde{\phi_{i,N_\mathrm{c}}}\rangle\langle\widetilde{\phi_{i,N_\mathrm{c}}}|-\sum_{k,l=1}^N({S_{N_\mathrm{c}}}^{-1})_{k,l}|\widetilde{\phi_{k,N_\mathrm{c}}}\rangle\langle\widetilde{\phi_{l,N_\mathrm{c}}}|\\
&= \sum_{k,l=1}^N(\delta_{k,l}-({S_{N_\mathrm{c}}}^{-1})_{k,l})|\widetilde{\phi_{k,N_\mathrm{c}}}\rangle\langle\widetilde{\phi_{l,N_\mathrm{c}}}|.
\end{aligned}$$

Taking the Hilbert–Schmidt norm, we obtain

$$\begin{aligned}
\|(1-\Delta)^{1/2}(\widetilde{\gamma_{N_\mathrm{c}}}-\widetilde{\widetilde{\gamma_{N_\mathrm{c}}}})\|^2_{\mathfrak{S}_2(L^2_\#)} &= \mathrm{Tr}\left((\widetilde{\gamma_{N_\mathrm{c}}}-\widetilde{\widetilde{\gamma_{N_\mathrm{c}}}})(1-\Delta)(\widetilde{\gamma_{N_\mathrm{c}}}-\widetilde{\widetilde{\gamma_{N_\mathrm{c}}}})\right)\\
&= \sum_{k,l=1}^N\sum_{m,n=1}^N(\delta_{k,l}-({S_{N_\mathrm{c}}}^{-1})_{k,l})(\delta_{m,n}-({S_{N_\mathrm{c}}}^{-1})_{m,n})\\
&\quad\times\langle\widetilde{\phi_{l,N_\mathrm{c}}}|(1-\Delta)|\widetilde{\phi_{m,N_\mathrm{c}}}\rangle\langle\widetilde{\phi_{n,N_\mathrm{c}}}|\widetilde{\phi_{k,N_\mathrm{c}}}\rangle\\
&= \sum_{k,l=1}^N\sum_{m,n=1}^N(\delta_{k,l}-({S_{N_\mathrm{c}}}^{-1})_{k,l})(\delta_{m,n}-({S_{N_\mathrm{c}}}^{-1})_{m,n})\\
&\quad\times\left(\langle\phi_{l,N_\mathrm{c}}|(1-\Delta)|\phi_{m,N_\mathrm{c}}\rangle+\langle\phi^{(1,1)}_{l,N_\mathrm{c}}|(1-\Delta)|\phi^{(1,1)}_{m,N_\mathrm{c}}\rangle\right)\\
&\quad\times\left(\delta_{n,k}+\langle\phi^{(1,1)}_{n,N_\mathrm{c}}|\phi^{(1,1)}_{l,N_\mathrm{c}}\rangle\right),
\end{aligned}$$

as for all $i=1,\dots,N$, $\phi_{i,N_\mathrm{c}}\in X_{N_\mathrm{c}}$ and $\phi^{(1,1)}_{i,N_\mathrm{c}}\in X^\perp_{N_\mathrm{c}}$, and the Laplace operator commutes with Π_{N_c} and $\Pi^\perp_{N_\mathrm{c}}$. Using the Cauchy–Schwarz inequality and noting that for $i=1,\dots,N$, ϕ_{i,N_c} and $\phi^{(1,1)}_{i,N_\mathrm{c}}$ are uniformly bounded in $H^1_\#$-norm independently of N_c, there exists $C\in\mathbb{R}^+$ such that

$$\begin{aligned}
\|(1-\Delta)^{1/2}(\widetilde{\gamma_{N_\mathrm{c}}}-\widetilde{\widetilde{\gamma_{N_\mathrm{c}}}})\|^2_{\mathfrak{S}_2(L^2_\#)} &\leq C\left(\sum_{k,l=1}^N|\delta_{k,l}-({S_{N_\mathrm{c}}}^{-1})_{k,l}|\right)^2\\
&\leq C\|1_N-({S_{N_\mathrm{c}}}^{-1})\|_F^2.
\end{aligned}$$

The matrix S_{N_c} being a perturbation of 1_N, there holds at first order ${1_N - S_{N_c}}^{-1} = S_{N_c} - 1_N + h.o.t$, *h.o.t.* meaning higher order terms. Using (9.5.14), we can therefore conclude in particular that there exist $C \in \mathbb{R}^+$ and $N_c^0 \in \mathbb{N}$ such that for $N_c \geq N_c^0$,

$$\|(1-\Delta)^{1/2}(\widetilde{\gamma_{N_c}} - \widetilde{\widetilde{\gamma_{N_c}}})\|_{\mathfrak{S}_2(L^2_\#)} \leq C N_c^{-2} \|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_c})\|_{\mathfrak{S}_2(L^2_\#)}.$$

□

Combining (9.5.1) with the estimation given in Lemma 9.5.13 allows to prove (9.5.2).

Proof of estimates (9.5.3) **and** (9.5.4)

We start by proving (9.5.4). Let us define $\widetilde{\widetilde{\Phi_{N_c}}}^0 \in \mathcal{M}^{\Phi^0}$ by

$$\min_{U \in \mathcal{U}(N)} \|U\widetilde{\widetilde{\Phi_{N_c}}} - \Phi^0\|_{L^2_\#} = \|\widetilde{\widetilde{\Phi_{N_c}}}^0 - \Phi^0\|_{L^2_\#}.$$

Since $\widetilde{\widetilde{\Phi_{N_c}}}^0 \in \mathcal{M}^{\Phi^0}$, it verifies [44, Lemma 4.7]. Combined with [44, (4.47)], we obtain that there exists $C \in \mathbb{R}^+$ and $N_c^0 \in \mathbb{N}$ such that for all $N_c \geq N_c^0$,

$$\begin{aligned} \mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\widetilde{\widetilde{\Phi_{N_c}}}^0) - \mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\Phi^0) &\leq C\|(1-\Delta)^{1/2}(\widetilde{\widetilde{\Phi_{N_c}}}^0 - \Phi^0)\|^2_{L^2_\#} \\ &\leq C\|(1-\Delta)^{1/2}(\widetilde{\widetilde{\gamma_{N_c}}} - \gamma_0)\|^2_{\mathfrak{S}_2(L^2_\#)}, \end{aligned}$$

from (9.3.15), and noting that the density matrix corresponding to $\widetilde{\widetilde{\Phi_{N_c}}}^0$ is $\widetilde{\widetilde{\gamma_{N_c}}}$. Moreover, from the invariance property (9.2.5), there holds $\mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\widetilde{\widetilde{\Phi_{N_c}}}^0) = \mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\widetilde{\widetilde{\Phi_{N_c}}})$. Using (9.5.2) and (9.3.16), we obtain (9.5.4).

Let us now prove (9.5.3). The same reasoning cannot be applied as the perturbed eigenvectors $\widetilde{\Phi_{N_c}}$ do not satisfy the constraint, *i.e.* $\widetilde{\Phi_{N_c}} \notin \mathcal{M}$. A second-order Taylor expansion on the energy gives

$$\begin{aligned} \mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\widetilde{\Phi_{N_c}}) - \mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\Phi^0) &= \langle (\mathcal{E}^{\mathrm{KS}}_{0,\Omega})'(\Phi^0), \widetilde{\Phi_{N_c}} - \Phi^0 \rangle_{H^{-1}_\#, H^1_\#} \\ &\quad + \frac{1}{2}(\mathcal{E}^{\mathrm{KS}}_{0,\Omega})''(\Phi^0)(\widetilde{\Phi_{N_c}} - \Phi^0, \widetilde{\Phi_{N_c}} - \Phi^0) + h.o.t. \end{aligned}$$

Noting that such a development is also valid for the energy difference $\mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\widetilde{\widetilde{\Phi_{N_c}}}) - \mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\Phi^0)$, we obtain, still up to second order

$$\begin{aligned} \mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\widetilde{\Phi_{N_c}}) - \mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\Phi^0) = \langle (\mathcal{E}^{\mathrm{KS}}_{0,\Omega})'(\Phi^0), \widetilde{\widetilde{\Phi_{N_c}}} - \Phi^0 \rangle_{H^{-1}_\#, H^1_\#} + \frac{1}{2}(\mathcal{E}^{\mathrm{KS}}_{0,\Omega})''(\Phi^0)(\widetilde{\widetilde{\Phi_{N_c}}} - \Phi^0, \widetilde{\widetilde{\Phi_{N_c}}} - \Phi^0) \\ + \langle (\mathcal{E}^{\mathrm{KS}}_{0,\Omega})'(\Phi^0), \widetilde{\Phi_{N_c}} - \widetilde{\widetilde{\Phi_{N_c}}} \rangle_{H^{-1}_\#, H^1_\#} \\ + \frac{1}{2}(\mathcal{E}^{\mathrm{KS}}_{0,\Omega})''(\Phi^0)(\widetilde{\Phi_{N_c}} - \widetilde{\widetilde{\Phi_{N_c}}}, \widetilde{\Phi_{N_c}} + \widetilde{\widetilde{\Phi_{N_c}}} - 2\Phi^0) + h.o.t. \\ = \mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\widetilde{\widetilde{\Phi_{N_c}}}) - \mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\Phi^0) + \langle (\mathcal{E}^{\mathrm{KS}}_{0,\Omega})'(\Phi^0), \widetilde{\Phi_{N_c}} - \widetilde{\widetilde{\Phi_{N_c}}} \rangle_{H^{-1}_\#, H^1_\#} \\ + \frac{1}{2}(\mathcal{E}^{\mathrm{KS}}_{0,\Omega})''(\Phi^0)(\widetilde{\Phi_{N_c}} - \widetilde{\widetilde{\Phi_{N_c}}}, \widetilde{\Phi_{N_c}} + \widetilde{\widetilde{\Phi_{N_c}}} - 2\Phi^0) + h.o.t. \end{aligned}$$

From the continuity of $(\mathcal{E}^{\mathrm{KS}}_{0,\Omega})''(\Phi^0)$ [44, (4.18), (4.47)]), and since $(\mathcal{E}^{\mathrm{KS}}_{0,\Omega})'(\Phi^0)$ is bounded in $H^{-1}_{\#}$-norm [44, Lemma 4.7], there exist $C \in \mathbb{R}^+$ and $N_{\mathrm{c}}^0 \in \mathbb{N}$ such that for $N_{\mathrm{c}} \geq N_{\mathrm{c}}^0$,

$$|\mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\widetilde{\Phi_{N_{\mathrm{c}}}}) - \mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\widetilde{\widetilde{\Phi_{N_{\mathrm{c}}}}})| \leq C\|(1-\Delta)^{1/2}(\widetilde{\Phi_{N_{\mathrm{c}}}} - \widetilde{\widetilde{\Phi_{N_{\mathrm{c}}}}})\|_{L^2_{\#}} + C\|(1-\Delta)^{1/2}(\widetilde{\Phi_{N_{\mathrm{c}}}} - \widetilde{\widetilde{\Phi_{N_{\mathrm{c}}}}})\|^2_{L^2_{\#}}. \quad (9.5.30)$$

Moreover, developing the expression $\widetilde{\Phi_{N_{\mathrm{c}}}} - \widetilde{\widetilde{\Phi_{N_{\mathrm{c}}}}}$, we obtain

$$\|(1-\Delta)^{1/2}(\widetilde{\Phi_{N_{\mathrm{c}}}} - \widetilde{\widetilde{\Phi_{N_{\mathrm{c}}}}})\|_{L^2_{\#}} \leq \|1_N - {S_{N_{\mathrm{c}}}}^{-1/2}\|_{\mathrm{F}} \|(1-\Delta)^{1/2}\widetilde{\Phi_{N_{\mathrm{c}}}}\|_{L^2_{\#}}. \quad (9.5.31)$$

Since there exists $N_{\mathrm{c}}^0 \in \mathbb{N}$ such that for $N_{\mathrm{c}} \geq N_{\mathrm{c}}^0$, $\|(1-\Delta)^{1/2}\widetilde{\Phi_{N_{\mathrm{c}}}}\|_{L^2_{\#}}$ is bounded independently of N_{c}, there holds at first order

$$\begin{aligned}
\|1_N - {S_{N_{\mathrm{c}}}}^{-1/2}\|_{\mathrm{F}} &= \frac{1}{2}\|1_N - S_{N_{\mathrm{c}}}\|_{\mathrm{F}} + h.o.t. \\
&\leq CN_{\mathrm{c}}^{-2}\|(1-\Delta)^{1/2}(\gamma_{0,N_{\mathrm{c}}} - \gamma_0)\|^2_{\mathfrak{S}_2(L^2_{\#})} \\
&\leq CN_{\mathrm{c}}^{-2}|\mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\Phi_{N_{\mathrm{c}}}) - \mathcal{E}^{\mathrm{KS}}_{0,\Omega}(\Phi^0)|, \qquad (9.5.32)
\end{aligned}$$

where $C \in \mathbb{R}^+$, from (9.5.14) and (9.3.16). Combining (9.5.30), (9.5.31) and (9.5.32), we obtain (9.5.3).

9.6 Appendix

Proof of Lemma 9.3.1. The proof of (9.3.14) is identical to the proof of (23) in [48]. This proof indeed relies on results in [44] which are also valid in the nonlinear case. In order to show (9.3.15), let us recall that the positive operator $|\mathcal{H} - \epsilon_F|$ defined by the functional calculus is $|\mathcal{H}-\epsilon_F| = -\gamma_0(\mathcal{H}-\epsilon_F)\gamma_0 + (1-\gamma_0)(\mathcal{H}-\epsilon_F)(1-\gamma_0)$, where γ_0 is the exact density matrix defined in (9.2.8). It is known (see e.g. [46]) that, under Assumption 9.2.1, there exist $0 < c \leq C < \infty$ such that

$$c(1-\Delta) \leq |\mathcal{H} - \epsilon_{\mathrm{F}}| \leq C(1-\Delta),$$

Moreover, as is classical (see e.g. [46, Lemma 1] for a proof) there exists $C \in \mathbb{R}^+$ such that

$$\||\mathcal{H} - \epsilon_F|^{1/2}(1-\Delta)^{-1/2}\| \leq C \quad \text{and} \quad \|(1-\Delta)^{1/2}|\mathcal{H} - \epsilon_F|^{-1/2}\| \leq C. \quad (9.6.1)$$

Let $\Psi^0 \in \mathcal{M}^{\Phi^0}$ with corresponding density matrix $\gamma_{\Psi^0} = \sum_{i=1}^N |\psi_i^0\rangle\langle\psi_i^0|$. First, it can be shown that

$$\begin{aligned}
\||\mathcal{H} - \epsilon_F|^{1/2}(\gamma_{\Psi^0} - \gamma_0)\|^2_{\mathfrak{S}_2(L^2_{\#})} =& \||\mathcal{H} - \epsilon_F|^{1/2}(\Psi^0 - \Phi^0)\|^2_{L^2_{\#}} \\
&+ 2\sum_{j=1}^N |\lambda_j^0 - \epsilon_F| \left[\frac{1}{2}\|\psi_j^0 - \phi_j^0\|^2_{L^2} - \sum_{i=1}^N |\langle\psi_i^0 - \phi_i^0|\phi_j^0\rangle|^2\right]. \qquad (9.6.2)
\end{aligned}$$

Indeed, expressing the density matrix in terms of orbitals and noting that for all $j = 1, \dots, N$, ϕ_j^0 is an eigenvector of $|\mathcal{H} - \epsilon_F|$ with eigenvalue $|\lambda_j^0 - \epsilon_F|$ leads to

$$\begin{aligned}
\||\mathcal{H} - \epsilon_F|^{1/2}(\gamma_{\Psi^0} - \gamma_0)\|_{\sigma_2}^2 =& \text{Tr } ((\gamma_{\Psi^0} - \gamma_0)|\mathcal{H} - \epsilon_F|(\gamma_{\Psi^0} - \gamma_0)) \\
=& \text{Tr } (\gamma_{\Psi^0}|\mathcal{H} - \epsilon_F|\gamma_{\Psi^0}) + \text{Tr } (\gamma_0|\mathcal{H} - \epsilon_F|\gamma_0) \\
& - \text{Tr } (\gamma_0|\mathcal{H} - \epsilon_F|\gamma_{\Psi^0}) - \text{Tr } (\gamma_{\Psi^0}|\mathcal{H} - \epsilon_F|\gamma_0) \\
=& \sum_{i=1}^N \langle \psi_i^0 || \mathcal{H} - \epsilon_F || \psi_i^0 \rangle + \sum_{i=1}^N \langle \phi_i^0 || \mathcal{H} - \epsilon_F || \phi_i^0 \rangle \\
& - \sum_{i,j=1}^N \langle \psi_i^0 | \phi_j^0 \rangle \langle \phi_j^0 || \mathcal{H} - \epsilon_F || \psi_i^0 \rangle - \sum_{i,j=1}^N \langle \psi_i^0 || \mathcal{H} - \epsilon_F || \phi_j^0 \rangle \langle \phi_j^0 | \psi_i^0 \rangle \\
=& \||\mathcal{H} - \epsilon_F|^{1/2}\Psi^0\|_{L^2}^2 + \||\mathcal{H} - \epsilon_F|^{1/2}\Phi^0\|_{\mathfrak{S}_2(L^2_\#)}^2 \\
& - 2\sum_{i,j=1}^N |\lambda_j^0 - \epsilon_F| \langle \psi_i^0 | \phi_j^0 \rangle \langle \phi_j^0 | \psi_i^0 \rangle.
\end{aligned}$$

Then, introducing the orbital error, we obtain

$$\begin{aligned}
\||\mathcal{H} - \epsilon_F|^{1/2}(\gamma_{\Psi^0} - \gamma_0)\|_{\sigma_2}^2 =& \||\mathcal{H} - \epsilon_F|^{1/2}(\Psi^0 - \Phi^0)\|_{L^2}^2 + \sum_{j=1}^N \langle \psi_j^0 || \mathcal{H} - \epsilon_F || \phi_j^0 \rangle \\
& + \sum_{j=1}^N \langle \phi_j^0 || \mathcal{H} - \epsilon_F || \psi_j^0 \rangle - 2\sum_{i,j=1}^N |\lambda_j^0 - \epsilon_F| \langle \psi_i^0 | \phi_j^0 \rangle \langle \phi_j^0 | \psi_i^0 \rangle \\
=& \||\mathcal{H} - \epsilon_F|^{1/2}(\Psi^0 - \Phi^0)\|_{L^2}^2 \\
& + 2\sum_{j=1}^N |\lambda_j^0 - \epsilon_F| \left[\langle \psi_j^0 | \phi_j^0 \rangle - \sum_{i=1}^N \langle \psi_i^0 | \phi_j^0 \rangle \langle \phi_j^0 | \psi_i^0 \rangle \right] \\
=& \||\mathcal{H} - \epsilon_F|^{1/2}(\Psi^0 - \Phi^0)\|_{L^2}^2 \\
& + 2\sum_{j=1}^N |\lambda_j^0 - \epsilon_F| \left[\frac{1}{2}\|\psi_j^0 - \phi_j^0\|_{L^2}^2 - \sum_{i=1}^N |\langle \psi_i^0 - \phi_i^0 | \phi_j^0 \rangle|^2 \right].
\end{aligned}$$

On the one hand,

$$\begin{aligned}
\||\mathcal{H} - \epsilon_F|^{1/2}(\gamma_{\Psi^0} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}^2 \leq& \||\mathcal{H} - \epsilon_F|(\Psi^0 - \Phi^0)\|_{L^2_\#}^2 + \sum_{j=1}^N |\lambda_j^0 - \epsilon_F| \|\psi_j^0 - \phi_j^0\|_{L^2_\#}^2 \\
\leq& \||\mathcal{H} - \epsilon_F|^{1/2}(\Psi^0 - \Phi^0)\|_{L^2_\#}^2 \\
& + \sup_{i=1,\dots,N} |\lambda_i^0 - \epsilon_F| \sum_{j=1}^N \|\psi_j^0 - \phi_j^0\|_{L^2_\#}^2 \\
\leq& \||\mathcal{H} - \epsilon_F|^{1/2}(\Psi^0 - \Phi^0)\|_{L^2_\#}^2 \\
& + \frac{\sup\limits_{i=1,\dots,N} |\lambda_i^0 - \epsilon_F|}{\min\limits_{i \in \mathbb{N}^*} |\lambda_i^0 - \epsilon_F|} \sum_{j=1}^N \||\mathcal{H} - \epsilon_F|^{1/2}(\psi_j^0 - \phi_j^0)\|_{L^2_\#}^2 \\
\leq& C\||\mathcal{H} - \epsilon_F|^{1/2}(\Psi^0 - \Phi^0)\|_{L^2_\#}^2,
\end{aligned}$$

where $C \in \mathbb{R}^+$. On the other hand, using the Cauchy–Schwarz inequality, and (9.3.14),

$$\begin{aligned}
\||\mathcal{H}-\epsilon_F|^{1/2}(\Psi^0-\Phi^0)\|_{L^2_\#}^2 \leq & \||\mathcal{H}-\epsilon_F|^{1/2}(\gamma_{\Psi^0}-\gamma_0)\|_{\sigma_2}^2 \\
& + 2\sum_{j=1}^N |\lambda_j^0-\epsilon_F| \sum_{i=1}^N |\langle \psi_i^0-\phi_i^0|\phi_j^0\rangle|^2 \\
\leq & \||\mathcal{H}-\epsilon_F|^{1/2}(\gamma_{\Psi^0}-\gamma_0)\|_{\sigma_2}^2 \\
& + 2\sup_{k=1,\dots,N} |\lambda_k^0-\epsilon_F| \sum_{j=1}^N \sum_{i=1}^N |\langle \psi_i^0-\phi_i^0|\phi_j^0\rangle|^2 \\
\leq & \||\mathcal{H}-\epsilon_F|^{1/2}(\gamma_{\Psi^0}-\gamma_0)\|_{\sigma_2}^2 \\
& + 2N\sup_{k=1,\dots,N} |\lambda_k^0-\epsilon_F| \sum_{i=1}^N \|\psi_i^0-\phi_i^0\|_{L^2}^2 \\
\leq & \||\mathcal{H}-\epsilon_F|^{1/2}(\gamma_{\Psi^0}-\gamma_0)\|_{\sigma_2}^2 \\
& + 2N\sup_{k=1,\dots,N} |\lambda_k^0-\epsilon_F| \|\gamma_0-\gamma_{\Psi^0}\|_{\mathfrak{S}_2(L^2_\#)}^2. \qquad (9.6.3)
\end{aligned}$$

Moreover,

$$\|\gamma_0-\gamma_{\Psi^0}\|_{\mathfrak{S}_2(L^2_\#)}^2 \leq \frac{1}{\min_{i\in\mathbb{N}^*}|\lambda_i^0-\epsilon_F|} \||\mathcal{H}-\epsilon_F|^{1/2}(\gamma_{\Psi^0}-\gamma_0)\|_{\sigma_2}^2.$$

Therefore, there exists $C \in \mathbb{R}^+$ such that

$$\||\mathcal{H}-\epsilon_F|^{1/2}(\Psi^0-\Phi^0)\|_{L^2}^2 \leq C\||\mathcal{H}-\epsilon_F|^{1/2}(\gamma_{\Psi^0}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}^2. \qquad (9.6.4)$$

Combining (9.6.3) and (9.6.4) with (9.6.1) finishes the proof of the lemma. □

Acknowledgements

The author is grateful to Eric Cancès, Yvon Maday, Benjamin Stamm and Martin Vohralík for helpful discussions and comments. This work was partially undertaken in the framework of CALSIMLAB, supported by the public grant ANR-11-LABX- 0037-01 overseen by the French National Research Agency (ANR) as part of the Investissements d'avenir program (reference: ANR-11-IDEX-0004-02).

Bibliography

[1] K. Aidas, C. Angeli, K. L. Bak, V. Bakken, et al., *The Dalton quantum chemistry program system*, Wiley Interdisciplinary Reviews: Computational Molecular Science, 4 (2014), pp. 269–284.

[2] M. Ainsworth, *Robust a posteriori error estimation for nonconforming finite element approximation*, SIAM J. Numer. Anal., 42 (2005), pp. 2320–2341.

[3] M. Ainsworth, *A posteriori error estimation for discontinuous Galerkin finite element approximation*, SIAM J. Numer. Anal., 45 (2007), pp. 1777–1798.

[4] M. Ainsworth and J. T. Oden, *A posteriori error estimation in finite element analysis*, Wiley-Interscience, New York, 2000.

[5] M. Aizenman and E. H. Lieb, *On semi-classical bounds for eigenvalues of Schrödinger operators*, in The Stability of Matter: From Atoms to Stars, vol. 66, Springer-Verlag, 2005, pp. 241–243.

[6] A. Anantharaman and E. Cancès, *Existence of minimizers for Kohn–Sham models in quantum chemistry*, Annales de l'Institut Henri Poincare (C) Non Linear Analysis, 26 (2009), pp. 2425–2455.

[7] P. F. Antonietti, A. Buffa, and I. Perugia, *Discontinuous Galerkin approximation of the Laplace eigenproblem*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 3483–3503.

[8] P. F. Antonietti, P. Houston, and I. Smears, *A note on optimal spectral bounds for nonoverlapping domain decomposition preconditioners for hp-version discontinuous Galerkin methods*, International Journal of Numerical Analysis and Modeling, 13 (2016), pp. 513–524.

[9] T. Arbogast and Z. Chen, *On the implementation of mixed methods as nonconforming methods for second-order elliptic problems*, Math. Comp., 64 (1995), pp. 943–972.

[10] M. G. Armentano and R. G. Durán, *Asymptotic lower bounds for eigenvalues by nonconforming finite element methods*, Electron. Trans. Numer. Anal., 17 (2004), pp. 93–101.

[11] D. N. Arnold and F. Brezzi, *Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 7–32.

[12] N. Aronszajn, *Approximation methods for eigenvalues of completely continuous symmetric operators*, in Proceedings of the Symposium on Spectral Theory and Differential Problems, 1951, pp. 179–202.

[13] F. Aviat, A. Levitt, B. Stamm, Y. Maday, et al., *Truncated conjugate gradient: an optimal strategy for the analytical evaluation of the many-body polarization energy and forces in molecular simulations*, Journal of Chemical Theory and Computation, 13 (2017), pp. 180–190.

[14] I. Babuška and J. Osborn, *Eigenvalue problems*, in Handbook of numerical analysis, Vol. II, North-Holland, Amsterdam, 1991, pp. 641–787.

[15] I. Babuška and T. Strouboulis, *The finite element method and its reliability*, Numerical Mathematics and Scientific Computation, The Clarendon Press Oxford University Press, New York, 2001.

[16] G. B. Bacskay, *A quadratically convergent Hartree-Fock (QC-SCF) method. Application to closed shell systems*, Chemical Physics, 61 (1981), pp. 385–404.

[17] N. W. Bazley, *Lower bounds for eigenvalues with application to the helium atom*, Physical Review, 120 (1960), pp. 144–149.

[18] N. W. Bazley and D. W. Fox, *Lower bounds for eigenvalues of Schrödinger's equation*, Phys. Rev. (2), 124 (1961), pp. 483–492.

[19] C. Beattie and F. Goerisch, *Methods for computing lower bounds to eigenvalues of self-adjoint operators*, Numerische Mathematik, 72 (1995), pp. 143–172.

[20] A. D. Becke, *Density-functional thermochemistry. III. The role of exact exchange*, The Journal of Chemical Physics, 98 (1993), pp. 5648–5652.

[21] H. Behnke, U. Mertins, M. Plum, and C. Wieners, *Eigenvalue inclusions via domain decomposition*, Proc. R. Soc. Lond. A, 456 (2000), pp. 2717–2730.

[22] T. D. Bhaarathi Natarajana, Mark E. Casidaa, Luigi Genoveseb, *Wavelets for density-functional theory and post-density-functional-theory calculations*, Science, (2011), pp. 1–45.

[23] X. Blanc, É. Cancès, and M.-S. Dupuy, *Variational projector augmented-wave method*, Comptes Rendus Mathématique, 355 (2017), pp. 665 – 670.

[24] F. Bloch, *Über die Quantenmechanik der Elektronen in Kristallgittern*, Zeitschrift für Physik, 52 (1929), pp. 555–600.

[25] P. E. Blöchl, *Projector augmented-wave method*, Physical Review B, 50 (1994), pp. 17953–17979.

[26] H. Blum and M. Dobrowolski, *On finite element methods for elliptic equations on domains with corners*, Computing, 28 (1982), pp. 53–63.

[27] V. Blum, R. Gehrke, F. Hanke, P. Havu, et al., *Ab initio molecular simulations with numeric atom-centered orbitals*, Computer Physics Communications, 180 (2009), pp. 2175–2196.

[28] D. Boffi, *Finite element approximation of eigenvalue problems*, Acta Numerica, 19 (2010), pp. 1–120.

[29] D. Boffi, F. Brezzi, and L. Gastaldi, *On the problem of spurious eigenvalues in the approximation of linear elliptic problems in mixed form*, Math. Comp., 69 (2000), pp. 121–140.

[30] M. Born and V. Fock, *Beweis des Adiabatensatzes*, Zeitschrift für Physik, 51 (1928), pp. 165–180.

[31] M. Born and R. Oppenheimer, *Zur Quantentheorie der Molekeln*, Annalen der Physik, 389 (1927), pp. 457–484.

[32] D. Braess, V. Pillwein, and J. Schöberl, *Equilibrated residual error estimates are p-robust*, Comput. Methods Appl. Mech. Engrg., 198 (2009), pp. 1189–1197.

[33] D. Braess and J. Schöberl, *Equilibrated residual error estimator for edge elements*, Math. Comp., 77 (2008), pp. 651–672.

[34] S. C. Brenner, *Discrete Sobolev and Poincaré inequalities for piecewise polynomial functions*, Electron. Trans. Numer. Anal., 18 (2004), pp. 42–48.

[35] S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, vol. 15 of Texts in Applied Mathematics, Springer, New York, third ed., 2008.

[36] F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*, vol. 15 of Springer Series in Computational Mathematics, Springer-Verlag, New York, 1991.

[37] F. Brezzi, J. Rappaz, and P. A. Raviart, *Finite dimensional approximation of nonlinear problems - Part I: Branches of nonsingular solutions*, Numerische Mathematik, 36 (1980), pp. 1–25.

[38] E. L. Briggs, D. J. Sullivan, and J. Bernholc, *Real-space multigrid-based approach to large-scale electronic structure calculations*, Physical Review B, 54 (1996), pp. 14362–14375.

[39] K. Burke, *Perspective on density functional theory*, The Journal of Chemical Physics, 136 (2012), p. 150901.

[40] G. Caloz and J. Rappaz, *Numerical analysis for nonlinear and bifurcation problems*, Handbook of Numerical Analysis, 5 (1997), pp. 487–637.

[41] E. Cancès, *Self-consistent field algorithms for Kohn–Sham models with fractional occupation numbers*, The Journal of Chemical Physics, 114 (2001), pp. 10616–10622.

[42] E. Cancès, R. Chakir, L. He, and Y. Maday, *Two-grid methods for a class of nonlinear elliptic eigenvalue problems*, IMA Journal of Numerical Analysis, (2017), p. drw053.

[43] E. Cancès, R. Chakir, and Y. Maday, *Numerical analysis of nonlinear eigenvalue problems*, Journal of Scientific Computing, 45 (2010), pp. 90–117.

[44] 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 (2012), pp. 341–388.

[45] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday, *Computational quantum chemistry: A primer*, in Handbook of Numerical Analysis, vol. 10, 2003, pp. 3–270.

[46] 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 (2008), pp. 129–177.

[47] E. Cancès and G. Dusson, *Discretization error cancellation in electronic structure calculation: toward a quantitative study*, ESAIM: Mathematical Modelling and Numerical Analysis, 51 (2017), pp. 1617–1636.

[48] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík, *Post-processing of the planewave approximation of Schrödinger equations. Part I: linear operators*, in preparation.

[49] 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 (2014), pp. 941–946.

[50] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík, *A perturbation-method-based post-processing for the planewave discretization of Kohn–Sham models*, Journal of Computational Physics, 307 (2016), pp. 446–459.

[51] 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*, 2017, `https://hal.inria.fr/hal-01483461/`.

[52] 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 (2017), pp. 2228–2254.

[53] E. Cancès, V. Ehrlacher, and Y. Maday, *Non-consistent approximations of self-adjoint eigenproblems: application to the supercell method*, Numerische Mathematik, 128 (2014), pp. 663–706.

[54] E. Cancès and C. Le Bris, *Can we outperform the DIIS approach for electronic structure calculations?*, International Journal of Quantum Chemistry, 79 (2000), pp. 82–90.

[55] E. Cancès and C. Le Bris, *On the convergence of SCF algorithms for the Hartree-Fock equations*, ESAIM: Mathematical Modelling and Numerical Analysis, 34 (2000), pp. 749–774.

[56] E. Cancès, C. Le Bris, and Y. Maday, *Méthodes mathématiques en chimie quantique. Une introduction*, vol. 53 of Mathématiques & Applications, Springer Berlin Heidelberg, 2006.

[57] E. Cancès and N. Mourad, *A mathematical perspective on density functional perturbation theory*, Nonlinearity, 27 (2014), pp. 1999–2033.

[58] E. Cancès and N. Mourad, *Existence of a type of optimal norm-conserving pseudopotentials for Kohn–Sham models*, Communications in Mathematical Sciences, 14 (2016), pp. 1315–1352.

[59] E. Cancès and G. Stoltz, *A mathematical formulation of the random phase approximation for crystals*, Annales de l'Institut Henri Poincare (C) Non Linear Analysis, 29 (2012), pp. 887–925.

[60] C. G. Canuto, M. Y. Hussaini, A. M. Quarteroni, and T. A. Zang, *Spectral methods: evolution to complex geometries and applications to fluid dynamics*, 2007.

[61] C. Carstensen and S. A. Funken, *Fully reliable localized error control in the FEM*, SIAM J. Sci. Comput., 21, pp. 1465–1484.

[62] C. Carstensen and D. Gallistl, *Guaranteed lower eigenvalue bounds for the biharmonic equation*, Numer. Math., 126 (2014), pp. 33–51.

[63] C. Carstensen and J. Gedicke, *An oscillation-free adaptive FEM for symmetric eigenvalue problems*, Numer. Math., 118 (2011), pp. 401–427.

[64] C. Carstensen and J. Gedicke, *Guaranteed lower bounds for eigenvalues*, Math. Comp., 83 (2014), pp. 2605–2629.

[65] C. Carstensen, J. Gedicke, and D. Rim, *Explicit error estimates for Courant, Crouzeix-Raviart and Raviart-Thomas finite element methods*, J. Comput. Math., 30 (2012), pp. 337–353.

[66] C. Carstensen and C. Merdon, *Computational survey on a posteriori error estimators for nonconforming finite element methods for the Poisson problem*, J. Comput. Appl. Math., 249 (2013), pp. 74–94.

[67] J. Cea, *Approximation variationnelle des problèmes aux limites*, Annales de L'Institut Fourier, 14 (1964), pp. 345–444.

[68] G. Chaban, M. W. Schmidt, and M. S. Gordon, *Approximate second order method for orbital optimization of SCF and MCSCF wavefunctions*, Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta), 97 (1997), pp. 88–95.

[69] F. Chatelin, *Spectral approximation of linear operators*, Academic Press, New York, 1983.

[70] H. Chen, X. Dai, X. Gong, L. He, and A. Zhou, *Adaptive finite element approximations for Kohn–Sham models*, Multiscale Modeling & Simulation, 12 (2014), pp. 1828–1869.

[71] H. Chen, X. Gong, L. He, Z. Yang, and A. Zhou, *Numerical analysis of finite dimensional approximations of Kohn–Sham models*, Advances in Computational Mathematics, 38 (2013), pp. 225–256.

[72] H. Chen, X. Gong, L. He, and A. Zhou, *Convergence of adaptive finite element approximations for nonlinear eigenvalue problems*, 2010, `http://arxiv.org/abs/1001.2344`.

[73] H. Chen, X. Gong, and A. Zhou, *Numerical approximations of a nonlinear eigenvalue problem and applications to a density functional model*, Mathematical Methods in the Applied Sciences, 33 (2010), pp. 1723–1742.

[74] H. Chen and R. Schneider, *Error estimates of some numerical atomic orbitals in molecular simulations*, Communications in Computational Physics, 18 (2015), pp. 125–146.

[75] H. Chen and R. Schneider, *Numerical analysis of augmented plane wave methods for full-potential electronic structure calculations*, ESAIM: Mathematical Modelling and Numerical Analysis, 49 (2015), pp. 755–785.

[76] P. G. Ciarlet, *The finite element method for elliptic problems*, vol. 4, North-Holland, Amsterdam, 1978.

[77] P. Ciarlet Jr. and M. Vohralík, *Robust a posteriori error control for transmission problems with sign changing coefficients using localization of dual norms*, 2015, `https://hal.inria.fr/hal-01148476/`.

[78] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. J. Probert, K. Refson, and M. C. Payne, *First principles methods using CASTEP*, Zeitschrift für Kristallographie, 220 (2005), pp. 567–570.

[79] M. Costabel and A. McIntosh, *On Bogovskii and regularized Poincaré integral operators for de Rham complexes on Lipschitz domains*, Mathematische Zeitschrift, 265 (2010), pp. 297–320.

[80] X. Dai, L. He, and A. Zhou, *Convergence and quasi-optimal complexity of adaptive finite element computations for multiple eigenvalues*, IMA Journal of Numerical Analysis, 35 (2015), pp. 1934–1977.

[81] X. Dai, J. Xu, and A. Zhou, *Convergence and optimal complexity of adaptive finite element eigenvalue computations*, Numerische Mathematik, 110 (2008), pp. 313–355.

[82] E. A. Dari, R. G. Durán, and C. Padra, *A posteriori error estimates for nonconforming approximation of eigenvalue problems*, Appl. Numer. Math., 62 (2012), pp. 580–591.

[83] L. Demkowicz, J. Gopalakrishnan, and J. Schöberl, *Polynomial extension operators I*, SIAM J. Numer. Anal., 46 (2008), pp. 3006–3031.

[84] L. Demkowicz, J. Gopalakrishnan, and J. Schöberl, *Polynomial extension operators. Part III*, Math. Comp., 81 (2012), pp. 1289–1326.

[85] J. Descloux, N. Nassaf, and J. Rappaz, *On spectral approximation. Part II. Error estimates for the Galerkin method*, R.A.I.R.O. Numerical Analysis, 12 (1978), pp. 113–119.

[86] P. Destuynder and B. Métivet, *Explicit error bounds for a nonconforming finite element method*, SIAM J. Numer. Anal., 35 (1998), pp. 2099–2115.

[87] P. Destuynder and B. Métivet, *Explicit error bounds in a conforming finite element method*, Math. Comp., 68 (1999), pp. 1379–1396.

[88] D. A. Di Pietro and A. Ern, *Mathematical aspects of discontinuous Galerkin methods*, vol. 69, Springer, Heidelberg, 2012.

[89] P. A. M. Dirac, *The quantum theory of the electron*, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 117 (1928), pp. 610–624.

[90] P. A. M. Dirac, *Quantum mechanics of many-electron systems*, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 123 (1929), pp. 714–733.

[91] V. Dolejší, A. Ern, and M. Vohralík, *hP-adaptation driven by polynomial-degree-robust a posteriori error estimates for elliptic problems*, SIAM J. Sci. Comput., 38 (2016), pp. A3220–A3246.

[92] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[93] R. M. Dreizler and E. K. U. Gross, *Density functional theory*, Springer Berlin Heidelberg, 1990.

[94] R. G. Durán, L. Gastaldi, and C. Padra, *A posteriori error estimators for mixed approximations of eigenvalue problems*, Math. Models Methods Appl. Sci., 9 (1999), pp. 1165–1178.

[95] R. G. Durán, C. Padra, and R. Rodríguez, *A posteriori error estimates for the finite element approximation of eigenvalue problems*, Math. Models Methods Appl. Sci., 13 (2003), pp. 1219–1229.

[96] G. Dusson, *A preliminary note on the post-processing of Rayleigh–Ritz eigenfunctions*, 2017, in preparation.

[97] G. Dusson, *Post-processing of the planewave approximation of Schrödinger equations. Part II: Kohn–Sham models*, 2017, in preparation.

[98] G. Dusson and Y. Maday, *A posteriori analysis of a nonlinear Gross–Pitaevskii-type eigenvalue problem*, IMA Journal of Numerical Analysis, (2016), p. drw001.

[99] G. Dusson and Y. Maday, *An overview of* a posteriori *estimation and post-processing methods for nonlinear eigenvalue problems*, 2017, in preparation.

[100] F. J. Dyson, *The S Matrix in Quantum Electrodynamics*, Physical Review, 75 (1949), pp. 1736–1755.

[101] L. El Alaoui, A. Ern, and M. Vohralík, *Guaranteed and robust a posteriori error estimates and balancing discretization and linearization errors for monotone nonlinear problems*, Computer Methods in Applied Mechanics and Engineering, 200 (2011), pp. 2782–2795.

[102] A. Ern, S. Nicaise, and M. Vohralík, *An accurate $H(div)$ flux reconstruction for discontinuous Galerkin approximations of elliptic problems*, C. R. Math. Acad. Sci. Paris, 345 (2007), pp. 709–712.

[103] A. Ern and M. Vohralík, *Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs*, SIAM Journal on Scientific Computing, 35 (2013), pp. A1761–A1791.

[104] A. Ern and M. Vohralík, *Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs*, SIAM J. Sci. Comput., 35 (2013), pp. A1761–A1791.

[105] A. Ern and M. Vohralík, *Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations*, SIAM J. Numer. Anal., 53 (2015), pp. 1058–1081.

[106] A. Ern and M. Vohralík, *Stable broken H^1 and $H(div)$ polynomial extensions for polynomial-degree-robust potential and flux reconstruction in three space dimensions*, 2016, `https://hal.inria.fr/hal-01422204`.

[107] E. Fermi, *Statistical method to determine some properties of atoms*, Rend. Accad. Naz. Lincei, 6 (1927), pp. 602–607.

[108] T. H. Fischer and J. Almlöf, *General methods for geometry and wave function optimization*, J. Phys. Chem., 96 (1992), pp. 9768–9774.

[109] G. E. Forsythe, *Asymptotic lower bounds for the fundamental frequency of convex membranes*, Pacific J. Math., 5 (1955), pp. 691–702.

[110] D. W. Fox and W. C. Rheinboldt, *Computational methods for determining lower bounds for eigenvalues of operators in Hilbert space*, SIAM Rev., 8 (1966), pp. 427–462.

[111] J. B. Francisco, J. M. Martinez, and L. Martinez, *Globally convergent trust-region methods for self-consistent field electronic structure calculations*, The Journal of Chemical Physics, 121 (2004), p. 10863.

[112] H. Frisch, MJ and Trucks, GW and Schlegel, HB and Scuseria, GE and Robb, MA and Cheeseman, JR and Scalmani, G and Barone, V and Petersson, GA and Nakatsuji, *Gaussian Inc 16, revision A. 03*, Wallingford CT, (2016).

[113] S. Giani and E. J. C. Hall, *An a posteriori error estimator for hp-adaptive discontinuous Galerkin methods for elliptic eigenvalue problems*, Math. Models Methods Appl. Sci., 22 (2012), pp. 35,1250030.

[114] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, et al., *QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials*, Journal of Physics: Condensed Matter, 21 (2009), p. 395502.

[115] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, vol. 224, Springer-Verlag, Berlin, second ed., 1983.

[116] S. Goedecker, *Wavelets and their application for the solution of partial differential equations in physics*, Presses polytechniques et universitaires romandes, 1998.

[117] S. Goedecker, M. Teter, and J. Hutter, *Separable dual-space Gaussian pseudopotentials*, Physical Review B, 54 (1996), pp. 1703–1710.

[118] F. Goerisch and Z. Q. He, *The determination of guaranteed bounds to eigenvalues with the use of variational methods I*, in Computer arithmetic and self-validating numerical methods (Basel, 1989), vol. 7, Academic Press, Boston, MA, 1990, pp. 137–153.

[119] X. Gonze, *Perturbation expansion of variational principles at arbitrary order*, Physical Review A, 52 (1995), pp. 1086–1095.

[120] X. Gonze, B. Amadon, P.-M. Anglade, J.-M. Beuken, et al., *ABINIT: First-principles approach to material and nanosystem properties*, Computer Physics Communications, 180 (2009), pp. 2582–2615.

[121] X. Gonze, J.-M. Beuken, R. Caracas, F. Detraux, et al., *First-principles computation of material properties: the ABINIT software project*, Computational Materials Science, 25 (2002), pp. 478–492.

[122] P. Grisvard, *Elliptic problems in nonsmooth domains*, vol. 24 of Monographs and Studies in Mathematics, Pitman Advanced Publishing Program, Boston, MA, 1985.

[123] L. Grubišić and J. S. Ovall, *On estimators for eigenvalue/eigenvector approximations*, Math. Comp., 78 (2009), pp. 739–770.

[124] B. Guo and I. Babuška, *The h-p version of the finite element method*, Computational Mechanics, 1 (1986), pp. 21–41.

[125] F. Gygi and G. Galli, *Real-space adaptive-coordinate electronic-structure calculations*, Physical Review B, 52 (1995), pp. R2229–R2232.

[126] G. A. Hagedorn and A. Joye, *Mathematical analysis of Born-Oppenheimer approximations*, Spectral Theory and Mathematical Physics: Quantum field theory, statistical mechanics, and nonrelativistic quantum systems, (2007), pp. 203–226.

[127] B. L. Hammond, J. W. A. Lester, and P. J. Reynolds, *Monte Carlo methods in ab initio quantum chemistry*, World Scientific, 1994.

[128] M. Hanrath, *Wavefunction quality and error estimation of single- and multi-reference coupled-cluster and CI methods: The H4 model system*, Chemical Physics Letters, 466 (2008), pp. 240–246.

[129] J. Harris, *Simplified method for calculating the energy of weakly interacting fragments*, Physical Review B, 31 (1985), pp. 1770–1779.

[130] R. J. Harrison, G. I. Fann, Z. Gan, T. Yanai, S. Sugiki, A. Beste, and G. Beylkin, *Multiresolution computational chemistry*, Journal of Physics: Conference Series, 16 (2005), pp. 243–246.

[131] F. Hecht, *New development in FreeFem++*, J. Numer. Math., 20 (2012), pp. 251–265.

[132] F. Hecht, O. Pironneau, J. Morice, A. Le Hyaric, and K. Ohtsuka, *FreeFem++*, tech. report, Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Paris, 2012.

[133] C. Heil, *Ten lectures on Wavelets (Ingrid Daubechies)*, SIAM Review, 35 (1993), pp. 666–669.

[134] B. Helffer, *Spectral theory and its applications*, vol. 139, Cambridge University Press, 2013.

[135] T. Helgaker, P. Jørgensen, and J. Olsen, *Molecular electronic-structure theory*, John Wiley & Sons, Ltd, Chichester, UK, 2000.

[136] V. Heuveline and R. Rannacher, *A posteriori error control for finite approximations of elliptic eigenvalue problems*, Adv. Comput. Math., 15 (2001), pp. 107–138.

[137] J. Hinze and C. C. J. Roothaan, *Multi-configuration self-consistent-field theory*, Progress of Theoretical Physics Supplement, 40 (1967), pp. 37–51.

[138] M. Hoffmann-Ostenhof and T. Hoffmann-Ostenhof, *"Schrödinger inequalities" and asymptotic behavior of the electron density of atoms and molecules*, Phys. Rev. A, 16 (1977), pp. 1782–1785.

[139] P. Hohenberg and W. Kohn, *Inhomogeneous electron gas*, Physical Review, 136 (1964), pp. B864–B871.

[140] J. Hu, Y. Huang, and Q. Lin, *Lower bounds for eigenvalues of elliptic operators: by nonconforming finite element methods*, J. Sci. Comput., 61 (2014), pp. 196–221.

[141] J. Hu, Y. Huang, and Q. Shen, *The lower/upper bound property of approximate eigenvalues by nonconforming finite element methods for elliptic operators*, J. Sci. Comput., 58 (2014), pp. 574–591.

[142] W. Hu, L. Lin, and C. Yang, *DGDFT: A massively parallel method for large scale density functional theory calculations*, The Journal of Chemical Physics, 143 (2015), p. 124110.

[143] S. Jia, H. Chen, and H. Xie, *A posteriori error estimator for eigenvalue problems by mixed finite element method*, Sci. China Math., 56 (2013), pp. 887–900.

[144] P. Jiránek, Z. Strakoš, and M. Vohralík, *A posteriori error estimates including algebraic error and stopping criteria for iterative solvers*, SIAM Journal on Scientific Computing, 32 (2010), pp. 1567–1590.

[145] T. Kato, *On the upper and lower bounds of eigenvalues*, J. Phys. Soc. Japan, 4 (1949), pp. 334–339.

[146] T. Kato, *Perturbation theory for linear operators*, Springer Berlin Heidelberg, 1976.

[147] J. Kaye, L. Lin, and C. Yang, *A posteriori error estimator for adaptive local basis functions to solve Kohn-Sham density functional theory*, Communications in Mathematical Sciences, 13 (2015), pp. 1741–1773.

[148] G. P. Kerker, *Non-singular atomic pseudopotentials for solid state applications*, Journal of Physics C: Solid State Physics, 13 (1980), pp. L189–L194.

[149] K.-Y. Kim, *A posteriori error analysis for locally conservative mixed methods*, Math. Comp., 76 (2007), pp. 43–66.

[150] K.-Y. Kim, *A posteriori error estimators for locally conservative methods of nonlinear elliptic problems*, Appl. Numer. Math., 57 (2007), pp. 1065–1080.

[151] W. Kohn, *Improvement of Rayleigh–Ritz eigenfunctions*, SIAM Review, 14 (1972), pp. 399–419.

[152] W. Kohn and L. J. Sham, *Self-consistent equations including exchange and correlation effects*, Physical Review, 140 (1965), pp. A1133–A1138.

[153] G. Kresse and J. Furthmüller, *Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set*, Computational Materials Science, 6 (1996), pp. 15–50.

[154] K. N. Kudin, G. E. Scuseria, and E. Cancès, *A black-box self-consistent field convergence algorithm: One step closer*, The Journal of Chemical Physics, 116 (2002), p. 8255.

[155] J. R. Kuttler and V. G. Sigillito, *Bounding eigenvalues of elliptic operators*, SIAM J. Math. Anal., 9 (1978), pp. 768–778.

[156] J. R. Kuttler and V. G. Sigillito, *Estimating eigenvalues with a posteriori/a priori inequalities*, vol. 135 of Research Notes in Mathematics, Pitman Advanced Publishing Program, Boston, MA, 1985.

[157] W. Kutzelnigg, *Error analysis and improvements of coupled-cluster theory*, Theoretica Chimica Acta, 80 (1991), pp. 349–386.

[158] W. Kutzelnigg, *Rate of convergence of basis expansions in quantum chemistry*, AIP Conference Proceedings, 1504 (2012), pp. 15–30.

[159] Y. A. Kuznetsov and S. I. Repin, *Guaranteed lower bounds of the smallest eigenvalues of elliptic differential operators*, J. Numer. Math., 21 (2013), pp. 135–156.

[160] P. Ladevèze and D. Leguillon, *Error estimate procedure in the finite element method and applications*, SIAM J. Numer. Anal., 20 (1983), pp. 485–509.

[161] B. Langwallner, C. Ortner, and E. Süli, *Existence and convergence results for the Galerkin approximation of an electronic density functional*, Mathematical Models and Methods in Applied Sciences, 20 (2010), pp. 2237–2265.

[162] M. G. Larson, *A posteriori and a priori error analysis for finite element approximations of self-adjoint elliptic eigenvalue problems*, SIAM J. Numer. Anal., 38 (2000), pp. 608–625.

[163] C. Lee, W. Yang, and R. G. Parr, *Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density*, Physical Review B, 37 (1988), pp. 785–789.

[164] K. Lejaeghere, G. Bihlmayer, T. Bjorkman, P. Blaha, et al., *Reproducibility in density functional theory calculations of solids*, Science, 351 (2016), p. aad3000.

[165] W. A. Lester, S. M. Rothstein, and S. Tanaka, *Recent Advances in Quantum Monte Carlo Methods - Part II*, World Scientific, 2002.

[166] W. Lester Jr. (ed.), *Recent advances in Quantum Monte Carlo methods*, World Scientific, 1997.

[167] 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 (1979), pp. 6062–6065.

[168] S. Li, K. Chen, M.-Y. Hsieh, N. Muralimanohar, et al., *System implications of memory reliability in exascale computing*, in Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis, New York, 2011, ACM, pp. 46:1–46:12.

[169] E. H. Lieb, *Density functionals for Coulomb systems*, International Journal of Quantum Chemistry, 24 (1983), pp. 243–277.

[170] E. H. Lieb, R. Seiringer, and J. Yngvason, *Bosons in a trap: A rigorous derivation of the Gross-Pitaevskii energy functional*, in The Stability of Matter: From Atoms to Stars, Springer Berlin Heidelberg, 2001, pp. 685–697.

[171] E. H. Lieb and B. Simon, *The Hartree-Fock theory for Coulomb systems*, Communications in Mathematical Physics, 53 (1977), pp. 185–194.

[172] L. Lin, J. Lu, L. Ying, and W. E, *Adaptive local basis set for Kohn–Sham density functional theory in a discontinuous Galerkin framework I: Total energy calculation*, Journal of Computational Physics, 231 (2012), pp. 2140–2154.

[173] L. Lin and B. Stamm, *A posteriori error estimates for discontinuous Galerkin methods using non-polynomial basis functions Part I: Second order linear PDE*, ESAIM: Mathematical Modelling and Numerical Analysis, 50 (2016), pp. 1193–1222.

[174] L. Lin and B. Stamm, *A posteriori error estimates for discontinuous Galerkin methods using non-polynomial basis functions. Part II: Eigenvalue problems*, ESAIM: Mathematical Modelling and Numerical Analysis, (2016).

[175] P. L. Lions, *Solutions of Hartree-Fock equations for Coulomb systems*, Communications in Mathematical Physics, 109 (1987), pp. 33–97.

[176] X. Liu, *A framework of verified eigenvalue bounds for self-adjoint differential operators*, Appl. Math. Comput., 267 (2015), pp. 341–355.

[177] X. Liu and F. Kikuchi, *Analysis and estimation of error constants for P_0 and P_1 interpolations over triangular finite elements*, J. Math. Sci. Univ. Tokyo, 17 (2010), pp. 27–78.

[178] X. Liu and S. Oishi, *Verified eigenvalue evaluation for the Laplacian over polygonal domains of arbitrary shape*, SIAM J. Numer. Anal., 51 (2013), pp. 1634–1654.

[179] P. O. Lowdin, *Studies in perturbation theory. X. Lower bounds to energy eigenvalues in perturbation-theory ground state*, Physical Review, 139 (1965).

[180] 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*, Sci. China Math., 55 (2012), pp. 1069–1082.

[181] D. I. Lyakh and R. J. Bartlett, *An adaptive coupled-cluster theory: @CC approach*, The Journal of Chemical Physics, 133 (2010), p. 244112.

[182] Y. Maday, *h-P Finite element approximation for full-potential electronic structure calculations*, Chinese Annals of Mathematics, Series B, 35 (2014), pp. 1–24.

[183] Y. Maday and A. T. Patera, *Numerical analysis of a posteriori finite element bounds for linear functional outputs*, Math. Models Methods Appl. Sci., 10 (2000), pp. 785–799.

[184] Y. Maday and G. Turinici, *Error bars and quadratically convergent methods for the numerical simulation of the Hartree-Fock equations*, Numerische Mathematik, 94 (2003), pp. 739–770.

[185] S. Mallat, *A wavelet tour of signal processing*, Elsevier, 2009.

[186] S. Mao and Z.-c. Shi, *Explicit error estimates for mixed and nonconforming finite elements*, J. Comput. Math., 27 (2009), pp. 425–440.

[187] C. Marcati, *hP discontinuous Galerkin finite element method for electronic structure calculation*, PhD thesis, 2017.

[188] M. A. L. Marques, A. Castro, G. F. Bertsch, and A. Rubio, *Octopus: A first-principles tool for excited electron-ion dynamics*, Computer Physics Communications, 151 (2003), pp. 60–78.

[189] V. Mehrmann and A. Miedlar, *Adaptive computation of smallest eigenvalues of self-adjoint elliptic partial differential equations*, Numer. Linear Algebra Appl., 18 (2011), pp. 387–409.

[190] B. Mercier, J. Osborn, J. Rappaz, and P.-A. Raviart, *Eigenvalue approximation by mixed and hybrid methods*, Math. Comp., 36 (1981), pp. 427–453.

[191] S. Mohr, L. E. Ratcliff, P. Boulanger, L. Genovese, D. Caliste, T. Deutsch, and S. Goedecker, *Daubechies wavelets for linear scaling density functional theory*, The Journal of Chemical Physics, 140 (2014), p. 204110.

[192] C. B. Moler and L. E. Payne, *Bounds for eigenvalues and eigenvectors of symmetric operators*, SIAM J. Numer. Anal., 5 (1968), pp. 64–70.

[193] P. Motamarri, M. Nowak, K. Leiter, J. Knap, and V. Gavini, *Higher-order adaptive finite-element methods for Kohn–Sham density functional theory*, Journal of Computational Physics, 253 (2013), pp. 308–343.

[194] F. Neese, A. Hansen, and D. G. Liakos, *Efficient and accurate approximations to the local coupled cluster singles doubles method using a truncated pair natural orbital basis*, Journal of Chemical Physics, 131 (2009), p. 064103.

[195] G. Panati, H. Spohn, and S. Teufel, *The time-dependent Born-Oppenheimer approximation*, ESAIM: Mathematical Modelling and Numerical Analysis, 41 (2007), pp. 297–314.

[196] J. Papež, U. Rüde, M. Vohralík, and B. Wohlmuth, *Sharp algebraic and total a posteriori error bounds via a multilevel approach*, 2017, in preparation.

[197] J. Papež, Z. Strakoš, and M. Vohralík, *Estimating and localizing the algebraic and total numerical errors using flux reconstructions*, 2016, `https://hal.inria.fr/hal-01312430`.

[198] J. E. Pask and P. A. Sterne, *Finite element methods in ab initio electronic structure calculations*, Modelling and Simulation in Materials Science and Engineering, 13 (2005), pp. R71–R96.

[199] J. P. Perdew, K. Burke, and M. Ernzerhof, *Generalized gradient approximation made simple*, Physical Review Letters, 77 (1996), pp. 3865–3868.

[200] J. P. Perdew and Y. Wang, *Accurate and simple analytic representation of the electron-gas correlation energy*, Physical Review B, 45 (1992), pp. 13244–13249.

[201] J. P. Perdew and A. Zunger, *Self-interaction correction to density-functional approximations for many-electron systems*, Physical Review B, 23 (1981), pp. 5048–5079.

[202] P. Pernot, B. Civalleri, D. Presti, and A. Savin, *Prediction uncertainty of density functional approximations for properties of crystals with cubic symmetry*, The Journal of Physical Chemistry A, 119 (2015), pp. 5288–5304.

[203] S. N. Pieniazek, F. R. Clemente, and K. N. Houk, *Sources of error in DFT computations of C-C bond formation thermochemistries: $\pi \rightarrow \sigma$ transformations and error cancellation by DFT methods*, Angewandte Chemie International Edition, 47 (2008), pp. 7746–7749.

[204] L. Pitaevskii and S. Stringari, *Bose-Einstein condensation (International series of monographs on physics)*, Oxford University Press, USA, 2003.

[205] M. Plum, *Guaranteed numerical bounds for eigenvalues*, in Spectral theory and computational methods of Sturm-Liouville problems (Knoxville, TN, 1996), vol. 191, Dekker, New York, 1997, pp. 313–332.

[206] S. Poncé, G. Antonius, P. Boulanger, E. Cannuccia, A. Marini, M. Côté, and X. Gonze, *Verification of first-principles codes: Comparison of total energies, phonon frequencies, electron–phonon coupling and zero-point motion correction to the gap between ABINIT and QE/Yambo*, Computational Materials Science, 83 (2014), pp. 341–348.

[207] W. Prager and J. L. Synge, *Approximations in elasticity based on the concept of function space*, Quart. Appl. Math., 5 (1947), pp. 241–269.

[208] P. Pulay, *Improved SCF convergence acceleration*, Journal of Computational Chemistry, 3 (1982), pp. 556–560.

[209] P. Pyykkö and J.-P. Desclaux, *Relativity and the periodic system of elements*, Accounts of Chemical Research, 12 (1979), pp. 276–281.

[210] R. Rannacher, A. Westenberger, and W. Wollner, *Adaptive finite element solution of eigenvalue problems: balancing of discretization and iteration error*, J. Numer. Math., 18 (2010), pp. 303–327.

[211] M. Reed and B. Simon, *Methods of modern mathematical physics*, Academic Press, New York, 1972.

[212] M. Reed and B. Simon, *Methods of modern mathematical physics. I. Functional analysis*, vol. 53, Academic Press, New York, 1972.

[213] S. I. Repin, *A posteriori estimates for partial differential equations*, Walter de Gruyter, Berlin, 2008.

[214] J. E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II, North-Holland, Amsterdam, 1991, pp. 523–639.

[215] T. Rohwedder and R. Schneider, *An analysis for the DIIS acceleration method used in quantum chemistry calculations*, Journal of Mathematical Chemistry, 49 (2011), pp. 1889–1914.

[216] T. Rohwedder and R. Schneider, *Error estimates for the Coupled Cluster method*, ESAIM: Mathematical Modelling and Numerical Analysis, 47 (2013), pp. 1553–1582.

[217] C. C. J. Roothaan, *New developments in molecular orbital theory*, Reviews of Modern Physics, 23 (1951), pp. 69–89.

[218] Y. Saad, *Numerical methods for large eigenvalue problems*, Manchester University Press, Manchester, UK, 1992.

[219] Y. Saad, J. R. Chelikowsky, and S. M. Shontz, *Numerical methods for electronic structure calculations of materials*, SIAM Review, 52 (2010), pp. 3–54.

[220] V. R. Saunders and I. H. Hillier, *A level-shifting method for converging closed shell Hartree-Fock wave functions*, International Journal of Quantum Chemistry, 7 (1973), pp. 699–705.

[221] R. Schneider, *Analysis of the projected coupled cluster method in electronic structure calculation*, Numerische Mathematik, 113 (2009), pp. 433–471.

[222] J. B. Schriber and F. A. Evangelista, *Communication: An adaptive configuration interaction approach for strongly correlated electrons with tunable accuracy*, The Journal of Chemical Physics, 144 (2016), p. 161106.

[223] E. Schrödinger, *An undulatory theory of the mechanics of atoms and molecules*, Physical Review, 28 (1926), pp. 1049–1070.

[224] I. Šebestová and T. Vejchodský, *Two-sided bounds for eigenvalues of differential operators with applications to Friedrichs, Poincaré, trace, and similar constants*, SIAM J. Numer. Anal., 52 (2014), pp. 308–329.

[225] I. Šebestová and T. Vejchodský, *Two-sided bounds of eigenvalues – local efficiency and convergence of adaptive algorithm*, 2016, `https://arxiv.org/abs/1606.01739`.

[226] J. C. Slater, *The theory of complex spectra*, Physical Review, 34 (1929), pp. 1293–1322.

[227] P. Solin and S. Giani, *An iterative adaptive finite element method for elliptic eigenvalue problems*, J. Comput. Appl. Math., 236 (2012), pp. 4582–4599.

[228] G. Still, *Computable bounds for eigenvalues and eigenfunctions of elliptic differential operators*, Numer. Math., 54 (1988), pp. 201–223.

[229] G. Strang and G. Fix, *An analysis of the finite element method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[230] A. Szabo and N. S. Ostlund, *Modern quantum chemistry: Introduction to advanced electronic structure theory*, Dover Publications, 1996.

[231] P. L. Taylor, *Quantum approach to the solid state*, Prentice-Hall, New Jersey, 1970.

[232] G. Temple, *The theory of Rayleigh's principle as applied to continuous systems*, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 119 (1928), pp. 276–293.

[233] L. H. Thomas, *The calculation of atomic fields*, Mathematical Proceedings of the Cambridge Philosophical Society, 23 (1927), p. 542.

[234] L. N. Trefethen and T. Betcke, *Computed eigenmodes of planar regions*, in Recent advances in differential equations and mathematical physics, vol. 412 of Contemp. Math., Amer. Math. Soc., Providence, RI, 2006, pp. 297–314.

[235] N. Troullier and J. Martins, *A straightforward method for generating soft transferable pseudopotentials*, Solid State Communications, 74 (1990), pp. 613–616.

[236] E. Tsuchida and M. Tsukada, *Electronic-structure calculations based on the finite-element method*, Physical Review B, 52 (1995), pp. 5573–5578.

[237] S. M. Valone, *Consequences of extending 1-matrix energy functionals from pure-state representable to all ensemble representable 1 matrices*, The Journal of Chemical Physics, 73 (1980), pp. 1344–1349.

[238] D. Vanderbilt, *Soft self-consistent pseudopotentials in a generalized eigenvalue formalism*, Physical Review B, 41 (1990), pp. 7892–7895.

[239] R. Verfürth, *A posteriori error estimates for nonlinear problems. Finite element discretizations of elliptic equations*, Math. Comp., 62 (1994), pp. 445–475.

[240] R. Verfürth, *A review of a posteriori error estimation and adaptive mesh-refinement techniques*, Teubner-Wiley, Stuttgart, 2013.

[241] M. Vohralík, *On the discrete Poincaré–Friedrichs inequalities for nonconforming approximations of the Sobolev space H^1*, Numer. Funct. Anal. Optim., 26 (2005), pp. 925–952.

[242] M. Vohralík, *A posteriori error estimates for lowest-order mixed finite element discretizations of convection-diffusion-reaction equations*, SIAM J. Numer. Anal., 45 (2007), pp. 1570–1599.

[243] C. F. von Weizsäcker, *Zur Theorie der Kernmassen*, Z. Phys., 96 (1935), pp. 431–458.

[244] H. Wang, M. Liao, P. Lin, and L. Zhang, *A posteriori error estimation and adaptive algorithm for the Atomistic/Continuum coupling in 2D*, `https://arxiv.org/pdf/1702.02701.pdf`.

[245] L. Wang, L. Chamoin, P. Ladevèze, and H. Zhong, *Computable upper and lower bounds on eigenfrequencies*, Computer Methods in Applied Mechanics and Engineering, 302 (2016), pp. 27–43.

[246] H. F. Weinberger, *Upper and lower bounds for eigenvalues by finite difference methods*, Comm. Pure Appl. Math., 9 (1956), pp. 613–623.

[247] A. Weinstein, *On the Sturm-Liouville theory and the eigenvalues of intermediate problems*, Numerische Mathematik, 5 (1963), pp. 238–245.

[248] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, and M. Schütz, *Molpro: a general-purpose quantum chemistry program package*, Wiley Interdisciplinary Reviews: Computational Molecular Science, 2 (2012), pp. 242–253.

[249] H.-J. Werner and W. Meyer, *A quadratically convergent multiconfiguration-self-consistent field method with simultaneous optimization of orbitals and CI coefficients*, The Journal of Chemical Physics, 73 (1980), pp. 2342–2356.

[250] H. Xie, M. Yue, and N. Zhang, *Fully computable error bounds for eigenvalue problem*, 2016, `https://arxiv.org/pdf/1601.01561v1.pdf`.

[251] J. Xu and A. Zhou, *A two-grid discretization scheme for eigenvalue problems*, Mathematics of Computation, 70 (1999), pp. 17–26.

[252] C. Yang, J. C. Meza, B. Lee, and L.-W. Wang, *KSSOLV–a MATLAB toolbox for solving the Kohn-Sham equations*, ACM Transactions on Mathematical Software, 36 (2009), pp. 1–35.

[253] Y. Yang, J. Han, H. Bi, and Y. Yu, *The lower/upper bound property of the Crouzeix–Raviart element eigenvalues on adaptive meshes*, J. Sci. Comput., 62 (2015), pp. 284–299.

[254] M. C. Zerner and M. Hehenberger, *A dynamical damping scheme for converging molecular scf calculations*, Chemical Physics Letters, 62 (1979), pp. 550–554.

[255] A. Zhou, *Finite dimensional approximations for the electronic ground state solution of a molecular system*, Mathematical Methods in the Applied Sciences, 30 (2007), pp. 429–447.