Some mathematical models in quantum chemistry and uncertainty quantification

Virginie Ehrlacher

UNIVERSITÉ PARIS-EST

Thèse présentée pour obtenir le grade de

Docteur de l'Université Paris-Est

Spécialité: Mathématiques

par

Virginie EHRLACHER

Ecole Doctorale : Mathématiques et Sciences et Technologies de l'Information et de la Communication

Quelques modèles mathématiques en chimie quantique et propagation d'incertitudes

Thèse soutenue le 12 juillet 2012 devant le jury composé de:

Eric Cancès	*Directeur de thèse*
Tony Lelièvre	*Co-Directeur de thèse*
Albert Cohen	*Rapporteur*
Gero Friesecke	*Rapporteur*
Patrick Joly	*Examinateur*
Patrice Hauret	*Examinateur*
Yvon Maday	*Examinateur*
Endre Süli	*Examinateur*

Remerciements

En premier lieu, je tiens à exprimer toute ma gratitude à Eric Cancès et Tony Lelièvre, dont j'ai pu d'abord apprécier le talent pédagogique comme professeurs au cours de ma scolarité, et qui ont accepté d'encadrer ce travail de thèse. Leurs qualités scientifiques et humaines exceptionnelles, leur enthousiasme pour la recherche et leur grande disponibilité m'ont beaucoup apporté au cours de ces trois ans. Je souhaiterais remercier également Yvon Maday pour le temps qu'il a consacré et les nombreuses discussions que nous avons eues, je suis très heureuse qu'il ait accepté de faire partie du jury. Je voudrais aussi témoigner ma reconnaissance à Claude Le Bris pour son soutien et ses conseils durant la thèse.

J'aimerais remercier l'entreprise Michelin, et plus particulièrement Patrice Hauret, pour une collaboration très enrichissante et pour tous les échanges qui en ont découlé.

Je suis très reconnaissante envers Albert Cohen et Gero Friesecke d'avoir accepté de rapporter ce travail, ainsi qu'à Patrick Joly et à Endre Süli d'avoir bien voulu faire partie du jury.

Je remercie Jean-François Guillemoles, Laurent Lombez et Amaury Delamarre, du laboratoire IRDEP, avec qui j'ai eu le grand plaisir de travailler sur des thématiques liées à la photoluminescence de cellules photovoltaïques.

Enfin, je souhaiterais remercier les autres chercheurs avec qui j'ai eu le plaisir d'échanger, Gabriel Stoltz, Frédéric Legoll, Ismaïla Dabo, Alexandre Ern, Philippe Courtier, Olivier Le Maitre, Matthieu Lewin, Adrien Leygue, Alexeï Lozinski, Anthony Nouy. Je souhaiterais remercier Kathrin Smetana ainsi que l'ensemble des membres du groupe de Mario Ohlberger à Münster, pour leur accueil ainsi que pour les échanges que nous avons eus á propos des algorithmes gloutons et des problèmes non symétriques. Merci également à Boris Khoromskij, Venera Khoromskaia, Mike Espig et Wolfgang Hackbusch pour l'invitation à Leipzig et les discussions très intéressantes que nous avons eues sur les formats de tenseurs. Enfin, je souhaiterais témoigner toute ma reconnaissance à Aihui Zhou, ainsi qu'à toute son équipe, de m'avoir accueillie pendant trois semaines à la Chinese Academy of Sciences à Pékin.

Le CERMICS est un endroit exceptionnel pour effectuer un doctorat et la bonne ambiance qui y règne y est pour beaucoup! Mes remerciements s'adressent tout d'abord aux doctorants qui ont eu la patience de partager le bureau B411 avec moi, Arnaud Anantharaman, Ronan Costaouec, Kimiya Minoukadeh, Salma Lahbabi, Nahia Mourad, ainsi que Florian Thomines en tant que membre d'honneur.

Un gros merci également à tous les autres, pour leur bonne humeur et leur gentillesse en toutes circonstances, Julie, Laurent, Raphaël, Abdel, Maxence, José, Olivier, Patrick, David D., David P., David B., David G. 1, David G. 2, Julien, Vincent, Adela, Christelle, Rémi, Fabien, Matthew, William, Cristina, Yan Li, Nadia, Nancy, Sébastien, Marie.

Bravo à Julien Sabin pour l'organisation des séances du groupe de travail des thésards en chimie quantique, et pour les discussions très intéressantes qui s'y sont

déroulées. Merci également à Séverine, Simona, Loïc, Antoine, Jérémy et Gaspard.

Ich bedanke auch Christian Mendl, mit dem ich gern zwei Monaten lang gearbeitet habe.

Un grand merci à Catherine, Nathalie et Martine pour leur grande efficacité pour toutes les formalités administratives. Je voudrais également remercier Jacques Daniel pour son aide précieuse pour tous les problèmes informatiques rencontrés pendant trois ans.

Mes parents et mon frère Charles, m'ont toujours soutenu, que ce soit avant ou pendant ce travail de thèse. J'aimerais les en remercier tous les trois du fond du coeur. Enfin, merci Edouard pour ta patience, ta compréhension, ton écoute, pour tes idées aussi qui m'ont parfois aidée à avancer, et surtout merci de supporter mes diverses absences...

A Edouard,
A mes parents,
A mon frère Charles.

Quelques modèles mathématiques en chimie quantique et propagation d'incertitudes

Ce travail comporte deux volets.

Le premier concerne l'étude de défauts locaux dans des matériaux cristallins. Le chapitre 1 donne un bref panorama des principaux modèles utilisés en chimie quantique pour le calcul de structures électroniques.

Dans le chapitre 2, nous présentons un modèle variationnel exact qui permet de décrire les défauts locaux d'un cristal périodique dans le cadre de la théorie de Thomas-Fermi-von Weiszäcker. Celui-ci est justifié à l'aide d'arguments de limite thermodynamique. On montre en particulier que les défauts modélisés par cette théorie ne peuvent pas être chargés électriquement.

Les chapitres 3 et 4 de cette thèse traitent du phénomène de pollution spectrale. En effet, lorsqu'un opérateur est discrétisé, il peut apparaître des valeurs propres parasites, qui n'appartiennent pas au spectre de l'opérateur initial. Dans le chapitre 3, nous montrons que des méthodes d'approximation de Galerkin via une discrétisation en éléments finis pour approcher le spectre d'opérateurs de Schrödinger périodiques perturbés sont sujettes au phénomène de pollution spectrale. Par ailleurs, les vecteurs propres associés aux valeurs propres parasites peuvent être interprétés comme des états de surface. Nous prouvons qu'il est possible d'éviter ce problème en utilisant des espaces d'éléments finis augmentés, construits à partir des fonctions de Wannier associées à l'opérateur de Schrödinger périodique non perturbé. On montre également que la méthode dite de supercellule, qui consiste à imposer des conditions limites périodiques sur un domaine de simulation contenant le défaut, ne produit pas de pollution spectrale. Dans le chapitre 4, nous établissons des estimations d'erreur a priori pour la méthode de supercellule. En particulier, nous montrons que l'erreur effectuée décroît exponentiellement vite en fonction de la taille de la supercellule considérée.

Un deuxième volet concerne l'étude d'algorithmes gloutons pour résoudre des problèmes de propagation d'incertitudes en grande dimension. Le chapitre 5 de cette thèse présente une introduction aux méthodes numériques classiques utilisées dans le domaine de la propagation d'incertitudes, ainsi qu'aux algorithmes gloutons. Dans le chapitre 6, nous prouvons que ces algorithmes peuvent être appliqués à la minimisation de fonctionnelles d'énergie fortement convexes non linéaires et que leur vitesse de convergence est exponentielle en dimension finie. Nous illustrons ces résultats par la résolution de problèmes de l'obstacle avec incertitudes via une formulation pénalisée.

Some mathematical models in quantum chemistry and uncertainty quantification

The contributions of this thesis work are two fold.

The first part deals with the study of local defects in crystalline materials. Chapter 1 gives a brief overview of the main models used in quantum chemistry for electronic structure calculations.

In Chapter 2, an exact variational model for the description of local defects in a periodic crystal in the framework of the Thomas-Fermi-von Weiszäcker theory is presented. It is justified by means of thermodynamic limit arguments. In particular, it is proved that the defects modeled within this theory are necessarily neutrally charged.

Chapters 3 and 4 are concerned with the so-called spectral pollution phenomenon. Indeed, when an operator is discretized, spurious eigenvalues which do not belong to the spectrum of the initial operator may appear. In Chapter 3, we prove that standard Galerkin methods with finite elements discretization for the approximation of perturbed periodic Schrödinger operators are prone to spectral pollution. Besides, the eigenvectors associated with spurious eigenvalues can be characterized as surface states. It is possible to circumvent this problem by using augmented finite element spaces, constructed with the Wannier functions of the periodic unperturbed Schrödinger operator. We also prove that the supercell method, which consists in imposing periodic boundary conditions on a large simulation domain containing the defect, does not produce spectral pollution. In Chapter 4, we give a priori error estimates for the supercell method. It is proved in particular that the rate of convergence of the method scales exponentially with respect to the size of the supercell.

The second part of this thesis is devoted to the study of greedy algorithms for the resolution of high-dimensional uncertainty quantification problems. Chapter 5 presents the most classical numerical methods used in the field of uncertainty quantification and an introduction to greedy algorithms. In Chapter 6, we prove that these algorithms can be applied to the minimization of strongly convex nonlinear energy functionals and that their convergence rate is exponential in the finite-dimensional case. We illustrate these results on obstacle problems with uncertainty via penalized formulations.

Quelques modèles mathématiques en chimie quantique et propagation d'incertitudes

Cette thèse est scindée en deux sujets.

Une première partie traite du travail qui a été effectué pour le calcul de structures électroniques en chimie quantique. Le chapitre 1 présente les principaux modèles utilisés pour les molécules, les cristaux parfaits et les cristaux avec défauts locaux, ainsi que le contexte dans lequel s'inscrivent les résultats obtenus au cours de cette thèse. Un intérêt particulier est porté à la modélisation des cristaux, ainsi qu'aux méthodes numériques utilisées pour simuler leurs propriétés.

Le chapitre 2 présente des résultats obtenus avec Eric Cancès sur la modélisation d'un défaut local dans un matériau cristallin dans le cadre de la théorie de Thomas-Fermi-von Weiszäcker. Ce travail a fait l'objet d'un article *Local defects are always neutral in the Thomas-Fermi-von Weiszäcker theory* qui a été publié dans le journal *Archive for Rational Mechanics and Analysis*. Le but est de dériver un modèle variationnel exact dans le cadre de cette théorie afin de décrire la réponse de la densité de charge électronique du cristal à une perturbation de la densité de charge des noyaux du cristal, initialement parfaitement périodique. L'idée pour y parvenir est identique à celle utilisée dans l'article de Cancès, Deleurence et Lewin, *A new approach to the modeling of local defects in crystals: The reduced Hartree-Fock case, Comm. Math. Phys., 2008*, où une démarche similaire est menée dans le cadre du modèle Hartree-Fock réduit. Il s'agit de soustraire l'énergie totale du cristal parfait à l'énergie totale du cristal perturbé. Bien sûr, ces deux quantités sont infinies et donner un sens rigoureux à cette soustraction n'est pas évident. Ceci est fait dans le chapitre 2 au moyen d'une limite thermodynamique. Par ailleurs, il est prouvé que le modèle de Thomas-Fermi-von Weiszäcker ne permet pas d'expliquer l'existence de défauts chargés électriquement, contrairement au cas de la théorie de Hartree. Ceci montre que ce modèle n'est pas adapté pour modéliser les défauts d'un matériau semiconducteur, qui peuvent avoir une charge électrique non nulle, mais uniquement les défauts au sein d'un matériau métallique. Des propriétés supplémentaires de la réponse électronique du cristal sont obtenues dans le cas particulier où le cristal parfait de référence est considéré comme un *jellium*, c'est-à-dire que la densité électronique et nucléaire de ce cristal sont supposées être constantes.

Les chapitres 3 et 4 de la thèse sont consacrés à l'étude du phénomène de pollution spectrale, plus particulièrement pour les opérateurs de Schrödinger périodiques perturbés. Il s'agit d'un travail commun avec Eric Cancès et Yvon Maday qui a donné lieu à deux articles, *Periodic Schrödinger operators with local defects and spectral pollution*, qui a été accepté pour publication dans le journal *SIAM Journal*

on Numerical Analysis et *Non-consistent approximations of self-adjoint eigenproblems: Application to the supercell method* , qui a été soumis au journal *Numerische Mathematik*. Let $d \in \mathbb{N}^*$. Un opérateur de Schrödinger périodique perturbé est de la forme $H = -\frac{1}{2}\Delta + V_{\text{per}} + W$ sur $L^2(\mathbb{R}^d)$, où Δ est l'opérateur de Laplace, V_{per} est un potentiel périodique qui appartient à $L^2_{\text{loc}}(\mathbb{R}^d)$ et W une perturbation de ce potentiel telle que $W \in L^\infty(\mathbb{R}^d)$ et $W(x) \underset{|x|\to\infty}{\longrightarrow} 0$. Dans le cadre de théories de champ moyen, comme Hartree ou Kohn-Sham, la structure électronique d'un cristal avec défaut est entièrement caractérisée par la décomposition spectrale d'un opérateur de ce type. L'opérateur de Schrödinger périodique non perturbé $H_{\text{per}} = -\frac{1}{2}\Delta + V_{\text{per}}$ forme un opérateur auto-adjoint, borné inférieurement, de domaine $H^2(\mathbb{R}^d)$ et dont le spectre est purement continu. Ce dernier est composé d'une réunion d'intervalles de $\mathbb{R}$, appelés *bandes*. Un intervalle de $\mathbb{R}$ situé entre deux bandes est appelé un *gap spectral*. L'opérateur perturbé H est aussi un opérateur auto-adjoint, borné inférieurement, de domaine $H^2(\mathbb{R}^d)$. D'après le théorème de Weyl, comme W est une perturbation compacte de l'opérateur H_{per}, les spectres essentiels des opérateurs H et H_{per} sont identiques. Par contre, le spectre de H peut comporter des valeurs propres discrètes, soit inférieures à l'infimum du spectre essentiel de H_{per}, soit situées dans un gap spectral.

Dans les chapitres 3 et 4, on étudie certaines méthodes numériques pour calculer ces valeurs propres discrètes, ainsi que les vecteurs propres associés. Le principal problème rencontré est le phénomène de pollution spectrale. En effet, lorsque l'opérateur H est approximé par une suite d'opérateurs discrets, il peut arriver qu'il existe une suite de valeurs propres de ces opérateurs discrets qui converge vers une limite qui n'appartient pas au spectre de H. Cette limite est alors appelée une valeur propre parasite de H. Il s'agit d'un phénomène bien connu et qui peut être constaté lorsque l'on utilise la méthode la plus naturelle pour définir la suite d'opérateurs discrétisés, c'est-à-dire une méthode de Galerkin sur un domaine de simulation fini avec une discrétisation en éléments finis et des conditions aux bords de Dirichlet. Ce phénomène avait déjà été constaté auparavant, notamment par Boulton et Levitin. Dans le chapitre 3, nous montrons que les vecteurs propres des opérateurs discrétisés associés à une suite de valeurs propres convergeant vers une valeur propre parasite de l'opérateur H sont localisés au bord du domaine de simulation et peuvent ainsi être interprétés comme des états de surface. Nous disposons par ailleurs d'une caractérisation plus précise des ces valeurs propres et vecteurs propres parasites en dimension 1.

D'autres méthodes numériques sont étudiées, en particulier la méthode de supercellule. Celle-ci consiste à imposer des conditions de bord périodiques sur le domaine de simulation au lieu de conditions de Dirichlet. Dans le chapitre 3, il est montré que lorsque cette méthode est utilisée avec une discrétisation en modes de Fourier, il n'y a pas de pollution spectrale: la suite des spectres des opérateurs discrétisés converge exactement vers le spectre de l'opérateur H. Par ailleurs des estimations d'erreur a priori de la méthode sont prouvées dans le chapitre 4. En particulier, l'erreur liée à la taille de la supercellule décroît exponentiellement vite.

Une dernière méthode numérique, basée sur l'utilisation d'espaces d'éléments finis augmentés, est étudiée dans le chapitre 3. Cette méthode consiste à enrichir les espaces de discrétisation d'éléments finis avec conditions de bord de Dirichlet avec

des fonctions de Wannier associées à l'opérateur non perturbé. Nous prouvons que l'utilisation de ces espaces d'éléments finis augmentés permet d'éviter le phénomène de pollution spectrale à l'intérieur d'un des gaps spectraux de l'opérateur H.

Cette thèse comporte également un second volet, axé sur l'utilisation d'algorithmes gloutons pour résoudre des problèmes de propagation d'incertitudes en grande dimension. Le chapitre 5 est conascré à l'état de l'art des méthodes numériques utilisées dans le domaine de la propagation d'incertitudes. Nous présentons également le principe des algorithmes gloutons, qui sont plus particulièrement étudiés dans cette thèse, leur lien avec la *Progressive Generalized Decomposition* et leur intérêt pour résoudre des problèmes en grande dimension. Une brève liste des différentes méthodes numériques utilisées actuellement pour éviter la malédiction de la dimensionalité est détaillée.

Dans le chapitre 6, on étudie et prouve la convergence d'un algorithme glouton pour résoudre un problème de minimisation d'une fonctionnelle fortement convexe, non-quadratique, définie sur un espace de fonctions dépendant d'un grand nombre de variables. On prouve également que dans le cas où les problèmes considérés sont définis sur des espaces de dimension finie, alors la vitesse de convergence de l'algorithme est exponentielle. Ce travail a été motivé par une collaboration avec l'entreprise Michelin, dont le but est de résoudre un problème de l'obstacle avec des coefficients aléatoires. Nous illustrons la convergence de l'algorithme sur ce problème, en utilisant une formulation pénalisée de l'inéquation variationnelle considérée.

Contents

Part I

Quantum chemistry

Chapter 1

Electronic structure calculations in solid state physics

1.1 Introduction

The aim of this chapter is to give a brief description of the main models used in quantum chemistry and solid state physics, for finite systems and periodic crystals with or without local defects. We will present the most common numerical methods used in the latter field. It is assumed here that the reader already has some knowledge about quantum mechanics. If not, a very good physical introduction can be found in [100]. The reader can also find a mathematical description of the models that are presented here in [42]. The contributions of this thesis are summarized in Sections 1.4.2 and 1.6.2.

There exists a wide variety of models to describe the electronic structure of materials at the atomic scale. We will focus here on nonlinear mean-field models on the one hand, and on linear empirical models on the other hand. In the rest of this Section, we will present the reference ab-initio model of nonrelativic quantum theory for a finite system, namely the many-body Schrödinger model. Unfortunately, this model leads to a very high-dimensional problem, except when the number of electrons is very small, and direct numerical approaches cannot be carried out in practice.

Nonlinear mean-field models, which are approximations of the many-body Schrödinger model, are easier to deal with from a numerical point of view. Among them, the Hartree-Fock (HF) and Density Functional Theory (DFT) models are the most widely used. The presentation of these models for finite systems, periodic crystals and crystals with local defects is done respectively in Sections 1.2, 1.3 and 1.4 of this chapter. Section 1.5 will be devoted to the description of linear semi-empirical models. Lastly, a short review of the numerical methods used for the discretization and the resolution of these problems is detailed in Section 1.6.

The contributions of this thesis work are two fold. A first part of the work is dedicated to the determination of an exact variational model to describe local defects in a host periodic crystalline material within the Thomas-Fermi-von Weiszäcker framework. This will be detailed in Section 1.4.2 and Chapter 2, which present the

results and proofs contained in the article [38] published in *Archive for Rational Mechanics and Analysis*.

A second part will be devoted to the study of numerical methods for linear semi-empirical mean-field models for crystals with local defects, and in particular to the study of the so-called *spectral pollution* phenomenon. Section 1.6.2 presents a short introduction to this issue along with a summary of our contributions. Chapters 3 and 4 detail all our theoretical results and proofs which were gathered in two articles [40] and [41] currently in review.

Many-body Schrödinger model

In the sequel, we will work in atomic units, which implies that

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

where $\hbar$ is the Planck constant, e the elementary charge, m_e the mass of the electron, and ε_0 the electric permittivity of the void.

In the vast majority of quantum chemistry computations, the nuclei are assumed to behave like classical particles. This approximation is called the *Born-Oppenheimer* approximation, and we will always work in this framework from now on.

Let us consider a physical system composed of

- M nuclei, that are assumed to be point charges, whose positions in $\mathbb{R}^3$ and electric charges are denoted by $R_1, \cdots, R_M$, and $z_1, \cdots, z_M$ respectively;
- N electrons, which are not described by their positions or velocities in $\mathbb{R}^3$, but, as they are considered as quantum particles, by their wavefunction $\psi_e(x_1, \cdots, x_n)$, where for all $1 \leq i \leq N$, x_i is a vector of $\mathbb{R}^3$ and $\psi_e \in \bigotimes_{i=1}^N L^2(\mathbb{R}^3) \approx L^2(\mathbb{R}^{3N})$ such that $\|\psi_e\|_{L^2(\mathbb{R}^{3N})} = 1$. Since electrons are fermionic particles, the Pauli exclusion principle implies that the wavefunction ψ_e must be antisymmetric with respect to the exchange of two particles. In other words, for any permutation p of the set $\{1, \cdots, N\}$, it holds that

  $$\psi_e(x_{p(1)}, \cdots, x_{p(N)}) = \varepsilon(p)\psi_e(x_1, \cdots, x_N),$$

  where $\varepsilon(p)$ is the signature of the permutation p. We will denote by $\bigwedge_{i=1}^N L^2(\mathbb{R}^3)$ the set of antisymmetric functions of $\bigotimes_{i=1}^N L^2(\mathbb{R}^3)$.

For any normalized $\psi_e \in \bigwedge_{i=1}^N L^2(\mathbb{R}^3)$, we introduce the *density operator* Γ_{ψ_e} which is the trace-class operator on $L^2(\mathbb{R}^{3N})$ defined by (using the bracket notation):

$$\Gamma_{\psi_e} := |\psi_e\rangle\langle\psi_e|.$$

The *density matrix* of ψ_e is denoted by $\tau_{\psi_e}(x, y)$ and defined for almost all $x, y \in \mathbb{R}^3$ as

$$\tau_{\psi_e}(x, y) = N \int_{\mathbb{R}^{3(N-1)}} \psi_e(x, x_2, \cdots, x_N)\psi_e(y, x_2, \cdots, x_N)\, dx_2 \cdots dx_N.$$

It is the kernel of a trace-class operator γ_{ψ_e} on $L^2(\mathbb{R}^3)$, called the *order 1 density operator* of ψ_e. Lastly, let us introduce the *electronic density* ρ_{ψ_e} associated with the wavefunction ψ_e, defined for almost all $x \in \mathbb{R}^3$ as

$$\rho_{\psi_e}(x) := N \int_{\mathbb{R}^{3(N-1)}} |\psi_e(x, x_2, \cdots, x_N)|^2 \, dx_2 \cdots dx_N = \tau_{\psi_e}(x, x).$$

We have voluntarily omitted to write the dependence of the wavefunction ψ_e on the spin variables. Actually, an electron is an elementary particle of spin 1/2. Thus, its spin variable, denoted by σ, can only take two different values, denoted by $|+\rangle$ (spin up) and $|-\rangle$ (spin down). Let $\Sigma := \{|+\rangle, \, |-\rangle\}$. Taking into account the spin would amount to considering a wavefunction ψ_e belonging to the space

$$\bigwedge_{i=1}^{N} L^2(\mathbb{R}^3 \times \Sigma).$$

For the sake of simplicity, we will ignore the spin in this thesis work.

The usual Born-Oppenheimer approximation consists in assuming that the nuclei are point classical particles that are evolving in the effective potential

$$W(R_1, \cdots, R_M) := I(R_1, \cdots, R_M) + \sum_{1 \leq k < l \leq M} \frac{z_k z_l}{|R_k - R_l|}.$$

There are two terms in the definition of the potential W:

- The term $\sum_{1 \leq k < l \leq M} \frac{z_k z_l}{|R_k - R_l|}$ is due to the repulsive Coulomb forces between the nuclei;
- The term $I(R_1, \cdots, R_M)$ corresponds to the effective potential created by the electrons.

In particular, the molecular configurations of minimal energy of the system are obtained by solving

$$\inf_{(R_1, \cdots R_M) \in \mathbb{R}^{3M}} W(R_1, \cdots, R_M). \tag{1.1}$$

Problem (1.1) is called the *geometry optimization* problem.

The value of the potential $I(R_1, \cdots R_M)$ for given fixed positions of the nuclei is obtained by solving the *electronic* problem:

$$I(R_1, \cdots, R_M) = \inf \left\{ \left\langle \psi_e, H_e^{(R_1, \cdots R_M)} \psi_e \right\rangle, \ \psi_e \in \mathcal{H}_e, \ \|\psi_e\|_{L^2(\mathbb{R}^{3N})} = 1 \right\}, \tag{1.2}$$

where

$$H_e^{(R_1, \cdots, R_M)} := -\sum_{i=1}^{N} \frac{1}{2} \Delta_{x_i} + \sum_{i=1}^{N} V(x_i) + \sum_{1 \leq i < j \leq N} \frac{1}{|x_i - x_j|},$$

$\mathcal{H}_e := \bigwedge_{i=1}^{N} H^1(\mathbb{R}^3)$ denotes the set of antisymmetric functions of $\bigotimes_{i=1}^{N} H^1(\mathbb{R}^3)$, and for almost all $x \in \mathbb{R}^3$,

$$V(x) = -\sum_{k=1}^{M} \frac{z_k}{|x - R_k|}.$$

Indeed, $I(R_1, \cdots, R_M)$ corresponds to the fundamental energy of the electronic Hamiltonian $H_e^{(R_1,\cdots,R_M)}$ on the space of all admissible wavefunctions. This Hamiltonian $H_e^{(R_1,\cdots,R_M)}$ is composed of three terms:

- $-\sum_{i=1}^N \frac{1}{2}\Delta_{x_i}$ corresponds to the kinetic energy of the electrons;
- $\sum_{i=1}^N V(x_i)$ models the Coulomb interactions induced by the nuclei on the electrons;
- $\sum_{1\leq i<j\leq N} \frac{1}{|x_i-x_j|}$ denotes the Coulomb interactions between electrons.

In the sequel, we will focus on the resolution of the *electronic* problem (1.2) for a given nuclear configuration $(R_1, \cdots, R_M)$ and denote by H_e the electronic Hamiltonian $H_e^{(R_1,\cdots,R_M)}$ in order to simplify the notation. Unfortunately, a direct approach for the resolution of (1.2) amounts to discretizing functions over the space $\mathbb{R}^{3N}$, which is impossible to do in practice, except when the number of electrons N in the system is very small. This is the reason why nonlinear or linear mean-field models have been introduced in Quantum Chemistry in order to reduce the dimension of the discretized problem while providing a good approximation of the many-body Schrödinger model. Any model which will be presented below writes as a minimization problem of a given energy functional. A minimizer is called a *ground state* and the corresponding value of the energy functional will be called the *ground state energy*.

1.2 Nonlinear mean-field models for finite systems

1.2.1 The Hartree-Fock model

Slater determinants

The Hartree-Fock model [139, 142] is a variational approximation of the electronic problem (1.2), which consists in reducing the minimization set of problem (1.2) to the set of wavefunctions ψ_e that can be written as a *Slater determinant*. Recall that a Slater determinant is a function of the form

$$\psi_e(x_1, \cdots, x_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(x_1) & \cdots & \phi_1(x_N) \\ \vdots & \ddots & \vdots \\ \phi_N(x_1) & \cdots & \phi_N(x_N) \end{vmatrix}$$

where $(\phi_1, \cdots, \phi_N)$ is an orthonormal family of $L^2(\mathbb{R}^3)$. Such a function is in $\mathcal{H}_e = \bigwedge_{i=1}^N H^1(\mathbb{R}^3)$ if and only if each ϕ_i is in $H^1(\mathbb{R}^3)$. In quantum chemistry, the functions $(\phi_1, \cdots, \phi_N)$ are called *molecular orbitals*.

Let us denote by

$$\mathcal{W}_N := \left\{ \Phi = (\phi_i)_{1\leq i\leq N},\ \phi_i \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij},\ 1 \leq i,j \leq N \right\}$$

the set of all families of N molecular orbitals and

$$\mathcal{S}_N := \left\{ \psi_e \in \mathcal{H}_e,\ \exists \Phi = (\phi_i)_{1\leq i\leq N} \in \mathcal{W}_N,\ \psi_e = \frac{1}{\sqrt{N!}}\det(\phi_i(x_j)) \right\},$$

the set of all functions of $\mathcal{H}_e$ that can be written as a Slater determinant of finite energy. Then, the Hartree-Fock problem reads

$$I^{HF} = \inf\{\langle \psi_e, H_e\psi_e\rangle,\ \psi_e \in \mathcal{S}_N\}. \tag{1.3}$$

It holds that for all $\Phi = (\phi_i)_{1\leq i\leq N} \in \mathcal{W}_N$, if $\psi_e \in \mathcal{S}_N$ is the Slater determinant associated with Φ, then

$$\langle \psi_e, H_e\psi_e\rangle = E^{HF}(\Phi),$$

where

$$\begin{aligned} E^{HF}(\Phi) \ &= \ \sum_{i=1}^{N} \frac{1}{2}\int_{\mathbb{R}^3} |\nabla\phi_i|^2 + \int_{\mathbb{R}^3} \rho_\Phi V \\ &\quad + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho_\Phi(x)\rho_\Phi(y)}{|x-y|}\,dx\,dy - \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{|\tau_\Phi(x,y)|^2}{|x-y|}\,dx\,dy. \end{aligned}$$

In the above expression,

$$\rho_\Phi(x) = \rho_{\psi_e}(x) = \sum_{i=1}^{N} |\phi_i(x)|^2$$

and

$$\tau_\Phi(x,y) = \tau_{\psi_e}(x,y) = \sum_{i=1}^{N} \phi_i(x)\phi_i(y).$$

The Hartree-Fock problem therefore reads

$$I^{HF} = \inf\{E^{HF}(\Phi),\ \Phi \in \mathcal{W}_N\}. \tag{1.4}$$

It is typically an ab-initio model since no empirical parameter needs to be tuned. The first term in the expression of $E^{HF}(\Phi)$ represents the kinetic energy of the electrons, the second the electrostatic interaction between nuclei and electrons. The third term corresponds to the classical Coulomb energy of the electronic charge distribution ρ_Φ. Lastly, the term $-\frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{|\tau_\Phi(x,y)|^2}{|x-y|}\,dx\,dy$ is called the *exchange energy* term and its origin is purely quantum, since it originates from the antisymmetry of the wavefunction. Since the minimization set $\mathcal{S}_N$ is smaller than the minimization set of the original electronic problem (1.2), the fundamental energy given by (1.3), or equivalently by (1.4), is an upper bound of the fundamental energy of (1.2). The difference between these two energies is called the *correlation energy*.

Let us also point out that the restriction of the minimization set to the Slater determinants $\mathcal{S}_N$ has disadvantages. Actually, the energy functional $E^{HF}(\Phi)$ is not quadratic with respect to Φ, whereas the energy functional $\langle \psi_e, H_e\psi_e\rangle$ appearing in (1.2) is quadratic with respect to the wavefunction ψ_e.

The existence of a minimizer to problem (1.4) was proved for neutral or positively charged systems, i.e. when $Z := \sum_{k=1}^{M} z_k \geq N$ (see [139, 142]). The question of uniqueness is an intricate one and some answers for closed-shell atoms may be found in [99].

In the case when there exists a minimizer $\Phi = (\phi_i)_{1\leq i\leq N} \in \mathcal{W}_N$ of (1.4), up to replacing Φ by $U\Phi$ with a well-chosen unitary matrix of $\mathbb{R}^{N\times N}$, the functions $(\phi_i)_{1\leq i\leq N}$ satisfy the following set of nonlinear eigenvalue equations

$$H_\Phi^{HF}\phi_i = \lambda_i\phi_i,\ 1 \leq i \leq N,$$

where $\lambda_1 \leq \cdots \leq \lambda_N$ are the lowest N eigenvalues (counting multiplicities) of the operator H_Φ^{HF} on $L^2(\mathbb{R}^3)$ which is defined as follows

$$H_\Phi^{HF}\psi = -\frac{1}{2}\Delta\psi + V\psi + \left(\rho_\Phi * |\cdot|^{-1}\right)\psi - \int_{\mathbb{R}^3} \frac{\gamma_\Phi(\cdot,y)}{|\cdot - y|}\psi(y)\,dy.$$

The essential spectrum of H_Φ^{HF} is equal to $[0,+\infty)$ and H_Φ^{HF} possesses at least N negative eigenvalues.

Density operators

For any Hilbert space $\mathcal{H}$, we denote by $\mathcal{B}(\mathcal{H})$ (respectively $\mathcal{S}(\mathcal{H})$, $\mathfrak{S}_1(\mathcal{H})$ and $\mathfrak{S}_2(\mathcal{H})$) the set of the operators on $\mathcal{H}$ which are bounded (respectively self-adjoint, trace-class and Hilbert-Schmidt).

The Hartree-Fock problem can also be rewritten using the *order 1 density operator* formalism. For $\Phi = (\phi_i)_{1\leq i\leq N} \in \mathcal{W}_N$, if $\psi_e \in \mathcal{S}_N$ is the Slater determinant associated with Φ, then

$$\gamma_\Phi = \gamma_{\psi_e} = \sum_{i=1}^{N} |\phi_i\rangle\langle\phi_i|.$$

For all $\Phi \in \mathcal{W}_N$, $\gamma_\Phi \in \mathfrak{S}_1(L^2(\mathbb{R}^3)) \cap \mathcal{S}(L^2(\mathbb{R}^3))$, τ_Φ is the kernel of γ_Φ and

$$\mathrm{Tr}(\gamma_\Phi) = \int_{\mathbb{R}^3} \rho_\Phi = N.$$

The Hartree-Fock problem can be rewritten in an equivalent manner as a minimization problem over density operators γ [136] as follows:

$$I^{HF} = \inf\left\{E^{HF}(\gamma),\ \gamma \in \mathcal{L}_N\right\}, \tag{1.5}$$

where

$$\begin{aligned}\mathcal{L}_N \ :=\ & \big\{\gamma \in \mathfrak{S}_1(L^2(\mathbb{R}^3)) \cap \mathcal{S}(L^2(\mathbb{R}^3)),\ 0 \leq \gamma \leq 1,\ \mathrm{Tr}(\gamma) = N,\\ & \mathrm{Tr}\left((1-\Delta)^{1/2}\gamma(1-\Delta)^{1/2}\right) < +\infty\big\},\end{aligned}$$

$$\begin{aligned}E^{HF}(\gamma) \ =\ & \frac{1}{2}\mathrm{Tr}\left(-\Delta\gamma\right) + \int_{\mathbb{R}^3} V\rho_\gamma + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho_\gamma(x)\rho_\gamma(y)}{|x-y|}\,dx\,dy\\ & -\frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{|\tau_\gamma(x,y)|^2}{|x-y|}\,dx\,dy,\end{aligned}$$

$\tau_\gamma(x,y)$ is the kernel of γ, and ρ_γ is the density associated with γ, namely $\rho_\gamma(x) = \tau_\gamma(x,x)$. In the expression of $E^{HF}(\gamma)$, we have used the following notation

$$\mathrm{Tr}(-\Delta\gamma) := \mathrm{Tr}(|\nabla|\gamma|\nabla|) = \mathrm{Tr}\left((-\Delta)^{1/2}\gamma(-\Delta)^{1/2}\right).$$

Then, γ is a minimizer of (1.5) if and only if $\gamma = \gamma_\Phi$ for some $\Phi \in \mathcal{W}_N$ which is a minimizer of (1.4).

Other models

The *reduced Hartree-Fock* (rHF) model is a simplified version of the Hartree-Fock problem using the density operator formalism. It reads

$$I^{rHF} = \inf\left\{E^{rHF}(\gamma),\ \gamma \in \mathcal{L}_N\right\},$$

where

$$E^{rHF}(\gamma) = \frac{1}{2}\mathrm{Tr}\left(-\Delta\gamma\right) + \int_{\mathbb{R}^3} V\rho_\gamma + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho_\gamma(x)\rho_\gamma(y)}{|x-y|}\,dx\,dy$$

is the Hartree-Fock energy without the exchange-term. For a study of the reduced Hartree-Fock theory, we refer to [171].

As explained above, the Hartree-Fock theory consists in restricting the minimization set of the full electronic problem (1.2) to the set of Slater determinants $\mathcal{S}_N$. One way to improve the model is to take finite linear combinations of Slater determinants as a minimization set. More precisely, the *multiconfigurational self-consistent field* (MCSCF) reads

$$I^{MCSCF} = \inf\{\langle\psi_e, H_e\psi_e\rangle,\ \psi_e \in \mathcal{S}_N^K\},$$

where

$$\mathcal{S}_N^K := \left\{\psi_e = \sum_{I=\{i_1,\cdots,i_n\}\subset\{1,\cdots,K\}} c_I \frac{1}{\sqrt{N!}}\det(\phi_{i_1},\cdots,\phi_{i_N}),\ (\phi_i)_{1\le i\le K} \in \mathcal{W}_K,\ \sum_I c_I^2 = 1\right\},$$

and K is an integer greater than N. The analysis of this model for finite systems was done in [129, 91, 133].

1.2.2 Density Functional Theory

The principle of Density Functional Theory (DFT), and of all the models which are derived from it, is the reformulation of problem (1.2) with the density (and not anymore the wavefunction) as the main variable. The key advantage of this method is that problems are then formulated over the domain $\mathbb{R}^3$ instead of $\mathbb{R}^{3N}$. The theoretical justification of this approach was first done by Hohenberg and Kohn [111] and was later developped by Levy and Lieb [137]. Indeed, the Hohenberg-Kohn theorem [111] states that the energy and the electronic density of the ground state of the electronic problem (1.2) can be found by solving a problem of the form

$$I(R_1,\cdots,R_M) = \inf\left\{F(\rho) + \int_{\mathbb{R}^3}\rho V,\ \rho \in L^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3}\rho = N\right\},$$

where F is a functional of the electronic density ρ.

Let us define

$$\mathcal{F}_N := \left\{\psi_e \in \mathcal{H}_e,\ \|\psi_e\|_{L^2(\mathbb{R}^{3N})} = 1\right\},$$

and rewrite H_e under the following form

$$H_e = H_V = H_1 + \sum_{i=1}^{N} V(x_i),$$

where

$$H_1 := -\sum_{i=1}^{N} \frac{1}{2}\Delta_{x_i} + \sum_{1\leq i<j\leq N} \frac{1}{|x_i - x_j|}, \tag{1.6}$$

in order to highlight the dependence of the Hamiltonian on the potential V. The minimization problem (1.2) can then be rewritten as

$$I(R_1, \cdots, R_M) = I(V) = \inf\left\{\langle\psi_e, H_V\psi_e\rangle,\ \psi_e \in \mathcal{F}_N\right\}. \tag{1.7}$$

We also denote by

$$\mathcal{I}_N := \{\rho,\ \exists\psi_e \in \mathcal{F}_N,\ \rho_{\psi_e} = \rho\}$$

the set of all electronic densities associated with some admissible wavefunction. It is proved in [137] that $\mathcal{I}_N$ can be characterized equivalently as

$$\mathcal{I}_N = \left\{\rho \geq 0,\ \sqrt{\rho} \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3} \rho = N\right\}.$$

The DFT relies on the following elementary calculus [111, 137]:

$$\begin{aligned} I(V) &= \inf\left\{\langle\psi_e, H_V\psi_e\rangle,\ \psi_e \in \mathcal{F}_N\right\} \\ &= \inf\left\{\inf\left\{\langle\psi_e, H_1\psi_e\rangle,\ \psi_e \in \mathcal{F}_N,\ \rho_{\psi_e} = \rho\right\} + \int_{\mathbb{R}^3} \rho V,\ \rho \in \mathcal{I}_N\right\} \\ &= \inf\left\{F_{LL}(\rho) + \int_{\mathbb{R}^3} \rho V,\ \rho \in \mathcal{I}_N\right\}, \end{aligned}$$

where

$$F_{LL}(\rho) := \inf\left\{\langle\psi_e, H_1\psi_e\rangle,\ \psi_e \in \mathcal{F}_N,\ \rho_{\psi_e} = \rho\right\}$$

is called the *Levy-Lieb functional*. It is universal in the sense that it does not depend on the molecular system under consideration (which only comes into play through the potential V and the number of electrons N).

Another approach to DFT, with density operators, is detailed below. Let us recall that for any wavefunction $\psi_e \in \mathcal{F}_N$, also called *pure state*, we can associate a density operator Γ_{ψ_e} on $L^2(\mathbb{R}^{3N})$ defined by

$$\Gamma_{\psi_e} := |\psi_e\rangle\langle\psi_e|.$$

The *mixed states* are defined as the set of convex combinations of pure states. They are described by density operators of the form

$$\Gamma = \sum_{i=1}^{+\infty} p_i |\psi_e^i\rangle\langle\psi_e^i|,\ 0 \leq p_i \leq 1,\ \sum_{i=1}^{+\infty} p_i = 1,\ \psi_e^i \in \mathcal{F}_N,\ i \in \mathbb{N}^*. \tag{1.8}$$

Let us denote by $\mathcal{D}_N$ the set of density operators of the form (1.8), which is the convex hull of the set of density operators associated with pure states. The electronic density which corresponds to a density operator Γ of the form (1.8) is then given by

$$\rho_\Gamma(x) = \sum_{i=1}^{+\infty} p_i \rho_{\psi_e^i}(x),$$

where $\rho_{\psi_e^i}$ is the electronic density of the pure state ψ_e^i. It can be easily checked that

$$\begin{aligned}
\mathrm{Tr}(\Gamma) &= \sum_{i=1}^{+\infty} p_i \|\psi_e^i\|^2_{L^2(\mathbb{R}^{3N})} = 1, \\
\mathrm{Tr}(H_1\Gamma) &= \sum_{i=1}^{+\infty} p_i \langle\psi_e^i, H_1\psi_e^i\rangle, \\
\mathrm{Tr}(H_V\Gamma) &= \sum_{i=1}^{+\infty} p_i \langle\psi_e^i, H_1\psi_e^i\rangle + \int_{\mathbb{R}^3} \rho_\Gamma V.
\end{aligned}$$

Actually, the minimization on pure states (1.7) is equivalent to a minimization over mixed states, in other words,

$$I(V) = \inf\left\{\mathrm{Tr}(H_V\Gamma), \Gamma \in \mathcal{D}_N\right\}.$$

Besides,

$$\{\rho,\ \exists\Gamma \in \mathcal{D}_N,\ \rho_\Gamma = \rho\} = \mathcal{I}_N.$$

Thus, with similar arguments as above, it holds that

$$I(V) = \inf\left\{F_L(\rho) + \int_{\mathbb{R}^3} \rho V,\ \rho \in \mathcal{I}_N\right\},$$

where the *Lieb functional* $F_L(\rho)$ is defined by

$$F_L(\rho) := \inf\left\{\mathrm{Tr}(H_1\Gamma),\ \Gamma \in \mathcal{D}_N,\ \rho_\Gamma = \rho\right\}.$$

These two approaches enable us to rewrite (1.2) as a minimization problem on the electronic density rather than on the wavefunction. However, the main drawback of the DFT theory is that there is no explicit expression for the Levy-Lieb and Lieb functionals F_{LL} and F_L. Very recent results by Cotar, Friesecke and Kluppelberg [62] show that in the case of two electrons and in the semiclassical limit, exact expressions of these functionals can be obtained with the use of an optimal transport map associated with a given density ρ. In practice though, approximations of the functionals F_{LL} or F_L have to be used, and most of them derive from exact evaluations of these functionals on reference systems. A wide variety of models exist in the literature, and we only present here two important classes of DFT models: the Thomas-Fermi and the Kohn-Sham type models.

1.2.3 Thomas-Fermi type models

Thomas-Fermi type models belong to the so-called class of *orbital-free* models, which means that the approximate functional $F(\rho)$ is expressed as an explicit function of ρ and its derivatives. The reference system in this case is a homogeneous electron gas. The basic three models of this type are:

- the Thomas-Fermi (TF) model

$$F(\rho) = C_{TF} \int_{\mathbb{R}^3} \rho^{5/3} + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|}\, dx\, dy;$$

- the Thomas-Fermi-von Weiszäcker (TFW) model

$$F(\rho) = C_W \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}|^2 + C_{TF} \int_{\mathbb{R}^3} \rho^{5/3} + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|}\, dx\, dy;$$

- the Thomas-Fermi-Dirac-von Weiszäcker (TFDW) model

$$F(\rho) = C_W \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}|^2 + C_{TF} \int_{\mathbb{R}^3} \rho^{5/3} - C_D \int_{\mathbb{R}^3} \rho^{4/3} + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|}\, dx\, dy;$$

where the values of the Thomas-Fermi constant C_{TF} and of the Dirac constant C_D are given by

$$C_{TF} = \frac{3^{5/3}\pi^{4/3}}{10} \quad \text{and} \quad C_D = \frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}.$$

Several values for the von Weiszäcker constant C_W have been proposed in the literature (see e.g. [80]). Thomas-Fermi type models are not very much used nowadays in quantum chemistry computations. However they are still of interest from a mathematical point of view, since they are simpler but contain some difficulties which are encountered in more realistic quantum models.

The minimization problem associated with Thomas-Fermi type models then writes

$$I^{TFt} = \inf \left\{ F(\rho) + \int_{\mathbb{R}^3} \rho V,\ \rho \geq 0,\ \sqrt{\rho} \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3} \rho = N \right\}. \tag{1.9}$$

Let us point out that, in this formulation, N does not need to be an integer but may be an arbitrary positive real number.

In all these models, the energy functional $F(\rho)$ is the sum of the self-interaction Coulomb energy of the system

$$\frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|}\, dx\, dy$$

and of an approximation of the kinetic energy of the electrons.

The oldest model of this family, the Thomas-Fermi model was proposed by Thomas [175] and Fermi [85] in 1927. The kinetic energy of the electrons is approximated by

$$C_{TF} \int_{\mathbb{R}^3} \rho^{5/3}$$

which is the electronic kinetic energy associated with a system of non-interacting electrons in a homogeneous gas [114, 146]. An exhaustive mathematical analysis of the Thomas-Fermi model may be read in [140]. The Thomas-Fermi model suffers from several major drawbacks. Firstly, it cannot account for the existence of negatively charged ions. Indeed, problem (1.9) admits a unique minimizer if and only if $N \leq Z$. Secondly, atomic bonding in molecules cannot be explained within this theory: the ground state energy of a system with two nuclei continuously decreases as the distance between the two nuclei increases.

The addition of an inhomogeneity correcting term

$$C_W \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}|^2$$

by von Weiszäcker [181] in the expression of the approximate kinetic energy enables one to get rid of these problems. This quantity is also the first order correction to the Thomas-Fermi kinetic energy in a semiclassical approximation to the Hartree-Fock theory [121]. In particular, there exists $N_c \in \mathbb{R}_+$, $N_c > Z$ such that (1.9) admits a unique minimizer if and only if $N \leq N_c$ [28].

In the TFDW model, the Dirac correction

$$-C_D \int_{\mathbb{R}^3} \rho^{4/3}$$

enables one to take into account exchange effects. However, due to this term, the TFDW functional is no more convex, unlike in the case of the TF and TFW theories, and thus the analysis of this problem is more difficult. It can be carried out by means of concentration-compactness arguments that are exposed in [141, 128]. A very good review of this family of models can be found in [135].

1.2.4 Kohn-Sham type models

The reference system for the Kohn-Sham type models [119] is a system of N electrons without interaction. The Hamiltonian H_1, which was defined by (1.6), is then replaced by

$$H_0 = -\sum_{i=1}^{N} \frac{1}{2} \Delta_{x_i}.$$

The Hamiltonian H_0 is then used to obtain a kinetic energy functional, which will have two different forms depending on whether the approach is the one by Levy-Lieb (pure states) or the one by Lieb (mixed states).

In the first case, the *Kohn-Sham functional* is given by

$$\widetilde{T}_{KS}(\rho) = \inf \left\{ \langle \psi_e, H_0 \psi_e \rangle,\ \psi_e \in \mathcal{F}_N,\ \rho_{\psi_e} = \rho \right\}. \tag{1.10}$$

Unfortunately, the function $\widetilde{T}_{KS}$ admits an expression that can be used in practice only in the case when the minimizer in (1.10) is a Slater determinant. It is proved that it is not always the case [137]. Nevertheless, the approach adopted in most

numerical methods is to restrict the minimization set $\mathcal{F}_N$ to the set of Slater determinants $\mathcal{S}_N$. An approximation of the Kohn-Sham functional $\widetilde{T}_{KS}(\rho)$ is considered, namely

$$T_{KS}(\rho) = \inf\left\{\frac{1}{2}\sum_{i=1}^{N}\int_{\mathbb{R}^3}|\nabla\phi_i|^2,\ \Phi=(\phi_i)_{1\leq i\leq N}\in\mathcal{W}_N,\ \rho_\Phi=\sum_{i=1}^{N}|\phi_i|^2=\rho\right\}.$$

This problem of representation of the minimizers is avoided if the kinetic energy functional is derived from the Lieb mixed states formulation. The resulting functional, called the *Janak functional*, reads

$$T_J(\rho) = \inf\left\{\mathrm{Tr}(H_0\Gamma),\ \Gamma\in\mathcal{D}_N,\ \rho_\Gamma=\rho\right\}.$$

Let us denote by

$$\mathcal{W} := \left\{\Phi=(\phi_i)_{i\in\mathbb{N}^*},\ \phi_i\in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3}\phi_i\phi_j=\delta_{ij}\right\},$$

and by

$$\mathcal{N}_N := \left\{\nu=(n_i)_{i\in\mathbb{N}^*},\ 0\leq n_i\leq 1,\ \sum_{i=1}^{+\infty}n_i=N\right\}.$$

Then, it is rigorously proved that $T_J(\rho)$ can be equivalently rewritten as

$$T_J(\rho) = \inf\left\{\frac{1}{2}\sum_{i=1}^{+\infty}n_i\int_{\mathbb{R}^3}|\nabla\phi_i|^2,\ \Phi=(\phi_i)_{i\in\mathbb{N}^*}\in\mathcal{W},\nu=(n_i)_{i\in\mathbb{N}^*}\in\mathcal{N}_N,\ \sum_{i=1}^{+\infty}n_i|\phi_i|^2=\rho\right\}.$$

The functionals T_{KS} and T_J associated with the Hamiltonian H_0 are reasonably good approximations of the kinetic energy of the electrons. However, they cannot be expressed as an explicit function of ρ or its derivatives, unlike in orbital-free models.

The *Coulomb energy*

$$J(\rho) = \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\rho(x)\rho(y)}{|x-y|}\,dx\,dy$$

represents the electrostatic energy of a classical charge distribution ρ and gives a reasonable approximation of the interaction energy between the electrons of a system of electronic density ρ.

In the case of the HF model, we have seen that the energy of a Slater determinant (for the Hamiltonian H_1) was composed of the kinetic energy, the Coulomb energy and the exchange energy, and that the difference between the Hartree-Fock energy and the exact fundamental energy of (1.2) was called the correlation energy. The same terminology is used in the Kohn-Sham type models, in which the errors made on the kinetic energy and electronic repulsion energy are gathered in a single functional $E_{xc}(\rho)$, called the *exchange-correlation functional*. It is defined by

$$E_{xc}(\rho) := F_{LL}(\rho) - T_{KS}(\rho) - J(\rho),$$

or by

$$E_{xc}(\rho) := F_L(\rho) - T_J(\rho) - J(\rho),$$

depending on whether the Levy-Lieb or Lieb DFT approaches are adopted.

The *standard Kohn-Sham* model originates from the Levy-Lieb approach and reads as the following minization problem

$$I^{SKS} = \inf\{E^{SKS}(\Phi),\ \Phi \in \mathcal{W}_N\}, \tag{1.11}$$

where for all $\Phi = (\phi_i)_{1\leq i\leq N} \in \mathcal{W}_N$,

$$\begin{aligned} E^{SKS}(\Phi) &= \sum_{i=1}^{N} \frac{1}{2}\int_{\mathbb{R}^3} |\nabla\phi_i|^2 + \int_{\mathbb{R}^3} \rho_\Phi V \\ &\quad + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho_\Phi(x)\rho_\Phi(y)}{|x-y|}\,dx\,dy + E_{xc}(\rho_\Phi). \end{aligned}$$

The Lieb approach leads to the *extended Kohn-Sham* model, which reads

$$I^{EKS} = \inf\{E^{EKS}(\nu,\Phi),\ \Phi \in \mathcal{W},\ \nu \in \mathcal{N}_N\},$$

where for all $\Phi = (\phi_i)_{i\in\mathbb{N}^*} \in \mathcal{W}$ and all $\nu = (n_i)_{i\in\mathbb{N}^*} \in \mathcal{N}_N$,

$$\begin{aligned} E^{EKS}(\nu,\Phi) &= \sum_{i=1}^{+\infty} \frac{1}{2}\int_{\mathbb{R}^3} n_i|\nabla\phi_i|^2 + \int_{\mathbb{R}^3} \rho_{\nu,\Phi} V \\ &\quad + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho_{\nu,\Phi}(x)\rho_{\nu,\Phi}(y)}{|x-y|}\,dx\,dy + E_{xc}(\rho_{\nu,\Phi}), \end{aligned}$$

where

$$\rho_{\nu,\Phi} := \sum_{i=1}^{+\infty} n_i|\phi_i|^2.$$

The extended Kohn-Sham model can be rewritten in an equivalent manner (like in the case of the HF model) using order 1 density operators associated to mixed states, which are defined by

$$\gamma_{\nu,\Phi} = \sum_{i=1}^{+\infty} n_i\phi_i(x)\phi_i(y),\ \Phi = (\phi_i)_{i\in\mathbb{N}^*} \in \mathcal{W},\ \nu = (n_i)_{i\in\mathbb{N}^*} \in \mathcal{N}_N.$$

It indeed holds that

$$I^{EKS} = \inf\left\{E^{EKS}(\gamma),\ \gamma \in \mathcal{L}_N\right\}, \tag{1.12}$$

where

$$E^{EKS}(\gamma) := \frac{1}{2}\mathrm{Tr}(-\Delta\gamma) + \int_{\mathbb{R}^3} \rho_\gamma V + J(\rho_\gamma) + E_{xc}(\rho_\gamma).$$

A huge number of candidate expressions for the exchange-correlation functional can be found in the literature. Among them, let us mention

- the Local Density Approximation (LDA) where $E_{xc}(\rho)$ is assumed to be of the form
$$E_{xc}(\rho) = \int_{\mathbb{R}^3} g(\rho(x))\,dx,$$
where $g : \mathbb{R}_+ \to \mathbb{R}$;

- the Gradient Corrected Approximation (GCA) where $E_{xc}(\rho)$ is also assumed to depend locally on the gradient of the density, i.e.
$$E_{xc}(\rho) = \int_{\mathbb{R}^3} h(\rho(x), \nabla\rho(x))\,dx,$$
with $h : \mathbb{R}_+ \times \mathbb{R}^3 \to \mathbb{R}$.

In practice, the LDA functional is mainly used with the definition of g which derives from interacting uniform electron gas calculations and was introduced by Kohn and Sham [119]. For the GGA functional, there exist a lot of different choices for the function h in the literature (see e.g. [13, 125, 159, 160]). Let us point out that the extended Kohn-Sham model with no exchange-correlation energy functional ($E_{xc} = 0$) is strictly equivalent to the reduced Hartree-Fock model introduced in Section 1.2.1.

It is known that for neutral or positively charged systems, the standard [127] and extended [6] Kohn-Sham LDA problem does have a minimizer and some partial results exist for the Kohn-Sham GGA problem in the case of $N = 2$ electrons [6].

Typically, for the extended Kohn-Sham LDA problem [6], provided that $Z \geq N$ and that g is a C^1 function from $\mathbb{R}_+$ to $\mathbb{R}$, twice differentiable on $\mathbb{R}_+^*$ and satisfying the following assumptions

- $g(0) = 0$;
- $g' \leq 0$;
- $\exists 0 < \beta_- \leq \beta_+ < \frac{2}{3}$ such that $\sup\limits_{\rho\in\mathbb{R}_+} \dfrac{|g'(\rho)|}{\rho^{\beta^-} + \rho^{\beta^+}} < \infty$;
- $\exists 1 \leq \alpha < \frac{3}{2}$ such that $\operatorname{limsup}_{\rho\to 0^+} \frac{g(\rho)}{\rho^\alpha} < 0$,

then, problem (1.12) admits a minimizer γ. This minimizer satisfies the self-consistent equation
$$\gamma = \mathbb{1}_{(-\infty,\epsilon_F)}(H_\gamma^{EKS}) + \delta,$$
for some $\epsilon_F \leq 0$, where
$$H_\gamma^{EKS} = -\frac{1}{2}\Delta + V + \rho_\gamma * |\cdot|^{-1} + g'(\rho_\gamma),$$
and where δ is a self-adjoint operator on $L^2(\mathbb{R}^3)$ such that $0 \leq \delta \leq 1$ and $\operatorname{Ran}(\delta) \subset \operatorname{Ker}(H_\gamma - \epsilon_F)$. The minimizer is unique if $\epsilon_F \notin \sigma(H_\gamma)$.

1.3 Nonlinear mean-field models for periodic crystals

1.3.1 Thermodynamic limit

In this section, we are going to present nonlinear mean-field models used in quantum chemistry for the computation of the electronic structure of crystalline materials. These are composed of an infinite set of nuclei arranged periodically in space. Hopefully, the ground state electronic density will inherit the same periodicity, so that it will be possible to reduce the problem posed over the entire space to a problem posed only on a unit cell of the lattice with periodic (for Thomas-Fermi type models) or quasiperiodic (for Hartree-Fock or Kohn-Sham type models) boundary conditions. However, the fact that the density has the same periodicity as the nuclear distribution (i.e. that there is no symmetry breaking) is not obvious at all. This question can, in some particular cases, be answered by means of *thermodynamic limit* arguments.

Let us first introduce some notation. Let $\mathcal{R}$ be a periodic lattice of $\mathbb{R}^d$ with $d \in \mathbb{N}^*$ and Γ a unit cell of $\mathcal{R}$ such that 0 is in the interior of Γ. The lattice $\mathcal{R}$ can be defined as follows:

$$\mathcal{R} = \left\{ \sum_{i=1}^{d} u_i \mathbf{a}_i,\ (u_i)_{1\leq i\leq d} \in \mathbb{Z}^d \right\},$$

where $(\mathbf{a}_i)_{1\leq i\leq d}$ forms a basis of $\mathbb{R}^d$ and is said to be a *basis of* $\mathcal{R}$. An admissible choice for the unit cell Γ is

$$\Gamma = \left\{ \sum_{i=1}^{d} x_i \mathbf{a}_i,\ (x_i)_{1\leq i\leq d} \in \left[-\frac{1}{2}, \frac{1}{2}\right)^d \right\}.$$

We define the *reciprocal lattice* of $\mathcal{R}$ as

$$\mathcal{R}^* = \left\{ \sum_{i=1}^{d} u_i \mathbf{a}'_i,\ (u_i)_{1\leq i\leq d} \in \mathbb{Z}^d \right\},$$

where the basis $(\mathbf{a}'_i)_{1\leq i\leq d}$ satisfies the following relationships

$$\mathbf{a}_i \cdot \mathbf{a}'_j = 2\pi\delta_{ij},\ 1 \leq i, j \leq d.$$

This definition of $\mathcal{R}^*$ does not depend on the choice of the basis of $\mathcal{R}$. The *Wigner-Seitz cell* of a lattice, or Dirichlet zone, is defined as the Voronoï cell associated with the point 0. In other words, it is defined as the set of points of $\mathbb{R}^d$ which are closer to the point 0 than to any other point of the lattice. The Wigner-Seitz cell of the reciprocal lattice is called the *first Brillouin zone* of the crystal and is denoted by Γ^*. We define

$$L^2_{\mathrm{per}}(\Gamma) := \left\{ u \in L^2_{\mathrm{loc}}(\mathbb{R}^d),\ u\ \mathcal{R}\text{-periodic} \right\},$$

and

$$H^1_{\mathrm{per}}(\Gamma) := \left\{ u \in L^2_{\mathrm{per}}(\Gamma),\ \nabla u \in \left(L^2_{\mathrm{per}}(\Gamma)\right)^d \right\}.$$

Now, in the case of a perfect crystal, $d = 3$ and $\mathcal{R}$ is the periodic crystalline lattice. We assume that there are M nuclei inside the unit cell Γ, and denote by $R_1, \cdots, R_M \in \Gamma$ and $z_1, \cdots, z_M \in \mathbb{R}_+$ their positions and charges. The nuclear charge density of the crystal is then

$$\mu_{\rm per} = \sum_{R\in\mathcal{R}} \sum_{k=1}^{M} z_k \delta_{R_k+R},$$

where δ_x denotes the Dirac mass at a point $x \in \mathbb{R}^3$. Let $Z := \sum_{k=1}^{M} z_k$ be the total nuclear charge per unit cell.

For all $L \in \mathbb{N}^*$, let $\Gamma_L := L\Gamma$ be an assembly of L^3 unit cells. We may consider the finite system composed of all the nuclei whose positions are located inside Γ_L and of $L^3 Z$ electrons. Let us assume that, for each $L \in \mathbb{N}^*$, this system admits an electronic ground state whose associated density and energy are denoted respectively by ρ_L and I_L.

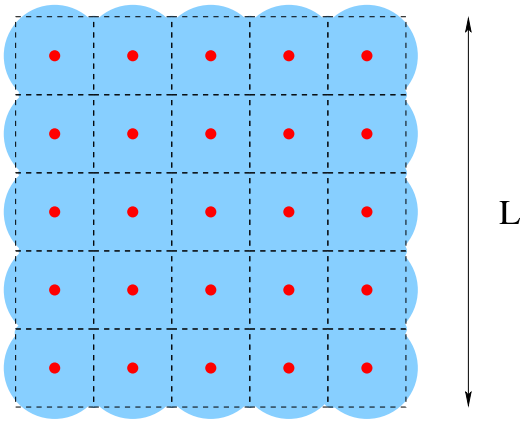

Figure 1.1: Thermodynamic limit for perfect periodic crystals

Then, the thermodynamic limit problem can be stated as follows: does there exist a scalar $I_{\rm per} \in \mathbb{R}$ and an electronic density $\rho_{\rm per}$ (which we expect to be $\mathcal{R}$-periodic) such that

$$\left.\begin{array}{ccc} \rho_L & \underset{L\to\infty}{\longrightarrow} & \rho_{\rm per} \\ \frac{I_L}{L^3} & \underset{L\to\infty}{\longrightarrow} & I_{\rm per} \end{array}\right\} \text{ in a certain sense?}$$

This question is difficult and there are very few models for which an answer has been found.

The periodic Coulomb potential is defined as the unique solution $G \in L^2_{\rm per}(\Gamma)$ of

$$\left\{\begin{array}{l} -\Delta G = 4\pi\left(-\frac{1}{|\Gamma|} + \sum_{R\in\mathcal{R}} \delta_R\right), \\ \int_\Gamma G = 0. \end{array}\right.$$

Let

$$m := \lim_{|x|\to 0} G(x) - \frac{1}{|x|}.$$

We define

$$V_{\rm per}(x) = -\sum_{k=1}^{M} z_k G(x - R_k), \text{ for almost all } x \in \mathbb{R}^3.$$

For all $L \in \mathbb{N}^*$, let us also denote by

$$U_L := \frac{1}{2} \sum_{\substack{R, R' \in \mathcal{R} \cap \Gamma_L, \\ 1 \leq k, l \leq M, \\ R + R_k \neq R' + R_l}} \frac{z_k z_l}{|(R_k + R) - (R_l + R')|},$$

the energy of the Coulomb interaction between the nuclei which are located inside Γ_L.

1.3.2 Thomas-Fermi models

This problem of the thermodynamic limit was first tackled by Lieb for the Thomas-Fermi model [135], then by Catto, Le Bris and Lions for the Thomas-Fermi-von Weiszäcker theory [50]. Let us state more precisely the results in [50], which are summarized in [49].

For $L \in \mathbb{N}^*$, the TFW energy functional of the finite system composed of the nuclei located inside Γ_L reads

$$\begin{aligned} E_L^{TFW}(\rho) &= C_W \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}|^2 + C_{TF} \int_{\mathbb{R}^3} \rho^{5/3} + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|}\, dx\, dy \\ &\quad + \int_{\mathbb{R}^3} \left(\sum_{R \in \mathcal{R} \cap \Gamma_L} V(x - R) \right) \rho(x)\, dx, \end{aligned}$$

where $V(x) = -\sum_{k=1}^{M} \frac{z_k}{|x - R_k|}$ for almost all $x \in \mathbb{R}^3$. Let

$$I_L^{TFW} = \inf \left\{ E_L^{TFW}(\rho) + U_L, \rho \geq 0, \sqrt{\rho} \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3} \rho = L^3 Z \right\},$$

and ρ_L be the unique minimizer of the above problem.

Let us also introduce the periodic TFW problem

$$I_{\rm per}^{TFW} = \inf \left\{ E_{\rm per}^{TFW}(\rho),\ \rho \geq 0,\ \sqrt{\rho} \in H^1_{\rm per}(\Gamma),\ \int_{\Gamma} \rho = Z \right\}, \tag{1.13}$$

where

$$\begin{aligned} E_{\rm per}^{TFW}(\rho) &= C_W \int_{\Gamma} |\nabla \sqrt{\rho}|^2 + C_{TF} \int_{\Gamma} \rho^{5/3} + \int_{\Gamma} \rho V_{\rm per} \\ &\quad + \frac{1}{2} \int_{\Gamma} \int_{\Gamma} \rho(x)\rho(y) G(x - y)\, dx\, dy. \end{aligned}$$

Then, there exists a unique minimizer $\rho_{\rm per}$ of problem (1.13) and the following convergence results hold [50]:

- **Convergence of the energy per unit volume:**

$$\lim_{L\to\infty}\frac{1}{L^3}I_L^{TFW}=I_{\rm per}^{TFW}+\frac{M}{2}.$$

- ***Convergence of the density:*** *As L goes to infinity, $\sqrt{\rho_L}$ converges to $\sqrt{\rho_{\rm per}}$ for the strong topology of $H^1_{\rm loc}(\mathbb{R}^3)$ and the strong topology of $L^p_{\rm loc}(\mathbb{R}^3)$, $1\le p\le\infty$.*

Besides, denoting by $u^0_{\rm per}=\sqrt{\rho_{\rm per}}$, the function $u^0_{\rm per}$ satisfies the Euler-Lagrange equation

$$H^{TFW}_{\rm per}u^0_{\rm per}=\epsilon^0_F u^0_{\rm per},$$

where

$$H^{TFW}_{\rm per}:=-C_W\Delta+\frac{5}{3}C_{TF}|u^0_{\rm per}|^{4/3}+\int_\Gamma G(\cdot-y)|u^0_{\rm per}(y)|^2\,dy+V_{\rm per} \tag{1.14}$$

is the periodic TFW Hamiltonian, and ϵ^0_F is the Lagrange multiplier of the constraint $\int_\Gamma|u^0_{\rm per}|^2=Z$. Besides, it holds that $\epsilon^0_F=\inf\sigma(H^{TFW}_{\rm per})$.

The thermodynamic limit problem for the TFDW model, and more generally for non-convex orbital-free models, remains open.

1.3.3 Bloch theory

Before describing the reduced Hartree-Fock and Kohn-Sham models used for perfect crystals, let us recall here the basics of Bloch theory, which was first introduced by Floquet [89] in the one-dimensional setting, then by Bloch [19] in the general case.

For all $q\in\Gamma^*$, we denote by

$$L^2_q(\Gamma):=\left\{\phi\in L^2_{\rm loc}(\mathbb{R}^d)\mid e^{-iq\cdot x}\phi(x)\in L^2_{\rm per}(\Gamma)\right\}.$$

This space, endowed with the scalar product

$$\langle\phi,\psi\rangle_{L^2_q(\Gamma)}:=\int_\Gamma\overline{\phi}\psi,$$

is a Hilbert space isomorphic to $L^2_{\rm per}(\Gamma)$. The associated norm is denoted by $\|\cdot\|_{L^2_q(\Gamma)}$. The set $\bigcup_{q\in\Gamma^*}L^2_q(\Gamma)$ is called the set of *Bloch functions*.

Let us now consider the space

$$\begin{aligned}&\frac{1}{|\Gamma^*|}\int_{\Gamma^*}^{\oplus}L^2_q(\Gamma)\,dq\\&:=\left\{(\phi_q)_{q\in\Gamma^*}\mid\phi_q\in L^2_q(\Gamma),\ q\mapsto\|\phi_q\|^2_{L^2_q(\Gamma)}\text{ measurable},\ \int_{\Gamma^*}\|\phi_q\|^2_{L^2_q(\Gamma)}\,dq<+\infty\right\}.\end{aligned}$$

Endowed with the scalar product defined by

$$\forall\phi=(\phi_q)_{q\in\Gamma^*},\psi=(\psi_q)_{q\in\Gamma^*}\in\frac{1}{|\Gamma^*|}\int_{\Gamma^*}^{\oplus}L^2_q(\Gamma)\,dq,\ (\phi,\psi):=\frac{1}{|\Gamma^*|}\int_{\Gamma^*}\langle\phi_q,\psi_q\rangle_{L^2_q(\Gamma)}\,dq,$$

this space is a Hilbert space.

Denoting by $\mathcal{S}(\mathbb{R}^d)$ the Schwartz space

$$\mathcal{S}(\mathbb{R}^d) := \left\{ f \in \mathcal{C}^\infty(\mathbb{R}^d),\ \forall \alpha, \beta \in \mathbb{N}^d,\ \sup_{x \in \mathbb{R}^d} |x^\alpha \partial^\beta f(x)| < +\infty \right\},$$

the Bloch decomposition theorem states that the linear map $U : \mathcal{S}(\mathbb{R}^d) \to \frac{1}{|\Gamma^*|} \int_{\Gamma^*}^{\oplus} L^2_q(\Gamma)\, dq$, defined by

$$\forall \phi \in \mathcal{S}(\mathbb{R}^3),\ U\phi = \left((U\phi)_q\right)_{q \in \Gamma^*},\ \text{ with } (U\phi)_q(x) = \sum_{R \in \mathcal{R}} e^{-iq\cdot R} \phi(x+R),$$

can be uniquely extended to a unitary operator U from $L^2(\mathbb{R}^d)$ onto $\frac{1}{|\Gamma^*|} \int_{\Gamma^*}^{\oplus} L^2_q(\Gamma)\, dq$.

Its inverse, denoted by U^*, is defined for all $\phi = (\phi_q)_{q \in \Gamma^*} \in \frac{1}{|\Gamma^*|} \int_{\Gamma^*}^{\oplus} L^2_q(\Gamma)\, dq$ as follows

$$\forall R \in \mathcal{R},\ (U^*\phi)(x-R) = \frac{1}{|\Gamma^*|} \int_{\Gamma^*} e^{-iq\cdot R} \phi_q(x)\, dq \text{ for almost all} x \in \Gamma.$$

The fact that U is an isometry then means that for all $\phi \in L^2(\mathbb{R}^d)$,

$$\phi(x) = \frac{1}{|\Gamma^*|} \int_{\Gamma^*} \phi_q(x)\, dq, \tag{1.15}$$

where for almost all $q \in \Gamma^*$ and $x \in \mathbb{R}^d$,

$$\phi_q(x) = \sum_{R \in \mathcal{R}} e^{-iq\cdot R} \phi(x+R) \in L^2_q(\Gamma),$$

and

$$\|\phi\|^2_{L^2(\mathbb{R}^d)} = \frac{1}{|\Gamma^*|} \int_{\Gamma^*} \|\phi_q\|^2_{L^2_q(\Gamma)}\, dq.$$

The decomposition (1.15) is called the *Bloch transform* of ϕ.

Now, let A be a bounded self-adjoint operator on $L^2(\mathbb{R}^d)$ which commutes with the translations of the periodic lattice, i.e. such that

$$\forall R \in \mathcal{R},\ \tau_R A = A \tau_R,$$

where for all $\phi \in L^2(\mathbb{R}^d)$ and $R \in \mathcal{R}$,

$$(\tau_R \phi)(x) = \phi(x-R).$$

Then, the operator A admits a Bloch transform in the sense that there exists a unique function $q \in \Gamma^* \mapsto A_q \in \mathcal{B}(L^2_q(\Gamma)) \cap \mathcal{S}(L^2_q(\Gamma))$ in $L^\infty(\Gamma^*, \mathcal{B}(L^2_q(\Gamma)))$, such that for all $\phi \in L^2(\mathbb{R}^3)$ and almost all $q \in \Gamma^*$,

$$(UA\phi)_q = A_q (U\phi)_q,$$

i.e.

$$(A\phi)(x) = \frac{1}{|\Gamma^*|} \int_{\Gamma^*} (A\phi)_q(x)\, dq = \frac{1}{|\Gamma^*|} \int_{\Gamma^*} (A_q \phi_q)(x)\, dq.$$

The following symbolic notation is used

$$A = \frac{1}{|\Gamma^*|} \int_{\Gamma^*}^{\oplus} A_q dq.$$

Besides, it holds that

$$\sup_{q \in \Gamma^*} \|A_q\|_{\mathcal{B}(L^2_q(\Gamma))} = \|A\|_{\mathcal{B}(L^2(\mathbb{R}^d))}.$$

If A is a positive self-adjoint operator on $L^2(\mathbb{R}^d)$, then for any $q \in \Gamma^*$, A_q is positive, and we can define

$$\mathrm{Tr}_\Gamma(A) := \frac{1}{\Gamma^*} \int_{|\Gamma^*|} \mathrm{Tr}_{L^2_q(\Gamma)}(A_q)\, dq \in \mathbb{R}_+ \cup \{+\infty\}. \tag{1.16}$$

1.3.4 Hartree-Fock and Kohn-Sham models

The Bloch formalism introduced in the preceding section is central in the definition of Kohn-Sham and Hartree-Fock type models for periodic crystals. In these settings, the electronic structure of a perfect crystal is described by density operators γ satisfying the following properties

- γ is a self-adjoint operator on $L^2(\mathbb{R}^3)$ satisfying $0 \leq \gamma \leq 1$ in the sense of operators (Pauli principle);
- γ commutes with all the translations of the lattice, i.e.

$$\forall R \in \mathcal{R},\ \tau_R \gamma = \gamma \tau_R.$$

According to the results in Section 1.3.3, such an operator γ can be decomposed as

$$\gamma = \frac{1}{|\Gamma^*|} \int_{\Gamma^*}^{\oplus} \gamma_q\, dq,$$

where $\gamma_q \in \mathcal{B}(L^2_q(\Gamma))$ satisfies $0 \leq \gamma_q \leq 1$ for almost all $q \in \Gamma^*$ in the sense of operators on $L^2_q(\Gamma)$. We can also define $\mathrm{Tr}_\Gamma(\gamma)$ as in (1.16) and this quantity will be equal to the number of electrons per unit cell. Let us point out that

$$\mathrm{Tr}_\Gamma(\gamma) = \int_\Gamma \rho_\gamma,$$

where the non-negative $\mathcal{R}$-periodic function

$$\rho_\gamma(x) = \frac{1}{|\Gamma^*|} \int_{\Gamma^*} \tau_{\gamma_q}(x,x)\, dq$$

is the electronic density associated with γ. We also set

$$\mathrm{Tr}_\Gamma(-\Delta\gamma) := \frac{1}{|\Gamma^*|} \int_{\Gamma^*} \mathrm{Tr}_{L^2_q(\Gamma)}(-\Delta_q \gamma_q)\, dq,$$

where Δ_q is the Laplace operator, seen as an operator on $L^2_q(\Gamma)$. Naturally, in the expression above, we have used the usual convention

$$\mathrm{Tr}_\Gamma(-\Delta\gamma) := \mathrm{Tr}_\Gamma((-\Delta)^{1/2}\gamma(-\Delta)^{1/2})$$

and

$$\mathrm{Tr}_{L^2_q(\Gamma)}(-\Delta_q\gamma_q) := \mathrm{Tr}_{L^2_q(\Gamma)}((-\Delta_q)^{1/2}\gamma_q(-\Delta_q)^{1/2}).$$

For all density operators γ such that $\mathrm{Tr}_\Gamma(-\Delta\gamma) < +\infty$, it holds that $\sqrt{\rho_\gamma} \in H^1_{\mathrm{per}}(\Gamma)$. We then define

$$\mathcal{L}_{\mathrm{per}} := \left\{\gamma \in \mathcal{S}(L^2(\mathbb{R}^3)) \mid 0 \le \gamma \le 1,\ \forall R \in \mathcal{R},\ \tau_R\gamma = \gamma\tau_R,\ \mathrm{Tr}_\Gamma((1-\Delta)\gamma) < +\infty\right\}.$$

For a perfect crystal with Z electrons per unit cell, the extended Kohn-Sham model then writes

$$I^{EKS}_{\mathrm{per}} = \inf\left\{E^{EKS}_{\mathrm{per}}(\gamma),\ \gamma \in \mathcal{L}_{\mathrm{per}},\ \mathrm{Tr}_\Gamma(\gamma) = \int_\Gamma \rho_\gamma = Z\right\}, \tag{1.17}$$

where the extended Kohn-Sham energy is given by

$$E^{EKS}_{\mathrm{per}}(\gamma) := \frac{1}{2}\mathrm{Tr}_\Gamma(-\Delta\gamma) + \int_\Gamma V_{\mathrm{per}}\rho_\gamma + \frac{1}{2}\int_\Gamma\int_\Gamma G(x-y)\rho_\gamma(x)\rho_\gamma(y)\,dx\,dy + E_{xc}(\rho_\gamma),$$

where E_{xc} is a well-chosen exchange-correlation functional.

In the case of the reduced Hartree-Fock model ($E_{xc} = 0$), Catto, Le Bris and Lions proved in [53] that problem (1.17) has a minimizer γ_{per} and that all minimizers of (1.17) share the same electronic density $\rho_{\mathrm{per}} = \rho_{\gamma_{\mathrm{per}}}$. Cancès, Deleurence and Lewin proved that this minimizer is actually unique [36], is an orthogonal projector on $L^2(\mathbb{R}^3)$ (i.e. $\gamma^2_{\mathrm{per}} = \gamma^*_{\mathrm{per}} = \gamma_{\mathrm{per}}$) and satisfies the self-consistent equation

$$\gamma_{\mathrm{per}} = \mathbb{1}_{(-\infty,\epsilon^0_F]}(H_{\mathrm{per}}),$$

where H_{per} is the periodic Schrödinger operator on $L^2(\mathbb{R}^3)$ defined by

$$H_{\mathrm{per}} = -\frac{1}{2}\Delta + \mathcal{V}_{\mathrm{per}}, \tag{1.18}$$

the mean-field potential $\mathcal{V}_{\mathrm{per}}$ being given by

$$\mathcal{V}_{\mathrm{per}}(x) = \int_\Gamma G(x-y)\rho_{\mathrm{per}}(y)\,dy + V_{\mathrm{per}}(x),$$

and where ϵ^0_F is called the *Fermi level*. The value of ϵ^0_F can be determined by using the closing relationship

$$\int_\Gamma \rho_{\mathrm{per}}(x)\,dx = Z.$$

Catto, Le Bris and Lions proved the convergence of the thermodynamic limit of the reduced Hartree-Fock model to the periodic problem (1.17). They also exhibited a periodic problem that is thought to correspond to the thermodynamic limit of the full Hartree-Fock model [51, 54, 53, 52, 93] (the exchange term requires some special treatment). We will not give here the form of the periodic Hartree-Fock problem for the sake of brevity. Proving the convergence of the thermodynamic limit in the full Hartree-Fock case remains an open problem.

Let us also mention the results by Fefferman[84], and by Hainzl, Lewin and Solovej [108, 109] who investigated the thermodynamic limit for the many-body Schrödinger equation (1.2). They proved the convergence of the energy per unit cell but the convergence of the density remains an open issue.

1.4 Nonlinear mean-field models for periodic crystals with local defects

Real crystalline materials are not perfectly periodic. Defects, such as vacancies, impurities, dislocations etc. are present inside the material and are accountable for some physical behaviors of the crystals, such as plasticity for instance. We will here only address the issue of modeling one single point defect in a crystalline material. Extended defects or randomly distributed point defects [124] will not be considered in this work.

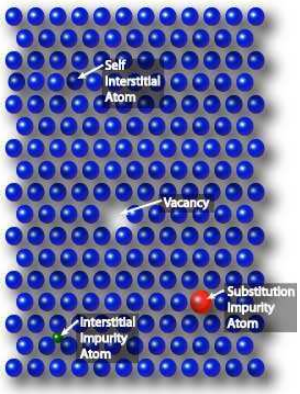

Figure 1.2: Crystal with local defects

For the sake of simplicity, we will consider *smeared nuclei*. In other words, the nuclear charge distribution $\mu_{\rm per}$ of the periodic host crystal will be approximated by

$$\mu_{\rm per}(x) = \sum_{R\in\mathcal{R}} \sum_{k=1}^{M} z_k \chi(x - (R_k + R)),$$

where $\chi \in C_c^\infty(\mathbb{R}^3)$ is a smooth, radial, non-negative function with compact support such that $\int_{\mathbb{R}^3} \chi = 1$. Let us then denote by

$$V_{\rm per}(x) = -\int_\Gamma G(x-y)\mu_{\rm per}(y)\,dy.$$

1.4.1 Reduced Hartree-Fock model

The mathematical analysis of the electronic structure of crystals with defects within the rHF framework has been initiated in [36]. This work is based on a formally simple idea, which is very similar to that used in [55, 105, 106] to properly define a quantum electrodynamical (QED) model for atoms and molecules. The general principle consists in considering the defect (the atom or the molecule in QED) as a quasiparticle embedded in a well-characterized background (a perfect crystal in our case, the polarized vacuum in QED), and to build a variational model allowing one to compute the ground state of the quasiparticle. In [36], such a variational model

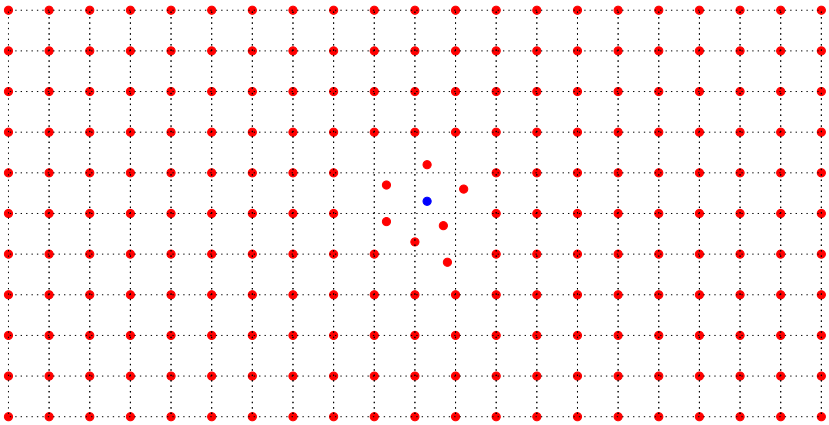

Figure 1.3: Crystal with local defects

is obtained by passing to the thermodynamic limit on the difference between the ground state density matrices obtained respectively with and without the defect.

The analysis done in [36] holds for insulating or semiconductor crystals. Such materials are characterized by the fact that the Fermi level ϵ_F lays in a gap of $\sigma(H_{\rm per})$, the spectrum of the mean-field Hamiltonian $H_{\rm per}$ of the periodic host crystal defined by (1.18). We therefore assume in this section that there exist $\Sigma^-, \Sigma^+ \in \mathbb{R}$ with $\Sigma^- < \Sigma^+$ such that $\epsilon_F \in (\Sigma^-, \Sigma^+)$ and $(\Sigma^-, \Sigma^+) \cap \sigma(H_{\rm per}) = \emptyset$.

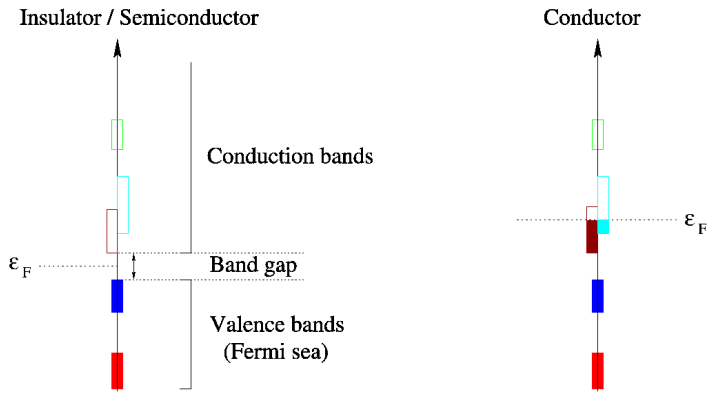

Figure 1.4: Insulating and conducting materials

The local defect we consider introduces a local perturbation $\nu(x)$ of the nuclear charge distribution. The rHF ground state density of the crystal in the presence of the defect can be written as

$$\gamma = \gamma_{\rm per} + Q$$

where $\gamma_{\rm per}$ is the ground state density operator of the host perfect crystal and Q a self-adjoint operator on $L^2(\mathbb{R}^3)$. Although Q is not trace-class in general, both the operators $Q^{++} := \gamma_{\rm per} Q \gamma_{\rm per}$ and $Q^{--} := (1 - \gamma_{\rm per}) Q (1 - \gamma_{\rm per})$ are trace-class.

Actually, Q belongs to the Banach space

$$\begin{aligned}\mathcal{Q} \;:=\; & \{Q \in \mathfrak{S}_2(L^2(\mathbb{R}^3)) \cap \mathcal{S}(L^2(\mathbb{R}^3)),\ Q^{++} \in \mathfrak{S}_1(L^2(\mathbb{R}^3)),\ Q^{--} \in \mathfrak{S}_1(L^2(\mathbb{R}^3)),\\ & |\nabla|Q \in \mathfrak{S}_2(L^2(\mathbb{R}^3)),\ |\nabla|Q^{++}|\nabla| \in \mathfrak{S}_1(L^2(\mathbb{R}^3)),\ |\nabla|Q^{--}|\nabla| \in \mathfrak{S}_1(L^2(\mathbb{R}^3))\}.\end{aligned}$$

For all $Q \in \mathcal{Q}$, we define $\mathrm{Tr}_0(Q)$ as

$$\mathrm{Tr}_0(Q) := \mathrm{Tr}(Q^{++} + Q^{--}).$$

The associated electronic density ρ_Q can be defined in a weak sense by

$$\forall \xi \in \mathcal{C}_c^\infty(\mathbb{R}^3),\ \mathrm{Tr}_0(Q\xi) = \int_{\mathbb{R}^3} \rho_Q \xi.$$

We also define for all $Q \in \mathcal{Q}$,

$$\mathrm{Tr}_0(H_{\mathrm{per}}Q) := \mathrm{Tr}(|H_{\mathrm{per}} - \kappa|^{1/2}(Q^{++} - Q^{--})|H_{\mathrm{per}} - \kappa|^{1/2}) + \kappa \mathrm{Tr}_0(Q),$$

where κ is an arbitrary real number belonging to the gap (Σ^-, Σ^+). Actually, this expression does not depend on the choice of κ.

The energy functional of the defect model can then be written as

$$E_\nu^{rHF}(Q) := \mathrm{Tr}_0(H_{\mathrm{per}}Q) - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho_Q(x)\nu(y)}{|x-y|}\,dx\,dy + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho_Q(x)\rho_Q(y)}{|x-y|}\,dx\,dy.$$

The minimization set where the density operator Q must be chosen is

$$\mathcal{K} := \{Q \in \mathcal{Q},\ -\gamma_{\mathrm{per}} \le Q \le 1 - \gamma_{\mathrm{per}}\}.$$

The condition $-\gamma_{\mathrm{per}} \le Q \le 1 - \gamma_{\mathrm{per}}$ is enforced to ensure that $0 \le \gamma \le 1$.

Cancès, Deleurence and Lewin [36] considered two types of physical settings. The first one gives rise to the minimization problem

$$I_\nu^{rHF}(\epsilon_F) := \inf\left\{E_\nu^{rHF}(Q) - \epsilon_F \mathrm{Tr}_0(Q),\ Q \in \mathcal{K}\right\}, \tag{1.19}$$

for a fixed value of $\epsilon_F \in (\Sigma^-, \Sigma^+)$. In this model, the charge of the defect is controlled by the Fermi level ϵ_F (which thus plays the role of a chemical potential). This problem admits a minimizer Q^{ν,ϵ_F}, which may not be unique. However, all the minimizers of (1.19) share the same density $\rho_{Q^{\nu,\epsilon_F}}$. Besides, Q^{ν,ϵ_F} satisfies the self-consistent equation

$$\begin{aligned} Q^{\nu,\epsilon_F} &= \mathbb{1}_{(-\infty,\epsilon_F)}(H_{Q^{\nu,\epsilon_F}}) - \gamma_{\mathrm{per}} + \delta,\\ H_{Q^{\nu,\epsilon_F}} &= H_{\mathrm{per}} + (\rho_{Q^{\nu,\epsilon_F}} - \nu) * |\cdot|^{-1}, \end{aligned} \tag{1.20}$$

where δ is a self-adjoint finite-rank operator such that $0 \le \delta \le 1$ and $\mathrm{Ran}(\delta) \subset \mathrm{Ker}(H_{Q^{\nu,\epsilon_F}} - \epsilon_F)$. The non-uniqueness of the minimizer of (1.19) relies in the finite-rank operator δ. Actually, if $\epsilon_F \notin \sigma(H_{Q^{\nu,\epsilon_F}})$, Q^{ν,ϵ_F} is unique. Besides, if $\nu \in L^1(\mathbb{R}^3) \cap L^2(\mathbb{R}^3)$, it is proved that $\rho_{Q^{\nu,\epsilon_F}} \in L^2(\mathbb{R}^3)$ and thus $\mathcal{W} := (\rho_{Q^{\nu,\epsilon_F}} - \nu) * |\cdot|^{-1}$ is such

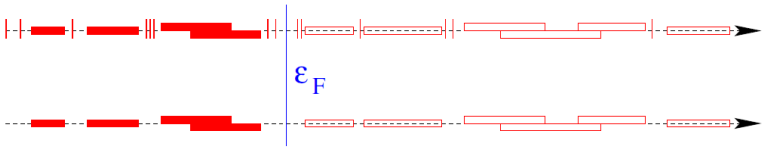

Figure 1.5: Spectra of the Hamiltonian for the perfect crystal (below) and for the crystal with a local defect (above)

that its Fourier transform belongs to $L^1(\mathbb{R}^3)$. Thus, $\mathcal{W} \in L^\infty(\mathbb{R}^3)$, $\mathcal{W}(x) \underset{|x|\to\infty}{\longrightarrow} 0$, and $\mathcal{W}$ is a compact perturbation of the operator $H_{\rm per}$.

The second setting consists in considering the minimization of the energy functional $E_\nu^{rHF}(Q)$ with a charge constraint, namely

$$I_\nu^{rHF}(q) := \inf\left\{E_\nu^{rHF}(Q),\ Q \in \mathcal{K},\ \mathrm{Tr}_0(Q) = q\right\}, \tag{1.21}$$

where the charge $q \in \mathbb{R}$ of the defect is given. Actually, the existence of a minimizer for problem (1.21) is equivalent to each of these conditions

- any minimizing sequence for (1.21) is precompact in $\mathcal{Q}$ and converges towards a minimizer $Q^{\nu,q}$;
- the HVZ type inequalities

$$\forall q' \in \mathbb{R} \setminus \{0\},\ I_\nu^{rHF}(q) < I_\nu^{rHF}(q-q') + I_0^{rHF}(q')$$

  are satisfied.

The set of real numbers $q \in \mathbb{R}$ which satisfy these properties form a nonempty closed interval of $\mathbb{R}$. When these conditions hold, the minimizer is not necessarily unique, but the associated density $\rho_{Q^{\nu,q}}$ is. Besides, there exists a Fermi level $\epsilon_F^{\nu,q} \in [\Sigma^-, \Sigma^+]$ such that $Q^{\nu,q}$ is a minimizer for problem (1.19) with $\epsilon_F = \epsilon_F^{\nu,q}$. This minimizer satisfies the self-consistent equations (1.20) for an operator δ such that $0 \le \delta \le 1$ and $\mathrm{Ran}(\delta) \subset \mathrm{Ker}(H_{Q^{\nu,q}} - \epsilon_F^{\nu,q})$. The operator δ is finite-rank if $\epsilon_F^{\nu,q} \in (\Sigma^-, \Sigma^+)$. Otherwise, it is trace-class.

We refer to [70] for a complete description of these models and of the theoretical results proved in [36]. Note that, still in the rHF setting, the dynamical version of the variational model obtained in [36] is nothing but the random phase approximation (RPA), widely used in solid-state physics. The well-posedness of the nonlinear RPA dynamics, as well as of each term of the Dyson expansion with respect to the external potential, is proved in [44].

1.4.2 Thomas-Fermi type models

The derivation of an exact variational model within the TFW framework was done by Eric Cancès and myself in [38] and the theoretical results and proofs of this article

are presented in Chapter 2 of this manuscript. Let us use the same notation as those used in Section 1.3.2 for the description of the periodic host crystal.

Let us first introduce the Coulomb space $\mathcal{C}$, which is the set of electronic densities whose Coulomb energy is finite:

$$\mathcal{C} := \left\{\rho \in \mathcal{S}'(\mathbb{R}^3) \mid \widehat{\rho} \in L^1_{\mathrm{loc}}(\mathbb{R}^3), |\cdot|^{-1}\widehat{\rho}(\cdot) \in L^2(\mathbb{R}^3)\right\}.$$

Endowed with the inner product,

$$D(\rho_1, \rho_2) := 4\pi \int_{\mathbb{R}^3} \frac{\overline{\widehat{\rho_1}(k)}\widehat{\rho_2}(k)}{|k|^2}\, dk,$$

$\mathcal{C}$ is a Hilbert space. Typically, in the case when $\sqrt{\rho_1},\ \sqrt{\rho_2} \in H^1(\mathbb{R}^3)$, it holds that

$$D(\rho_1, \rho_2) = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho_1(x)\rho_2(y)}{|x-y|}\, dx\, dy,$$

and we recognize the more familiar expression of the Coulomb energy already encountered in previous contexts.

The general approach is essentially the same as the one adopted in the rHF setting. The crystal is assumed to have a nuclear distribution given by $\mu = \mu_{\mathrm{per}} + \nu$ where $\nu(x)$ is the perturbation induced by the local defect.

The TFW electronic state of this system is described by a function v related to the electronic density ρ by the relation

$$v = \sqrt{\rho} - \sqrt{\rho_{\mathrm{per}}} = \sqrt{\rho} - u^0_{\mathrm{per}}.$$

We will see in Chapter 2 that the function v belongs to the minimization set

$$\mathcal{Q}_+ := \left\{v \in H^1(\mathbb{R}^3),\ v \geq -u^0_{\mathrm{per}},\ u^0_{\mathrm{per}} v \in \mathcal{C}\right\}.$$

The ground state electronic density of the crystal with a local defect ν is then given by

$$\rho_\nu = (u^0_{\mathrm{per}} + v_\nu)^2$$

where v_ν is the only minimizer of

$$I^{TFW}_\nu = \inf\left\{E^{TFW}_\nu(v),\ v \in \mathcal{Q}_+\right\},$$

the energy functional $E^{TFW}_\nu(v)$ being defined as follows

$$\begin{aligned} E^{TFW}_\nu(v) := {} & \langle (H^{TFW}_{\mathrm{per}} - \epsilon^0_F)v, v\rangle_{H^{-1}(\mathbb{R}^3), H^1(\mathbb{R}^3)} \\ & + C_{TF}\int_{\mathbb{R}^3}\left(|u^0_{\mathrm{per}} + v|^{10/3} - |u^0_{\mathrm{per}}|^{10/3} - \frac{5}{3}|u^0_{\mathrm{per}}|^{4/3}(2u^0_{\mathrm{per}}v + v^2)\right) \\ & + \frac{1}{2}D(2u^0_{\mathrm{per}}v + v^2 - \nu, 2u^0_{\mathrm{per}}v + v^2 - \nu), \end{aligned}$$

with H^{TFW}_{per} given by (1.14).

Actually, within this theory, and unlike in the rHF case, local defects are always neutrally charged, i.e. the nuclear charge of the defect is fully screened by the crystal in a weak sense. If we denote by $\rho_\nu^0 = \nu + \rho_{\rm per} - \rho_\nu = \nu - (2u_{\rm per}^0 v_\nu + v_\nu^2)$, it holds that

$$\lim_{r\to 0} \frac{1}{|B_r|} \int_{B_r} |\widehat{\rho_\nu^0}(k)|\, dk = 0, \tag{1.22}$$

i.e. 0 is a Lebesgue point of the Fourier transform $\widehat{\tau}_\nu$ of ρ_ν^0. In the case when $\rho_\nu^0 \in L^1(\mathbb{R}^3)$, then $\widehat{\rho_\nu^0} \in \mathcal{C}(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3)$ and (1.22) implies that the charge of the defect $\int_{\mathbb{R}^3} \rho_\nu^0 = \widehat{\rho_\nu^0}(0) = 0$. This means in particular that the TFW model cannot be used to model insulating or semiconductor crystals, in which charged defects can be observed.

A similar full screening effect had already been proved for the Thomas-Fermi model in the case when the host crystal is an homogeneous medium [135].

1.5 Linear models

For any operator A on $L^2(\mathbb{R}^3)$, we denote by $\sigma_d(A)$ the *discrete spectrum* of A, i.e. the set of the isolated eigenvalues of A of finite multiplicity. The *essential spectrum* of A is defined as $\sigma_{\rm ess}(A) = \sigma(A) \setminus \sigma_d(A)$ where $\sigma(A)$ denotes the spectrum of A.

1.5.1 Linear models for finite systems

The mean-field Hamiltonian of a model of Hartree-Fock or Kohn-Sham type can always be rewritten as a self-adjoint operator on $L^2(\mathbb{R}^3)$ of the form

$$H = -\frac{1}{2}\Delta + \mathcal{V}$$

where $\mathcal{V}$ is an operator (not necessarily local) modeling the nuclei-electrons and electrons-electrons interactions.

In linear models, the operator $\mathcal{V}$ is a multiplication operator by an effective potential $\mathcal{V}(x)$ designed by physicists to qualitatively reproduce some of the properties of the system under study. Such operators are called *Schrödinger operators* and the eigenvectors associated with the discrete spectrum are called *bound states*. For finite systems, the effective potential $\mathcal{V}$ is a compact perturbation of the operator $-\Delta$. Since the spectrum of $-\Delta$ as an operator on $L^2(\mathbb{R}^3)$ with domain $H^2(\mathbb{R}^3)$ is purely absolutely continuous and equal to $[0, +\infty)$, it results from the Weyl's theorem [163] that $\sigma_{\rm ess}(H) = \sigma_{\rm ess}(-\Delta) = [0, +\infty)$. If the system is stable, H also has some negative discrete eigenvalues.

In the case of an extended Kohn-Sham LDA or Hartree-Fock type model, the ground state of a finite system is given by

$$\gamma = \mathbb{1}_{(-\infty, \epsilon_F]}(H),$$

where ϵ_F is a negative Fermi level chosen to ensure that $\mathrm{Tr}(\gamma) = N$. Thus, if H has at least N discrete negative eigenvalues, a ground state γ is given by

$$\gamma = \sum_{i=1}^{N} |\phi_i\rangle\langle\phi_i|,$$

where $\Phi = (\phi_i)_{1\leq i\leq N} \in \mathcal{W}_N$ satisfies

$$H\phi_i = \lambda_i \phi_i,\ 1 \leq i \leq N, \tag{1.23}$$

$\lambda_1 \leq \cdots \leq \lambda_N$ being the lowest N discrete eigenvalues of H (counting multiplicities).

There exists a ground state if and only if the operator H possesses at least N negative discrete eigenvalues (counting multiplicities). This ground state is unique if and only if one of the two following conditions is satisfied:

- If the operator H has at least $N+1$ discrete negative eigenvalues, its $(N+1)^{st}$ eigenvalue λ_{N+1} is such that $\lambda_{N+1} > \lambda_N$;
- The operator H has at most N negative discrete eigenvalues.

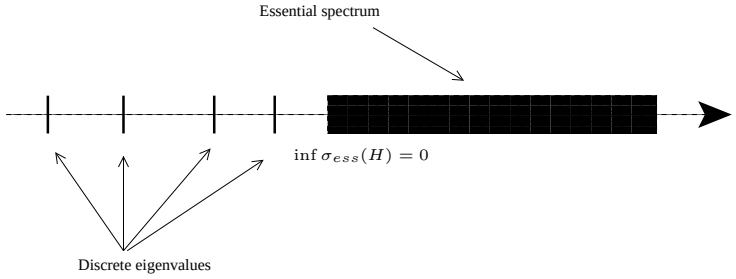

Figure 1.6: Spectrum of H

1.5.2 Linear models for perfect crystals

In the case of Hartree-Fock or Kohn-Sham models for periodic crystalline materials (see Section 1.3), the corresponding mean-field Hamiltonian can be written under the form

$$H_{\rm per} = -\frac{1}{2}\Delta + \mathcal{V}_{\rm per} \tag{1.24}$$

with $\mathcal{V}_{\rm per}$ an $\mathcal{R}$-periodic operator, where $\mathcal{R}$ is the crystalline lattice.

In linear models, the operator $\mathcal{V}_{\rm per}$ is replaced with a multiplication operator by a $\mathcal{R}$-periodic function $\mathcal{V}_{\rm per}(x)$ chosen a priori, like in [60] for instance. It is then important to know precisely the spectral decomposition of the operator $H_{\rm per} = -\frac{1}{2}\Delta + \mathcal{V}_{\rm per}$ in order to have access to the ground state density operator $\gamma_{\rm per} = \mathbb{1}_{(-\infty,\epsilon_F]}(H_{\rm per})$.

Operators of the form (1.24), with $\mathcal{V}_{\rm per}$ a periodic function, are called periodic Schrödinger operators and appear in many physical contexts, apart from the study of crystalline materials in solid state physics. They can also model the properties of photonic crystals when they are considered as operators on $L^2(\mathbb{R}^2)$. Thus, it is interesting to study their properties as operators on $L^2(\mathbb{R}^d)$, with $d \in \mathbb{N}^*$ an arbitrary dimension.

If $\mathcal{V}_{\rm per} \in L^p_{\rm loc}(\mathbb{R}^d)$ with $p = 2$ if $d \leq 3$, $p > 2$ if $d = 4$ and $p > d/2$ if $d \geq 5$, then $H_{\rm per}$ is a bounded from below, self-adjoint operator on $L^2(\mathbb{R}^d)$, with domain $H^2(\mathbb{R}^d)$.

Besides, $H_{\rm per}$ commutes with all the translations of the lattice, and can therefore be decomposed using Bloch theory (which is also valid for unbounded operators, see [163]):

$$H_{\rm per} = \frac{1}{|\Gamma^*|}\int_{\Gamma^*}^{\oplus} H_q\, dq,$$

where for all $q \in \Gamma^*$, H_q is an unbounded self-adjoint operator on $L^2_q(\Gamma)$. Actually, it is proved that for all $q \in \Gamma^*$,

$$H_q = -\frac{1}{2}\Delta + \mathcal{V}_{\rm per},$$

and H_q has a compact resolvent on $L^2_q(\Gamma)$. Thus, for all $q \in \Gamma^*$, there exists a non-decreasing sequence $(\varepsilon_{n,q})_{n\in\mathbb{N}^*}$ and an orthonormal basis $(\psi_{n,q})_{n\in\mathbb{N}^*}$ of $L^2_q(\Gamma)$ such that for all $n \in \mathbb{N}^*$ and all $q \in \Gamma^*$,

$$H_q\psi_{n,q} = \varepsilon_{n,q}\psi_{n,q}.$$

Denoting by $\phi_{n,q}(x) = e^{-iq\cdot x}\psi_{n,q}(x)$, it can be easily checked that $(\phi_{n,q})_{n\in\mathbb{N}^*}$ is an orthonormal basis of $L^2_{\rm per}(\Gamma)$ composed of the eigenvectors of the self-adjoint operator H^q on $L^2_{\rm per}(\Gamma)$ defined by

$$H^q = -\frac{1}{2}\Delta - iq\cdot\nabla + \frac{|q|^2}{2} + \mathcal{V}_{\rm per}.$$

More precisely, for all $q \in \Gamma^*$ and $n \in \mathbb{N}^*$,

$$H^q\phi_{n,q} = \varepsilon_{n,q}\phi_{n,q}.$$

As the application $q \in \Gamma^* \mapsto H^q$ is analytic, the functions $q \in \Gamma^* \mapsto \varepsilon_{n,q}$ are Lipschitz. Besides, it can be proved [163] that the spectrum of $H_{\rm per}$ is purely absolutely continuous and

$$\sigma(H_{\rm per}) = \bigcup_{n=1}^{\infty} b_n, \quad \text{where } b_n = \bigcup_{q\in\Gamma^*}\{\varepsilon_{n,q}\} = [\Sigma_n^-, \Sigma_n^+],\ \Sigma_n^- = \min_{q\in\Gamma^*}\varepsilon_{n,q},\ \Sigma_n^+ = \max_{q\in\Gamma^*}\varepsilon_{n,q}.$$

An interval (a, b) with $a, b \in \mathbb{R}$ such that $a < b$, $a, b \in \sigma(H_{\rm per})$ and $(a, b)\cap\sigma(H_{\rm per}) = \emptyset$ is called a *gap* of the operator $H_{\rm per}$.

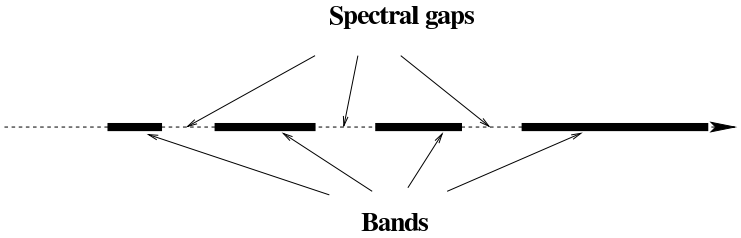

Figure 1.7: Spectrum of $H_{\rm per}$

Besides, since $\gamma_{\rm per} = \mathbb{1}_{(-\infty,\epsilon_F]}(H_{\rm per})$, it holds that

$$\gamma_{\rm per} = \frac{1}{|\Gamma^*|}\int_{\Gamma^*}^{\oplus} \gamma_q\, dq,$$

where for all $q \in \Gamma^*$,

$$\gamma_q = \mathbb{1}_{(-\infty,\epsilon_F]}(H_q) = \sum_{n=1}^{+\infty} \mathbb{1}_{(-\infty,\epsilon_F]}(\varepsilon_{n,q}) |\psi_{n,q}\rangle\langle\psi_{n,q}|.$$

The electronic density associated with the operator γ_{per} is then given by

$$\rho_{\mathrm{per}}(x) = \frac{1}{|\Gamma^*|} \int_{\Gamma^*} \sum_{n=1}^{+\infty} \mathbb{1}_{(-\infty,\epsilon_F]}(\varepsilon_{n,q}) |\psi_{n,q}(x)|^2 \, dq,$$

and the Fermi level ϵ_F is chosen in order to ensure that $\int_\Gamma \rho_{\mathrm{per}} = Z$.

1.5.3 Linear models for crystals with local defects

In the case when $d = 3$, when a local defect is present inside the periodic crystalline material, the perturbation of the periodicity of the arrangement of the nuclei induces a perturbation of the potential $\mathcal{V}_{\mathrm{per}}$. For an arbitray dimension $d \in \mathbb{N}^*$, the associated perturbed Schrödinger operator reads

$$H_{\mathrm{def}} = -\frac{1}{2}\Delta + \mathcal{V}_{\mathrm{per}} + \mathcal{W} = H_{\mathrm{per}} + \mathcal{W},$$

where $\mathcal{W}$ is a potential such that $\mathcal{W} \in L^\infty(\mathbb{R}^d)$ and $\mathcal{W}(x) \underset{|x|\to+\infty}{\longrightarrow} 0$.

The potential $\mathcal{W}$ is a compact perturbation of H_{per}. Thus, the operator H_{def} is self-adjoint and bounded from below on $L^2(\mathbb{R}^d)$, with domain $H^2(\mathbb{R}^d)$. Furthermore, from Weyl's theorem [163], $\sigma_{\mathrm{ess}}(H_{\mathrm{def}}) = \sigma_{\mathrm{ess}}(H_{\mathrm{per}})$. However, there may appear some discrete eigenvalues inside the spectral gaps of the periodic Schrödinger operator H_{per}. Chapters 3 and 4 of this thesis will be concerned with the computation of these discrete eigenvalues.

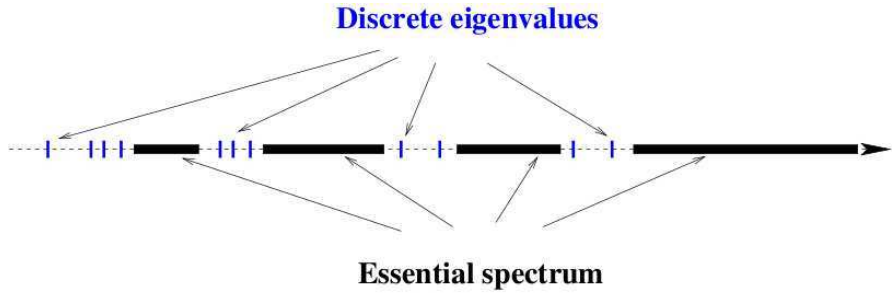

Figure 1.8: Spectrum of H_{def}

1.6 Numerical methods

In this section, we present some of the most common numerical methods used for electronic calculations of crystals with or without local defects. For the sake of brevity, we do not deal here with computations done for finite systems and refer to [42, 126, 186].

1.6.1 Electronic structure calculations of perfect periodic crystals

Density Functional Theory models

Let us describe here the most common numerical methods used to discretize the Thomas-Fermi-von Weiszäcker model. Let us recall that the periodic TFW problem for a perfect periodic crystal reads as follows:

$$I_{\rm per}^{TFW} = \left\{ E_{\rm per}^{TFW}(\rho),\ \rho \geq 0,\ \sqrt{\rho} \in H^1_{\rm per}(\Gamma),\ \int_\Gamma \rho = Z \right\},$$

where

$$\begin{aligned} E_{\rm per}^{TFW}(\rho) &= C_W \int_\Gamma |\nabla \sqrt{\rho}|^2 + C_{TF} \int_\Gamma \rho^{5/3} \\ &\quad + \int_\Gamma \rho V_{\rm per} + \frac{1}{2} \int_\Gamma \int_\Gamma G(x-y)\rho(x)\rho(y). \end{aligned}$$

This minimization problem has a unique minimizer $\rho_{\rm per}$ and $u^0_{\rm per} = \sqrt{\rho_{\rm per}}$ is such that $u^0_{\rm per} \in H^1_{\rm per}(\Gamma)$, $u^0_{\rm per} \geq 0$ on $\mathbb{R}^3$ and

$$H_{\rm per}^{TFW} u^0_{\rm per} = -C_W \Delta u^0_{\rm per} + \left(\frac{5}{3} C_{TF} |u^0_{\rm per}|^{4/3} + V_{\rm per} + \int_\Gamma G(\cdot - y) |u^0_{\rm per}(y)|^2 \, dy \right) u^0_{\rm per} = \epsilon_F^0 u^0_{\rm per},$$

where ϵ_F^0 is the infimum of the spectrum of $H_{\rm per}^{TFW}$.

A first standard way of discretizing this minimization problem is to use finite element methods. We will not detail this method and concentrate more on the planewave discretization method, whose numerical analysis has recently been published in [35].

Let us assume for the sake of simplicity that $\mathcal{R} = \mathbb{Z}^3$. We denote by $\mathcal{R}^* = 2\pi\mathbb{Z}^3$ the dual lattice of $\mathcal{R}$, and by $e_k(x) = \frac{1}{|\Gamma|^{1/2}} e^{ik\cdot x}$ the planewave with wavevector $k \in \mathcal{R}^*$. The family $(e_k)_{k\in\mathcal{R}^*}$ forms an orthonormal basis of $L^2_{\rm per}(\Gamma, \mathbb{C})$ and for all $u \in L^2_{\rm per}(\Gamma, \mathbb{C})$,

$$u(x) = \sum_{k\in\mathcal{R}^*} \widehat{u}_k e_k(x) \quad \text{with} \quad \widehat{u}_k = (e_k, u)_{L^2_{\rm per}} = \frac{1}{|\Gamma|^{1/2}} \int_\Gamma u(x) e^{-ik\cdot x} \, dx.$$

For $N_c \in \mathbb{N}$, we denote by

$$V_{N_c} = \left\{ \sum_{k\in\mathcal{R}^*,\ |k|\leq 2\pi N_c} c_k e_k,\ \forall k, c_{-k} = c_k^* \right\}.$$

For all $v \in L^2_{\rm per}(\Gamma, \mathbb{C})$, the orthogonal projection of v on V_{N_c} is given by

$$\Pi_{N_c} v = \sum_{k \in \mathcal{R}^*,\ |k| \leq 2\pi N_c} \widehat{v}_k e_k.$$

For $N_g \in \mathbb{N}^*$, we denote by $\widehat{\phi}^{FFT,N_g}$ the discrete Fourier transform on the cartesian grid $\mathcal{G}_{N_g} := \frac{1}{N_g}\mathbb{Z}^3$ of the continuous $\mathcal{R}$-periodic function ϕ. In other words, $\widehat{\phi}^{FFT,N_g} = \left(\widehat{\phi}^{FFT,N_g}_k\right)_{k \in \mathcal{R}^*}$ where

$$\widehat{\phi}^{FFT,N_g}_k = \frac{1}{N_g^3} \sum_{x \in \mathcal{G}_{N_g} \cap \Gamma} \phi(x) e^{-ik\cdot x} = |\Gamma|^{-1/2} \sum_{K \in \mathcal{R}^*} \widehat{\phi}_{k+N_g K}.$$

We now introduce the subspaces

$$W^{1D}_{N_g} = \begin{cases} \text{Span}\left\{e^{ily},\ l \in 2\pi\mathbb{Z},\ |l| \leq 2\pi\left(\frac{N_g-1}{2}\right)\right\} & (N_g \text{ odd}), \\ \text{Span}\left\{e^{ily},\ l \in 2\pi\mathbb{Z},\ |l| \leq 2\pi\left(\frac{N_g}{2}\right)\right\} \oplus \mathbb{C}\left(e^{i\pi N_g y} + e^{-i\pi N_g y}\right) & (N_g \text{ even}), \end{cases}$$

and $W^{3D}_{N_g} = W^{1D}_{N_g} \otimes W^{1D}_{N_g} \otimes W^{1D}_{N_g}$. It is possible to define the interpolation projector $\mathcal{I}_{N_g}$ from the set of $\mathcal{R}$-periodic continuous functions onto $W^{3D}_{N_g}$ by $[\mathcal{I}_{N_g}(\phi)](x) = \phi(x)$ for all $x \in \mathcal{G}_{N_g}$. In particular, when N_g is odd, it holds

$$\mathcal{I}_{N_g}(\phi) = |\Gamma|^{1/2} \sum_{k \in \mathcal{R}^*,\ |k|_\infty \leq 2\pi\left(\frac{N_g-1}{2}\right)} \widehat{\phi}^{FFT,N_g}_k e_k.$$

The planewave discretization of the TFW model is obtained by choosing

- an integer $N_c \in \mathbb{N}^*$, and consequently a finite dimensional Fourier space V_{N_c};
- a cartesian grid $\mathcal{G}_{N_g}$ with step size $\frac{1}{N_g}$ where $N_g \geq 4N_c + 1$;

and by considering the finite dimensional minimization problem

$$I^{TFW}_{N_c,N_g} = \inf\left\{E^{TFW}_{N_g}(v_{N_c}),\ v_{N_c} \in V_{N_c}, \int_\Gamma |v_{N_c}|^2 = N\right\}, \tag{1.25}$$

where

$$\begin{aligned} E^{TFW}_{N_g}(v_{N_c}) \ &= \ C_W \int_\Gamma |\nabla v_{N_c}|^2 + C_{TF} \int_\Gamma \mathcal{I}_{N_g}(|v_{N_c}|^{10/3}) + \int_\Gamma \mathcal{I}_{N_g}(V_{\rm per})|v_{N_c}|^2 \\ &\quad + \frac{1}{2}\int_\Gamma \int_\Gamma G(x-y)|v_{N_c}(x)|^2 |v_{N_c}(y)|^2 \, dx\, dy. \end{aligned}$$

The minimization problem (1.25) has at least one minimizer, and any minimizer u_{N_c,N_g} of (1.25) satisfies the Euler-Lagrange equation

$$-C_W \Delta u_{N_c,N_g} + \Pi_{N_c}\left[\left(\mathcal{I}_{N_g}\left(\frac{5}{3}C_{TF}|u_{N_c,N_g}|^{4/3} + V_{\rm per}\right) + V^{Coulomb}_{|u_{N_c,N_g}|^2}\right) u_{N_c,N_g}\right] = \epsilon_F^{N_c,N_g} u_{N_c,N_g},$$

where for almost all $x \in \mathbb{R}^3$, $V^{Coulomb}_{|u_{N_c,N_g}|^2}(x) = \int_\Gamma G(x-y)|u_{N_c,N_g}(y)|^2\,dy$ and where $\epsilon_F^{N_c,N_g}$ is the smallest eigenvalue of the discrete finite dimensional Hamiltonian

$$H^{TFW,N_g}_{|u_{N_c,N_g}|^2} = -C_W\Delta + \mathcal{I}_{N_g}\left(\frac{5}{3}C_{TF}|u_{N_c,N_g}|^{4/3} + V_{\rm per} + V^{Coulomb}_{|u_{N_c,N_g}|^2}\right)$$

defined on V_{N_c} by the Fourier matrix

$$\begin{aligned} M_{kl} &= C_W|k|^2\delta_{kl} + \frac{5}{3}C_{TF}\big(\widehat{|u_{N_c,N_g}|^{4/3}}\big)^{FFT,N_g}_{k-l} + \widehat{(V_{\rm per})}^{FFT,N_g}_{k-l} \\ &\quad + 4\pi\frac{(\widehat{|u_{N_c,N_g}|^2})^{FFT,N_g}_{k-l}}{|k-l|^2}(1-\delta_{kl}). \end{aligned}$$

The following a priori error estimates hold [35]:

Theorem 1.6.1. *Assume that*

$$\exists m > 3,\ C \geq 0,\ \forall k \in \mathcal{R}^*,\ |\widehat{V}_{\rm per,k}| \leq C|k|^{-m}. \tag{1.26}$$

For each $N_c \in \mathbb{N}$, and $N_g \geq 4N_c + 1$, let u_{N_c,N_g} be a minimizer to (1.25) such that $\langle u_{N_c,N_g}, u^0_{\rm per}\rangle_{L^2_{\rm per}(\Gamma)} \geq 0$. Then, for N_c large enough, u_{N_c,N_g} is unique and the following estimates hold true

$$\begin{aligned} \|u_{N_c,N_g} - u^0_{\rm per}\|_{H^s_{\rm per}(\Gamma)} &\leq C_{s,\varepsilon}\left(N_c^{-(m-s+1/2-\varepsilon)} + N_c^{3/2+(s-1)}N_g^{-m}\right), \\ |\epsilon_F^{N_c,N_g} - \epsilon_F^0| &\leq C_\varepsilon\left(N_c^{-(2m-1-\varepsilon)} + N_c^{3/2}N_g^{-m}\right), \\ |I^{TFW}_{\rm per} - I^{TFW}_{N_c,N_g}| &\leq C_\varepsilon\left(N_c^{2m-1-2\varepsilon} + N_c^{3/2}N_g^{-m}\right), \end{aligned}$$

for all $-m+3/2 < s < m+1/2$ and $\varepsilon > 0$, and for some constants $C_{s,\varepsilon} > 0$ and $C_\varepsilon > 0$ independent on N_c and N_g.

Let us point out that similar a priori error estimates were derived in [35] for the Kohn-Sham LDA model without numerical integration.

Linear models

Recall that

$$\gamma_{\rm per} = \mathbb{1}_{(-\infty,\varepsilon_F]}(H_{\rm per}) = \frac{1}{|\Gamma^*|}\int_{\Gamma^*}^{\oplus}\sum_{n=1}^{+\infty}\mathbb{1}_{(-\infty,\varepsilon_F]}(H_q)\,dq.$$

Thus, computing the spectral decomposition of the operator $H_{\rm per}$ on $L^2(\mathbb{R}^3)$ is equivalent to computing, for all $q \in \Gamma^*$, the spectral decomposition of H_q on $L^2_q(\Gamma)$. But, the latter is an easy task, since H_q is a bounded below self-adjoint operator on $L^2_q(\Gamma)$ with purely discrete spectrum.

Here again for the sake of simplicity, we will assume that $\mathcal{R} = \mathbb{Z}^3$ with unit cell $\Gamma = \left[-\frac{1}{2},\frac{1}{2}\right)^3$. Let $q = (q_x,q_y,q_z) \in \Gamma^* = [-\pi,\pi)^3$. Let us mention here two classical

methods for computing the spectral decomposition of the operator H_q. The first one consists in considering a variational approximation of the operator H_q on $L^2_q(\Gamma)$ using finite elements discretization spaces with quasi-periodic boundary conditions. In other words, if $\mathcal{T}$ denotes a mesh of $\mathbb{R}^3$ (a properly defined set of tetrahedron $(T)_{T\in\mathcal{T}}$) invariant with respect to translations of the lattice $\mathcal{R}$, then, for $k \in \mathbb{N}^*$, we consider the discretization space

$$\mathbb{P}_k = \{u \in \mathcal{C}(\Gamma, \mathbb{C}),\ \forall T \in \mathcal{T},\ u|_T \text{ is a polynomial function of degree at most } k\}.$$

Then, computing the spectral decomposition of H_q can be done in practice by computing the eigenvalues and associated eigenvectors of the finite-dimensional generalized eigenvalue problem

$$\left\{ \begin{array}{l} \text{find } (\psi_k, \lambda_k) \in V_k \times \mathbb{R} \text{ such that } \|\psi_k\|_{L^2_q(\Gamma)} = 1 \text{ and} \\ \forall \phi_k \in V_k,\ h(\psi_k, \phi_k) = \lambda_k \langle \psi_k, \phi_k \rangle_{L^2_q(\Gamma)}, \end{array} \right.$$

where the sesquilinear form h is defined as

$$\forall \phi, \psi \in H^1_{\rm loc}(\mathbb{R}^3, \mathbb{C}),\ h(\phi, \psi) = \int_\Gamma \nabla \phi^* \cdot \nabla \psi + \int_\Gamma \mathcal{V}_{\rm per} \phi^* \psi,$$

and where $V_k \subset L^2_q(\Gamma) \cap H^1_{\rm loc}(\mathbb{R}^3, \mathbb{C})$ is the finite dimensional space

$$\begin{aligned} V_k \ = \ & \left\{ u \in \mathbb{P}_k,\ \forall x, y, z \in \mathbb{R},\ u\left(-\frac{1}{2}, y, z\right) = e^{-iq_x} u\left(\frac{1}{2}, y, z\right), \right. \\ & \left. u\left(x, -\frac{1}{2}, z\right) = e^{-iq_y} u\left(x, \frac{1}{2}, z\right),\ u\left(x, y, -\frac{1}{2}\right) = e^{-iq_z} u\left(x, y, \frac{1}{2}\right) \right\}. \end{aligned}$$

Recall that the spectral decomposition of H_q on $L^2_q(\Gamma)$ is closely related to the spectral decomposition of the operator

$$H^q = -\frac{1}{2}\Delta - iq \cdot \nabla + \frac{1}{2}|q|^2 + \mathcal{V}_{\rm per}$$

on $L^2_{\rm per}(\Gamma)$. Thus, computing the spectral decomposition of the operator H_q on $L^2_q(\Gamma)$ is equivalent to computing the spectral decomposition of H^q on $L^2_{\rm per}(\Gamma)$. The latter is usually done by using a planewave (Fourier) discretization of $L^2_{\rm per}(\Gamma)$.

1.6.2 Electronic structure calculations of perfect periodic crystals with defects

Galerkin approximations

A first intuitive way of carrying out electronic structure computations on periodic crystals with local defects is to consider Galerkin approximations of the spectrum of the associated periodic perturbed Schrödinger operator. The major problem of this method is that it may lead to the phenomenon of *spectral pollution*, see [131] for instance.

This phenomenon is well-known in quantum and relativistic physics [21, 22, 24]. A prototypical example was given by Szegö [97]: let $f \in L^\infty_{\text{per}}(0,1)$ be a real-valued, piecewise continuous function. Consider the bounded self-adjoint operator T on $L^2_{\text{per}}(0,1)$ defined as $(Tu)(x) = f(x)u(x)$. The operator T then has a band spectrum $\sigma(T) = \text{ess-range}(f)$. Let us now denote by T_N the matrix of T in the Fourier basis $(e^{2i\pi nx})_{-N\leq n\leq N}$. Then,

$$\lim_{N_0\to\infty} \overline{\bigcup_{N\geq N_0} \sigma(T_N)} = \text{Conv}(\sigma(T)),$$

where $\text{Conv}(B)$ denotes the convex hull of a subset of $B \subset \mathbb{R}$. This means that, even if a natural discretization space (here a Fourier space) is chosen to approximate an operator, the discretization may produce *spurious eigenvalues*, i.e. real numbers, that do not belong to the spectrum of the operator, but are the limit of a sequence of eigenvalues of the discretized operators.

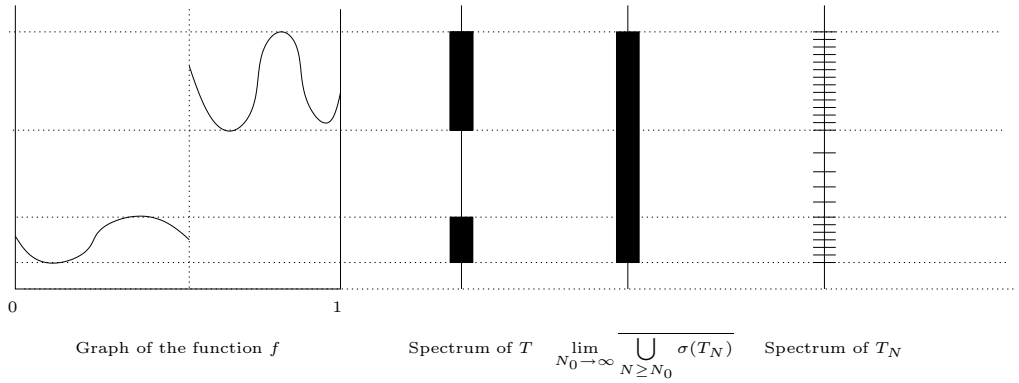

Figure 1.9: Example of spectral pollution

Chapter 3 is concerned with the results proved in [40]. We show that a standard finite element discretization for linear perturbed periodic Schrödinger operator indeed leads to spectral pollution, and that the spurious states can be characterized as surface states in a certain sense.

A way to avoid spectral pollution while using Galerkin approximations is to consider a method inspired by a recent work by Lewin and Séré [134]: provided that the discretization spaces $(V_n)_{n\in\mathbb{N}}$ are chosen so that $V_n = V_n^+ \oplus V_n^-$ with $V_n^+ \subset \text{Ran}(P)$ and $V_n^- \subset \text{Ker}(P)$ with P a well-chosen orthogonal projector, it is possible to avoid spectral pollution in some regions of the spectrum.

This method can be applied in practice to perturbed periodic Schrödinger operators by considering the so-called *Wannier functions*, which form an orthonormal basis of the range of the spectral projector $\mathbb{1}_{(-\infty,\epsilon_F)}(H_{\text{per}})$. It will also be studied in Chapter 3 of this thesis.

Supercell method

The so-called *supercell model* is the current state-of-the-art method to compute the electronic structure of a crystal with a local defect. In this approach, the defect and as many atoms of the host crystal as the available computer resources can accomodate, are put in a large box, called the supercell, and Born-von-Karman periodic

boundary conditions are imposed to the single particle orbitals (and consequently to the electronic density).

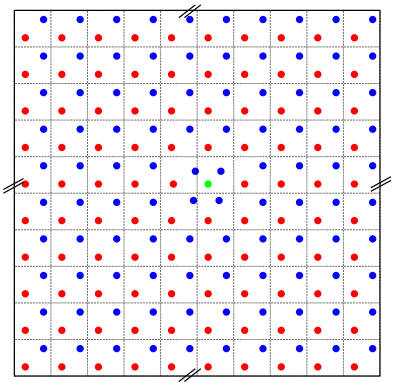

Figure 1.10: The supercell method: periodic boundary conditions

The proof of the convergence of the supercell method was carried out by Cancès, Deleurence and Lewin in [36] for the rHF model and by Cancès and myself in [38] for the TFW model. The proof of the convergence in the framework of the Thomas-Fermi-von Weiszäcker theory will be presented in Chapter 2.

In a joint work with Eric Cancès and Yvon Maday [41], we proved that the supercell method for linear perturbed Schrödinger operators with a planewave discretization, is spectral pollution free for the computation of the electronic properties of periodic crystals in the presence of a local defect. A priori estimates are derived for the approximation of discrete eigenvalues and associated eigenvectors, taking planewave discretization and numerical integration into account. This is the subject of Chapter 4 of this thesis work.

So far, the convergence rate of the supercell method has not been studied for nonlinear mean-field models.

We have seen that in electronic structure calculations, the state-of-the-art techniques for computing electronic properties of local defects in periodic crystals are to use periodic boundary conditions. Let us mention that other boundary conditions based on Dirichlet-to-Neumann operators are used in other fields, such as wave propagation in periodic media containing a local perturbation [87, 88].

Chapter 2

Local defects are always neutral in the Thomas-Fermi-von Weiszäcker theory of crystals

The results of this chapter were the object of an article which appeared in *Archive for Rational Mechanics and Analysis*. An exact variational model for the description of local defects in crystalline materials in the Thomas-Fermi-von Weiszäcker theory of crystals is derived and justified by means of thermodynamic limit arguments. We prove that, within this theory, local defects are necessary electrically neutral and additional results are proved in the case when the host crystal is homogeneous.

Local defects are always neutral in the Thomas-Fermi-von Weiszäcker theory of crystals

Eric Cancès and Virginie Ehrlacher
Université Paris-Est, CERMICS, Project-team Micmac, INRIA-Ecole des Ponts,
6 & 8 avenue Blaise Pascal, 77455 Marne-la-Vallée Cedex 2, France

Abstract

The aim of this article is to propose a mathematical model describing the electronic structure of crystals with local defects in the framework of the Thomas-Fermi-von Weizsäcker (TFW) theory. The approach follows the same lines as that used in *E. Cancès, A. Deleurence and M. Lewin, Commun. Math. Phys., 281 (2008), pp. 129–177* for the reduced Hartree-Fock model, and is based on thermodynamic limit arguments. We prove in particular that it is not possible to model charged defects within the TFW theory of crystals.

2.1 Introduction

The modelling and simulation of the electronic structure of crystals is a prominent topic in solid-state physics, materials science and nano-electronics [116, 161, 173]. Besides its importance for the applications, it is an interesting context for mathematicians for it gives rise to many interesting mathematical and numerical questions.

The mathematical difficulties originate from the fact that crystals consist of infinitely many charged particles (positivitely charged nuclei and negatively charged electrons) interacting with Coulomb potential. Of course, a real crystal contains a finite number of electrons and nuclei, but in order to understand and compute the macroscopic properties of a crystal from first principles, it is in fact easier, or at least not more complicated, to consider that we are dealing with an infinite system.

The first mathematical studies of the electronic structure of crystals were concerned with the so-called (zero-temperature) thermodynamic limit problem for perfect crystals. As opposed to real crystals, which contain local defects (vacancies, interstitial atoms, impurities) and/or extended defects (dislocations, grain boundaries), perfect crystals are periodic arrangements of nuclei and electrons, in the sense that both the nuclear density and the electronic density are $\mathcal{R}$-periodic distributions, $\mathcal{R}$ denoting some discrete periodic lattice of $\mathbb{R}^3$. The thermodynamic limit problem for perfect crystals can be stated as follows. Starting from a given electronic structure model for finite molecular systems, find out an electronic structure model for perfect crystals, such that when a cluster grows and "converges" (in some sense, see [50]) to some $\mathcal{R}$-periodic perfect crystal, the ground state electronic density of the cluster converges to the $\mathcal{R}$-periodic ground state electronic density of the perfect crystal.

For Thomas-Fermi like (orbital-free) models, it is not difficult to guess what should be the corresponding models for perfect crystals. On the other hand, solving the thermodynamic limit problem, that is proving the convergence property discussed above, is a much more difficult task. This program was carried out for the Thomas-Fermi (TF) model in [140] and for the Thomas-Fermi-von Weizsäcker (TFW) model in [50]. Note that these two models are strictly convex in the density, and that the uniqueness of the ground state density is an essential ingredient of the proof. The thermodynamic limit problem for perfect crystals remains open for the Thomas-Fermi-Dirac-von Weisäcker model, and more generally for nonconvex orbital-free models.

The case of Hartree-Fock and Kohn-Sham like models is more difficult. In these models, the electronic state is described in terms of electronic density matrices. For a finite system, the ground state density matrix is a non-negative trace-class self-adjoint operator, with trace N, the number of electrons in the system. For infinite systems, the ground state density matrix is no longer trace-class, which significantly complicates the mathematical arguments. Yet, perfect crystals being periodic, it is possible to make use of Bloch-Floquet theory and guess the structure of the periodic Hartree-Fock and Kohn-Sham models. These models are widely used in solid-state physics and materials science. Here also, the thermodynamic limit problem seems out of reach with state-of-the-art mathematical tools, except in the special case of the restricted Hartree-Fock (rHF) model, also called the Hartree model in the physics literature. Thoroughly using the strict convexity of the rHF energy functional with respect to the electronic density, Catto, Le Bris and Lions were able to solve the thermodynamic limit problem for the rHF model [53].

Very little is known about the modelling of perfect crystals within the framework of the N-body Schrödinger model. To the best of our knowledge, the only available results [84, 108] state that the energy per unit volume is well defined in the thermodynamic limit. So far, the Schrödinger model for periodic crystals is still an unknown mathematical object.

The mathematical analysis of the electronic structure of crystals with defects has been initiated in [36] for the rHF model. This work is based on a formally simple idea whose rigorous implementation however requires some effort. This idea is very similar to that used in [106, 107] to properly define a no-photon quantum electrodynamical (QED) model for atoms and molecules. Loosely speaking, it consists in considering the defect (the atom or the molecule in QED) as a quasiparticle embedded in a well-characterized background (a perfect crystal in our case, the polarized vacuum in QED), and to build a variational model allowing to compute the ground state of the quasiparticle.

In [36], such a variational model is obtained by passing to the thermodynamic limit in the difference between the ground state density matrices obtained respectively with and without the defect. In order to avoid additional technical difficulties, the thermodynamic limit argument in [36] is not carried out on clusters of atoms of increasing sizes, with vanishing boundary conditions at infinity (as in [50, 53]), but on the supercell model. Recall that the supercell model is the current state-of-the-art method to compute the electronic structure of a crystal with a local defect. In this approach, the defect and as many atoms of the host crystal as the available computer resources can accomodate, are put in a large, usually cubic, box, called the supercell, and artificial periodic boundary conditions are imposed to the single particle orbitals (and consequently to the electronic density). The limitations of the supercell methods are well-known: first, it gives rise to spurious interactions between the defect and its periodic images, and second, it requires that the total charge contained in the supercell is neutral (otherwise, the energy per unit volume would be infinite). In the case of charged defects, the extra amount of charge must be compensated in one way or another, for instance by adding to the total physical charge distribution of the system a uniformly charged background (called a jellium). It

is well-known that this procedure generates unphysical screening effects. Other charge compensation methods have been proposed, but none of them is completely satisfactory. Note that the above mentioned sources of error vanish in the thermodynamic limit, when the size of the supercell goes to infinity: both the interaction between a defect and its periodic images and the density of the jellium go to zero in the thermodynamic limit.

The variational model for the defect, considered as a quasiparticle, obtained in [36] has a quite unusual mathematical structure. The rHF ground state density matrix of an insulating or semiconducting crystal in the presence of a local defect can be written as

$$\gamma = \gamma^0_{\rm per} + Q$$

where $\gamma^0_{\rm per}$ is the density matrix of the host perfect crystal (an orthogonal projector on $L^2(\mathbb{R}^3)$ with infinite rank which commutes with the translations of the lattice) and Q a self-adjoint Hilbert-Schmidt operator on $L^2(\mathbb{R}^3)$. Although Q is not trace-class in general [43], it is possible to give a sense to its generalized trace

$$\mathrm{Tr}_0(Q) := \mathrm{Tr}\,(Q^{++}) + \mathrm{Tr}\,(Q^{--}) \quad \text{where} \quad Q^{++} := (1-\gamma^0_{\rm per})Q(1-\gamma^0_{\rm per}) \text{ and } Q^{--} := \gamma^0_{\rm per} Q \gamma^0_{\rm per}$$

(as $\gamma^0_{\rm per}$ is an orthogonal projector, $\mathrm{Tr} = \mathrm{Tr}_0$ on the space of the trace-class operators on $L^2(\mathbb{R}^3)$), as well as to its density ρ_Q. The latter is defined in a weak sense

$$\forall W \in C^\infty_{\rm c}(\mathbb{R}^3), \quad \mathrm{Tr}_0(QW) = \int_{\mathbb{R}^3} \rho_Q W.$$

The function ρ_Q is not in $L^1(\mathbb{R}^3)$ in general, but only in $L^2(\mathbb{R}^3) \cap \mathcal{C}$, where $\mathcal{C}$ is the Coulomb space, that is the space of charge distributions with finite Coulomb energy. An important consequence of these results is that

- in general, the electronic charge of the defect can be defined neither as $\mathrm{Tr}\,(Q)$ nor as $\int_{\mathbb{R}^3} \rho$;
- it may happen that $\rho_Q \in L^1(\mathbb{R}^3)$ but $\mathrm{Tr}_0(Q) \neq \int_{\mathbb{R}^3} \rho_Q$ (while we would have $\rho_Q \in L^1(\mathbb{R}^3)$ and $\mathrm{Tr}_0(Q) = \mathrm{Tr}\,(Q) = \int_{\mathbb{R}^3} \rho_Q$ if Q were a trace-class operator). In this case, $\mathrm{Tr}_0(Q)$ and $\int_{\mathbb{R}^3} \rho_Q$ can be interpreted respectively as the *bare* and *renormalized* electronic charges of the defect [43].

For a given nuclear charge of the defect, the bare (resp. the renormalized) electronic charge of the defect can *a priori* take several values [37], depending on the choice of the Fermi level (i.e. of the chemical potential of the electrons). Yet, if the Coulomb energy of the nuclear charge ν of the defect is small enough and if m is integrable, the bare and renormalized electronic charges of the defect are independent of the choice of the Fermi level, and are respectively equal to 0 and $\frac{L_0}{1+L_0}\int_{\mathbb{R}^3} \nu$, where $0 < L_0 < \infty$ is a constant depending only on the host crystal [43]. Consequenly the renormalized total charge is given by

$$\int_{\mathbb{R}^3} \nu - \frac{L_0}{1+L_0} \int_{\mathbb{R}^3} \nu = \frac{\int_{\mathbb{R}^3} \nu}{(1+L_0)}.$$

This means that the charge ν is partially screened by a factor $1 < (1+L_0) < \infty$. Full screening would correspond to $L_0 = +\infty$; in this case, the renormalized total charge would be equal to zero (neutral defect).

Note that the reason why, in general, Q is not trace-class and ρ_Q is not in $L^1(\mathbb{R}^3)$, is a consequence of both the infinite number of particles and the long-range of the Coulomb interaction. No such singular behavior arises if the long-range Coulomb potential with

kernel $\frac{1}{|x-x'|}$ is replaced by the short-range Yukawa potential with kernel $\frac{e^{-\kappa|x-x'|}}{|x-x'|}$ ($\kappa > 0$ being a fixed paramater).

Note also that, still in the rHF setting, the dynamical version of this variational model is nothing but the random phase approximation (RPA), widely used in solid-state physics. The well-posedness of the nonlinear RPA dynamics, as well as of each term of the Dyson expansion with respect to the external potential, is proved in [44].

Let us emphasize that the results in [36, 43] are limited to insulators and semiconductors, characterized in the rHF setting by the fact that there is a positive gap between the Z^{th} and $(Z+1)^{\text{st}}$ bands of the spectrum of the mean-field Hamiltonian of the perfect crystal, where Z is the number of electrons per unit cell. The mathematical arguments in [36, 43] cannot be straightforwardly adapted to the "metallic" case (absence of gap between the Z^{th} and $(Z+1)^{\text{st}}$ bands). In [140], Lieb and Simon have proved full screening for the TF model, under the assumption that the host crystal is a homogeneous medium. As far as we know, the mathematical study of the electronic structure of crystals with local defects has not been completed for the TFW model [135]. We prove in the present work that defects are always neutral in the TFW theory of crystals (i.e. that the nuclear charge ν of the defect is fully screened by the crystal). This means in particular that the TFW model cannot be used to model insulating or semiconducting crystals, for which the screening effect is only partial, and in which charged defects can be observed.

The article is organized as follows. In Section 2.2, we present the periodic TFW model used in condensed phase computations. After recalling the mathematical structure of the TFW model for perfect crystals (Section 2.3.1), we propose a variational TFW model for crystals with local defects (Section 2.3.2). We prove that this model is well-posed and that the nuclear charge of the defect is fully screened, so that the defect is globally neutral. In Section 2.3.3, we provide a mathematical justification of the model introduced in Section 2.3 based on bulk limit arguments. In Section 2.3.4, we focus on the special case when the host crystal is a homogeneous medium, that is when both the nuclear and electronic densities of the host crystal are uniform (and opposite one another). The technical parts of the proofs are collected in Section 2.4.

2.2 The periodic Thomas-Fermi-von Weiszäcker model

In this section, we describe the Thomas-Fermi-von Weiszäcker (TFW) model with periodic boundary conditions, used to perform computations in the condensed phase. In Section 2.3.3, we will use this periodic model to pass to the thermodynamic limit and construct a rigorously founded TFW model for crystals with local defects.

Let $\mathcal{R}$ be a periodic lattice of $\mathbb{R}^3$, that is a subgroup of $\mathbb{R}^3$ of the form

$$\mathcal{R} = \mathbb{Z}a_1 + \mathbb{Z}a_2 + \mathbb{Z}a_3,$$

where (a_1, a_2, a_3) is a triplet of linearly independent vectors of $\mathbb{R}^3$. The reciprocal (or dual) lattice $\mathcal{R}^*$ associated with $\mathcal{R}$ is defined as

$$\mathcal{R}^* = \mathbb{Z}a_1^* + \mathbb{Z}a_2^* + \mathbb{Z}a_3^*$$

where (a_1^*, a_2^*, a_3^*) is the triplet of linearly independent vectors of $\mathbb{R}^3$ characterized by $a_i \cdot a_j^* = 2\pi\delta_{ij}$. A unit cell is a semi-open convex polyhedron $\Gamma \subset \mathbb{R}^3$ such that $(\Gamma + R)_{R\in\mathcal{R}}$ forms a partition of the space $\mathbb{R}^3$. The Wigner-Seitz cell of $\mathcal{R}$ is the unit cell defined as the Voronoi cell of the origin, that is the set of points of $\mathbb{R}^3$ which are closer to 0 than to any other

point of $\mathcal{R}$. Lastly, we denote by Γ^* the first Brillouin zone, that is the Wigner-Seitz cell of the reciprocal lattice $\mathcal{R}^*$. If for instance $\mathcal{R} = a\mathbb{Z}^3$ (cubic lattice of edge length $a > 0$), then $\mathcal{R}^* = \frac{2\pi}{a}\mathbb{Z}^3$. The Wigner-Seitz cell of $\mathcal{R}$ is $\Gamma = (-\frac{a}{2}, \frac{a}{2}]^3$ and its first Brillouin zone is $\Gamma^* = (-\frac{\pi}{a}, \frac{\pi}{a}]^3$.

We introduce the usual $\mathcal{R}$-periodic L^p spaces defined by

$$L^p_{\rm per}(\Gamma) := \left\{ v \in L^p_{\rm loc}(\mathbb{R}^3) \mid v \ \mathcal{R}\text{-periodic} \right\},$$

and endow them with the norms

$$\|v\|_{L^p_{\rm per}(\Gamma)} := \left(\int_\Gamma |v|^p \right)^{1/p} \quad \text{for } 1 \le p < \infty \quad \text{and} \quad \|v\|_{L^\infty_{\rm per}(\Gamma)} := \text{ess-sup}|v|.$$

In particular,

$$\|v\|_{L^2_{\rm per}(\Gamma)} = (v,v)^{1/2}_{L^2_{\rm per}(\Gamma)} \quad \text{where} \quad (v,w)_{L^2_{\rm per}(\Gamma)} := \int_\Gamma vw.$$

Any function $v \in L^2_{\rm per}(\Gamma)$ can be expanded in Fourier modes as

$$v(x) = \sum_{k \in \mathcal{R}^*} c_k(v) \frac{e^{ik\cdot x}}{|\Gamma|^{1/2}} \quad \text{where} \quad c_k(v) = \frac{1}{|\Gamma|^{1/2}} \int_\Gamma v(x) e^{-ik\cdot x}\, dx.$$

The convergence of the above series holds in $L^2_{\rm per}(\Gamma, \mathbb{C})$, the space of locally square integrable $\mathcal{R}$-periodic $\mathbb{C}$-valued functions.

For each $s \in \mathbb{R}$, the $\mathcal{R}$-periodic Sobolev space of index s is defined as

$$H^s_{\rm per}(\Gamma) := \left\{ v(x) = \sum_{k \in \mathcal{R}^*} c_k(v) \frac{e^{ik\cdot x}}{|\Gamma|^{1/2}} \;\middle|\; \sum_{k \in \mathcal{R}^*} (1+|k|^2)^s |c_k(v)|^2 < \infty,\ \forall k \in \mathcal{R}^*,\ c_{-k} = \overline{c_k} \right\}$$

(throughout this article, $\overline{z}$ denotes the complex conjugate of the complex number z), and endowed with the inner product

$$(v,w)_{H^s_{\rm per}(\Gamma)} := \sum_{k \in \mathcal{R}^*} (1+|k|^2)^s \overline{c_k(v)} c_k(w).$$

Recall that $H^0_{\rm per}(\Gamma) = L^2_{\rm per}(\Gamma)$, $(\cdot,\cdot)_{H^0_{\rm per}(\Gamma)} = (\cdot,\cdot)_{L^2_{\rm per}(\Gamma)}$,

$$H^1_{\rm per}(\Gamma) = \left\{ v \in L^2_{\rm per}(\Gamma) \;\middle|\; \nabla v \in \left(L^2_{\rm per}(\Gamma)\right)^3 \right\}, \quad (v,w)_{H^1_{\rm per}(\Gamma)} = \int_\Gamma vw + \int_\Gamma \nabla v \cdot \nabla w,$$

and $(H^{-\sigma}_{\rm per}(\Gamma))' = H^\sigma_{\rm per}(\Gamma)$. The condition $\forall k \in \mathcal{R}^*,\ c_{-k} = \overline{c_k}$ implies that the functions of $H^s_{\rm per}(\Gamma)$ are real-valued; this is the reason why there is no complex conjugates in the physical space definitions of the inner products $(\cdot,\cdot)_{L^2_{\rm per}(\Gamma)}$ and $(\cdot,\cdot)_{H^1_{\rm per}(\Gamma)}$.

We also introduce the $\mathcal{R}$-periodic Coulomb kernel $G_\mathcal{R}$ defined as the unique function of $L^2_{\rm per}(\Gamma)$ solution of the elliptic problem

$$\begin{cases} -\Delta G_\mathcal{R} = 4\pi \left(\displaystyle\sum_{R \in \mathcal{R}} \delta_R - |\Gamma|^{-1} \right), \\ G_\mathcal{R}\ \mathcal{R}\text{-periodic}, \ \min\limits_{\mathbb{R}^3} G_\mathcal{R} = 0. \end{cases}$$

It is easy to check that

$$G_{\mathcal{R}}(x) = \frac{1}{|\Gamma|}\int_\Gamma G_{\mathcal{R}} + \sum_{k\in\mathcal{R}^*\setminus\{0\}} \frac{4\pi}{|k|^2}\frac{e^{ik\cdot x}}{|\Gamma|}.$$

The $\mathcal{R}$-periodic Coulomb energy is then defined for all f and g in $H^{-1}_{\rm per}(\Gamma)$ by

$$D_{\mathcal{R}}(f,g) \;=\; \left(\int_\Gamma G_{\mathcal{R}}\right)\overline{c_0(f)}c_0(g) + \sum_{k\in\mathcal{R}^*\setminus\{0\}} \frac{4\pi}{|k|^2}\overline{c_k(f)}c_k(g).$$

For all $(f,g)\in L^{6/5}_{\rm per}(\Gamma)\subset H^{-1}_{\rm per}(\Gamma)$, it holds

$$\begin{aligned} D_{\mathcal{R}}(f,g) &= \int_\Gamma\int_\Gamma G_{\mathcal{R}}(x-y)f(x)g(y)\,dx\,dy \\ &= \int_\Gamma (G_{\mathcal{R}}\star_{\mathcal{R}} f)(y)g(y)\,dy = \int_\Gamma (G_{\mathcal{R}}\star_{\mathcal{R}} g)(x)f(x)\,dx, \end{aligned}$$

where $\star_{\mathcal{R}}$ denotes the $\mathcal{R}$-periodic convolution product:

$$\forall (h,k)\in L^1_{\rm per}(\Gamma)\times L^1_{\rm per}(\Gamma),\quad (h\star_{\mathcal{R}} k)(x) = \int_\Gamma h(x-y)k(y)\,dy = \int_\Gamma h(y)k(x-y)\,dy.$$

Let $\rho^{\rm nuc}$ be a function of $H^{-1}_{\rm per}(\Gamma)$ modelling a $\mathcal{R}$-periodic nuclear charge distribution. The corresponding $\mathcal{R}$-periodic TFW energy functional is defined on $H^1_{\rm per}(\Gamma)$ and reads

$$E^{\rm TFW}_{\mathcal{R}}(\rho^{\rm nuc},v) = C_{\rm W}\int_\Gamma |\nabla v|^2 + C_{\rm TF}\int_\Gamma |v|^{10/3} + \frac{1}{2}D_{\mathcal{R}}(\rho^{\rm nuc}-v^2,\rho^{\rm nuc}-v^2), \qquad (2.1)$$

where

$$C_{\rm TF} = \frac{10}{3}(3\pi^2)^{2/3}\quad \text{(Thomas-Fermi constant)}\quad \text{and}\quad C_{\rm W}>0$$

(several values for $C_{\rm W}$ have been proposed in the literature, see e.g. [80]). From a physical viewpoint, $\rho = v^2$ represents the electronic density. The first two terms of $E^{\rm TFW}_{\mathcal{R}}(\rho^{\rm nuc},v)$ model the kinetic energy per simulation cell and the third term the Coulomb energy of the total $\mathcal{R}$-periodic charge distribution $\rho^{\rm tot} = \rho^{\rm nuc} - v^2$.

The electronic ground state with Q electrons in the simulation cell is obtained by solving the minimization problem

$$I_{\mathcal{R}}(\rho^{\rm nuc},Q) = \inf\left\{E^{\rm TFW}_{\mathcal{R}}(\rho^{\rm nuc},v),\ v\in H^1_{\rm per}(\Gamma),\ \int_\Gamma v^2 = Q\right\}. \qquad (2.2)$$

For the sake of simplicity, we assume here that the nuclear charge density is in $H^{-1}_{\rm per}(\Gamma)$. This allows us to collect all the Coulomb interactions in a single, non-negative term (the third term in the right hand side of (2.1)). On the other hand, this excludes point-like charges represented by Dirac measures. It is however possible to extend our analysis to point-like nuclei, by reasoning as in [18, section 3.2.2].

The following result is classical. We will however provide a proof of it in Section 2.4 for the sake of completeness.

Proposition 2.2.1. *Let $\rho^{\rm nuc}\in H^{-1}_{\rm per}(\Gamma)$ and $Q>0$.*

1. *Problem (2.2) has a minimizer u such that $u \in H^3_{\rm per}(\Gamma) \hookrightarrow C^1(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3)$ and $u > 0$ in $\mathbb{R}^3$. The function u satisfies the Euler equation*

$$-C_W \Delta u + \frac{5}{3} C_{\rm TF} u^{7/3} + \left(G_{\mathcal{R}} \star_{\mathcal{R}} (u^2 - \rho^{\rm nuc})\right) u = \epsilon_{\rm F} u, \tag{2.3}$$

where $\epsilon_{\rm F}$ is the Lagrange multiplier of the constraint $\int_\Gamma u^2 = Q$.

2. *The functions u and $-u$ are the only two minimizers of problem (2.2).*

As a consequence of Proposition 2.2.1, the ground state electronic density is always uniquely defined in the framework of the periodic TFW model.

2.3 The Thomas-Fermi-von Weiszäcker model for crystals

We now focus on crystals. More precisely, we consider two kind of systems:

- a reference $\mathcal{R}_1$-periodic perfect crystal with nuclear distribution

$$\rho^{\rm nuc}_{\rm per} \in H^{-1}_{\rm per}(\Gamma_1),$$

  where Γ_1 is the Wigner-Seitz cell of $\mathcal{R}_1$;

- a perturbation of the previous system characterized by the nuclear distribution

$$\rho^{\rm nuc} = \rho^{\rm nuc}_{\rm per} + \nu \quad \text{with} \quad \nu \in \mathcal{C}, \tag{2.4}$$

  $\mathcal{C}$ denoting the Coulomb space already mentioned in the introduction.

It is defined as

$$\mathcal{C} := \left\{ f \in \mathcal{S}'(\mathbb{R}^3) \mid \widehat{f} \in L^1_{\rm loc}(\mathbb{R}^3),\ |\cdot|^{-1} \widehat{f}(\cdot) \in L^2(\mathbb{R}^3) \right\}, \tag{2.5}$$

where $\widehat{f}$ is the Fourier transform of f, normalized in such a way that $\|\widehat{f}\|_{L^2(\mathbb{R}^3)} = \|f\|_{L^2(\mathbb{R}^3)}$ for all $f \in L^2(\mathbb{R}^3)$. Endowed with the inner product

$$D(f,g) := 4\pi \int_{\mathbb{R}^3} \frac{\overline{\widehat{f}(k)}\, \widehat{g}(k)}{|k|^2}\, dk,$$

$\mathcal{C}$ is a Hilbert space. It holds $L^{6/5}(\mathbb{R}^3) \subset \mathcal{C}$ with dense embedding and

$$\forall (f,g) \in L^{6/5}(\mathbb{R}^3) \times L^{6/5}(\mathbb{R}^3), \quad D(f,g) = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{f(x)\, g(x')}{|x - x'|}\, dx\, dx'.$$

We recognize in the right-hand-side the usual expression of the Coulomb interaction of the charge distributions f and g. Let

$$\mathcal{C}' := \left\{ V \in L^6(\mathbb{R}^3) \mid \nabla V \in L^2(\mathbb{R}^3) \right\}.$$

Endowed with the inner product

$$(V,W)_{\mathcal{C}'} := \frac{1}{4\pi} \int_{\mathbb{R}^3} \nabla V \cdot \nabla W = \frac{1}{4\pi} \int_{\mathbb{R}^3} |k|^2\, \overline{\widehat{V}(k)}\, \widehat{W}(k)\, dk,$$

$\mathcal{C}'$ is a Hilbert space, which can be identified with the dual of $\mathcal{C}$ by extention to $\mathcal{C}' \times \mathcal{C}$ of the bilinear form $\langle \cdot, \cdot \rangle_{\mathcal{C}',\mathcal{C}}$ defined on $\mathcal{C}' \times L^{6/5}(\mathbb{R}^3)$ by

$$\forall (V,\rho) \in \mathcal{C}' \times \mathcal{C}, \quad \langle V, \rho \rangle_{\mathcal{C}',\mathcal{C}} = \int_{\mathbb{R}^3} \rho V.$$

The operator $-(4\pi)^{-1}\Delta$ then is a bijective isometry from $\mathcal{C}'$ to $\mathcal{C}$.

2.3.1 Reference perfect crystal

It is shown in [50] that the ground state electronic density $\rho^0_{\rm per}$ of a crystal with nuclear charge distribution $\rho^{\rm nuc}_{\rm per} \in H^{-1}_{\rm per}(\Gamma_1)$ can be identified by a thermodynamic limit argument. It is given by $\rho^0_{\rm per} = |u^0_{\rm per}|^2$ where $u^0_{\rm per} \geq 0$ is obtained by solving the minimization problem

$$I_{\mathcal{R}_1}(\rho^{\rm nuc}_{\rm per}, Z) = \inf \left\{ E^{\rm TFW}_{\mathcal{R}_1}(\rho^{\rm nuc}_{\rm per}, v),\ v \in H^1_{\rm per}(\Gamma_1),\ \int_{\Gamma_1} v^2 = Z \right\}, \tag{2.6}$$

where

$$Z = \int_{\Gamma_1} \rho^{\rm nuc}_{\rm per}. \tag{2.7}$$

Note that problem (2.6) has a unique minimizer (up to the sign) for any value of Z. The correct value of Z given by (2.7) is obtained in [50] by a thermodynamic limit argument. As expected, this value implies the charge neutrality condition

$$\int_{\Gamma_1} (\rho^{\rm nuc}_{\rm per} - \rho^0_{\rm per}) = 0. \tag{2.8}$$

The unique non-negative minimizer $u^0_{\rm per}$ to (2.6)-(2.7) satisfies the Euler equation

$$-C_{\rm W}\Delta u^0_{\rm per} + \frac{5}{3} C_{\rm TF} (\rho^0_{\rm per})^{2/3} u^0_{\rm per} + \left(G_{\mathcal{R}_1} \star_{\mathcal{R}_1} (\rho^0_{\rm per} - \rho^{\rm nuc}_{\rm per}) \right) u^0_{\rm per} = \epsilon^0_{\rm F} u^0_{\rm per}, \tag{2.9}$$

where $\epsilon^0_{\rm F}$, the Lagrange multiplier of the charge constraint, which is uniquely defined, is called the Fermi level of the crystal. From (2.8), we infer that the Coulomb potential $V^0_{\rm per} = G_{\mathcal{R}_1} \star_{\mathcal{R}_1} (\rho^0_{\rm per} - \rho^{\rm nuc}_{\rm per})$ is the unique solution in $H^1_{\rm per}(\Gamma_1)$ to the $\mathcal{R}_1$-periodic Poisson problem

$$\left\{ \begin{array}{l} -\Delta V^0_{\rm per} = \dfrac{4\pi}{|\Gamma_1|} \left(\rho^0_{\rm per} - \rho^{\rm nuc}_{\rm per} \right), \\ V^0_{\rm per}\ \mathcal{R}_1\text{-periodic}, \displaystyle\int_{\Gamma_1} V^0_{\rm per} = 0. \end{array} \right.$$

Using Proposition 2.2.1, we obtain that $u^0_{\rm per} \in C^1(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3)$, and that $u^0_{\rm per} > 0$ in $\mathbb{R}^3$. We thus have the following bounds, that will be useful in our analysis:

$$\exists 0 < m \leq M < +\infty \quad \text{s.t.} \quad \forall x \in \mathbb{R}^3, \quad m \leq u^0_{\rm per}(x) \leq M. \tag{2.10}$$

Let us denote by $H^0_{\rm per}$ the periodic Schrödinger operator on $L^2(\mathbb{R}^3)$ with domain $H^2(\mathbb{R}^3)$ and form domain $H^1(\mathbb{R}^3)$ defined by

$$\forall v \in H^2(\mathbb{R}^3), \quad H^0_{\rm per} v = -C_{\rm W}\Delta v + \frac{5}{3} C_{\rm TF} (\rho^0_{\rm per})^{2/3} v + V^0_{\rm per} v.$$

It is classical (see e.g. [163]) that $H^0_{\rm per}$ is self-adjoint and bounded from below, and that its spectrum is purely absolutely continuous and made of a union of bands. For convenience, we will make the abuse of notation consisting in denoting by $H^0_{\rm per} v$ the distribution

$$H^0_{\rm per} v := -C_{\rm W}\Delta v + \frac{5}{3} C_{\rm TF} (\rho^0_{\rm per})^{2/3} v + V^0_{\rm per} v,$$

which is well-defined for any $v \in L^{6/5}_{\rm loc}(\mathbb{R}^3)$, and belongs to $H^{-1}(\mathbb{R}^3)$ if $v \in H^1(\mathbb{R}^3)$ and to $H^{-1}_{\rm per}(\Gamma)$ if $v \in H^1_{\rm per}(\Gamma)$. We can thus rewrite equation (2.9) under the form

$$H^0_{\rm per} u^0_{\rm per} = \epsilon^0_F u^0_{\rm per}. \tag{2.11}$$

Actually, $\epsilon_{\rm F}^0$ is in fact the minimum of the spectrum σ of the periodic Schrödinger operator $H_{\rm per}^0$ (that is the bottom of the lowest energy band). To see that, let us introduce, for $L \in \mathbb{N}^*$, the spectrum σ_L of $H_{\rm per}^0$, considered as an operator on $L^2_{\rm per}(L\Gamma_1)$. For each L, $u_{\rm per}^0$ is in $L^2_{\rm per}(L\Gamma_1)$ and therefore ϵ_F^0 belongs to σ_L. Using the fact that $u_{\rm per}^0 > 0$ in $\mathbb{R}^3$, it is easy to see that $u_{\rm per}^0$ is actually a ground state, which means that $\epsilon_F^0 = \min \sigma_L$. It then readily follows from Bloch-Floquet theory (see e.g. [163]) that

$$\sigma = \overline{\bigcup_{L\in\mathbb{N}^*} \sigma_L}.$$

Therefore $\epsilon_{\rm F}^0 = \min \sigma$. As a consequence,

$$\forall v \in H^1(\mathbb{R}^3), \quad \langle (H_{\rm per}^0 - \epsilon_{\rm F}^0)v, v\rangle_{H^{-1}(\mathbb{R}^3),H^1(\mathbb{R}^3)} \geq 0. \tag{2.12}$$

2.3.2 Crystals with local defects

We now consider a crystal with a local defect, whose nuclear charge distribution is given by (2.4). It is convenient to describe the TFW electronic state of this system by a function v related to the electronic density ρ by the relation

$$v = \sqrt{\rho} - u_{\rm per}^0. \tag{2.13}$$

Denoting by

$$E^{\rm TFW}(\rho^{\rm nuc}, w) = C_{\rm W}\int_{\mathbb{R}^3} |\nabla w|^2 + C_{\rm TF}\int_{\mathbb{R}^3} |w|^{10/3} + \frac{1}{2}D(\rho^{\rm nuc} - w^2, \rho^{\rm nuc} - w^2)$$

the TFW energy functional of a finite molecular system *in vacuo* with nuclear charge $\rho^{\rm nuc}$, we can formally define the relative energy (with respect to the perfect crystal) of the system with nuclear charge density $\rho_{\rm per}^{\rm nuc} + \nu$ and electronic density $\rho = (u_{\rm per}^0 + v)^2$ as

$$\begin{aligned} &E^{\rm TFW}(\rho_{\rm per}^{\rm nuc} + \nu, u_{\rm per}^0 + v) - E^{\rm TFW}(\rho_{\rm per}^{\rm nuc}, u_{\rm per}^0) \\ &\quad = \langle (H_{\rm per}^0 - \epsilon_{\rm F}^0)v, v\rangle + C_{\rm TF}\int_{\mathbb{R}^3} \left(|u_{\rm per}^0 + v|^{10/3} - |u_{\rm per}^0|^{10/3} - \frac{5}{3}|u_{\rm per}^0|^{4/3}(2u_{\rm per}^0 v + v^2)\right) \\ &\qquad + \frac{1}{2}D\left(2u_{\rm per}^0 v + v^2 - \nu, 2u_{\rm per}^0 v + v^2 - \nu\right) - \int_{\mathbb{R}^3} \nu V_{\rm per}^0 + \epsilon_{\rm F}^0 q, \end{aligned} \tag{2.14}$$

where

$$q = \int_{\mathbb{R}^3} \left(|u_{\rm per}^0 + v|^2 - |u_{\rm per}^0|^2 \right). \tag{2.15}$$

Of course, the left-hand side of (2.14) is a formal expression since it is the difference of two quantities taking the value plus infinity. On the other hand, the right-hand side of (2.14) is mathematically well-defined as soon as q is a fixed real number and $v \in \mathcal{Q}_+$, where

$$\mathcal{Q}_+ := \left\{ v \in H^1(\mathbb{R}^3) \mid v \geq -u_{\rm per}^0,\ u_{\rm per}^0 v \in \mathcal{C} \right\}.$$

Indeed, the first three terms of the right-hand side of (2.14) are non-negative, and are finite if and only if $v \in H^1(\mathbb{R}^3)$ and $u_{\rm per}^0 v \in \mathcal{C}$. Lastly, the requirement $v \geq -u_{\rm per}^0$ follows from (2.13). The set $\mathcal{Q}_+$ is a closed convex subset of the Hilbert space

$$\mathcal{Q} := \left\{ v \in H^1(\mathbb{R}^3) \mid u_{\rm per}^0 v \in \mathcal{C} \right\},$$

endowed with the inner product defined by

$$(v, w)_{\mathcal{Q}} := (v, w)_{H^1(\mathbb{R}^3)} + D(u_{\rm per}^0 v, u_{\rm per}^0 w).$$

This formal analysis leads us to propose the following model, which will be justified in the following section by means of thermodynamic limit arguments: the ground state electronic density of the perturbed crystal characterized by the nuclear charge density (2.4) is given by

$$\rho_\nu = (u^0_{\rm per} + v_\nu)^2,$$

where v_ν is a minimizer of

$$I^\nu = \inf\left\{\mathcal{E}^\nu(v),\ v \in \mathcal{Q}_+\right\} \tag{2.16}$$

with

$$\begin{aligned}\mathcal{E}^\nu(v) \ :=\ & \langle (H^0_{\rm per} - \epsilon^0_{\rm F})v, v\rangle_{H^{-1}(\mathbb{R}^3),H^1(\mathbb{R}^3)} \\ & + C_{\rm TF}\int_{\mathbb{R}^3}\left(|u^0_{\rm per}+v|^{10/3} - |u^0_{\rm per}|^{10/3} - \frac{5}{3}|u^0_{\rm per}|^{4/3}(2u^0_{\rm per}v + v^2)\right) \\ & + \frac{1}{2}D\left(2u^0_{\rm per}v + v^2 - \nu, 2u^0_{\rm per}v + v^2 - \nu\right).\end{aligned} \tag{2.17}$$

The following result, whose proof is postponed until Section 2.4, shows that our model is well-posed.

Theorem 2.3.1. *Let $\nu \in \mathcal{C}$. Then,*

1. Existence and uniqueness of the ground state density. *Problem (2.16) has a unique minimizer v_ν. The function v_ν is such that $u^0_{\rm per} + v_\nu > 0$ in $\mathbb{R}^3$ and satisfies the Euler equation*

$$\begin{aligned}(H^0_{\rm per} - \epsilon^0_F)v_\nu + \frac{5}{3}C_{\rm TF}\left(|u^0_{\rm per}+v_\nu|^{7/3} - |u^0_{\rm per}|^{7/3} - |u^0_{\rm per}|^{4/3}v_\nu\right) \\ + \left((2u^0_{\rm per}v_\nu + v_\nu^2 - \nu)\star|\cdot|^{-1}\right)(u^0_{\rm per}+v_\nu) = 0.\end{aligned} \tag{2.18}$$

It holds $v_\nu \in \mathcal{Q}\cap H^3(\mathbb{R}^3)$ and there exists some constant $C \in \mathbb{R}_+$ such that

$$\forall \nu \in \mathcal{C}, \qquad \|v_\nu\|_{\mathcal{Q}} \le C\left(\|\nu\|_{\mathcal{C}} + \|\nu\|^2_{\mathcal{C}}\right), \tag{2.19}$$

$$\|v_\nu\|_{H^1(\mathbb{R}^3)} \le C\|\nu\|_{\mathcal{C}}, \tag{2.20}$$

$$\|v_\nu\|_{H^2(\mathbb{R}^3)} \le C\left(\|\nu\|_{\mathcal{C}} + \|\nu\|^3_{\mathcal{C}}\right), \tag{2.21}$$

$$\|v_\nu\|_{H^3(\mathbb{R}^3)} \le C\left(\|\nu\|_{\mathcal{C}} + \|\nu\|^5_{\mathcal{C}}\right). \tag{2.22}$$

2. Neutrality of local defects. *Let us denote by $\rho^0_\nu = \nu - (2u^0_{\rm per}v_\nu + v_\nu^2)$ the total density of charge of the defect. It holds*

$$\lim_{r\to 0}\frac{1}{|B_r|}\int_{B_r}|\widehat{\rho^0_\nu}(k)|\,dk = 0. \tag{2.23}$$

In addition, the Coulomb potential $\Phi^0_\nu = \rho^0_\nu \star |\cdot|^{-1}$ generated by ρ^0_ν is in $L^2(\mathbb{R}^3)\cap\mathcal{C}'$, and there exists a constant $C \in \mathbb{R}_+$ such that

$$\forall \nu \in \mathcal{C}, \qquad \|\Phi^0_\nu\|_{\mathcal{C}'} \le C\left(\|\nu\|_{\mathcal{C}} + \|\nu\|^2_{\mathcal{C}}\right), \tag{2.24}$$

$$\|\Phi^0_\nu\|_{L^2(\mathbb{R}^3)} \le C\left(\|\nu\|_{\mathcal{C}} + \|\nu\|^3_{\mathcal{C}}\right), \tag{2.25}$$

3. Compactness of the minimizing sequences. *Any minimizing sequence* $(v_n)_{n\in\mathbb{N}}$ *for (2.16) converges to* v_ν *weakly in* $H^1(\mathbb{R}^3)$ *and strongly in* $L^p_{\rm loc}(\mathbb{R}^3)$ *for* $1 \leq p < 6$. *Besides,* $(u^0_{\rm per}v_n)_{n\in\mathbb{N}}$ *converges to* $u^0_{\rm per}v_\nu$ *weakly in* $\mathcal{C}$.

For any $q \in \mathbb{R}$, *there exists a minimizing sequence* $(v_n)_{n\in\mathbb{N}}$ *for (2.16) consisting of functions of* $\mathcal{Q}_+ \cap L^1(\mathbb{R}^3)$ *such that*

$$\forall n \in \mathbb{N}, \quad \int_{\mathbb{R}^3} \left(|u^0_{\rm per} + v_n|^2 - |u^0_{\rm per}|^2\right) = q. \tag{2.26}$$

We conclude this section with some physical considerations.

Let $\nu \in L^1(\mathbb{R}^3) \cap L^2(\mathbb{R}^3)$. Assuming that $v_\nu \in L^1(\mathbb{R}^3) \cap L^2(\mathbb{R}^3)$ (a property satisfied at least in the special case of a homogeneous host crystal, see Section 2.3.4), then $\widehat{\rho^0_\nu} \in C^0(\mathbb{R}^3)$ and (2.23) simply means that the continuous function $\widehat{\rho^0_\nu}$ vanishes at $k = 0$, or equivalently that

$$\int_{\mathbb{R}^3} \rho^0_\nu = 0. \tag{2.27}$$

The property (2.23) means that 0 is a Lebesgue point of $\widehat{\rho^0_\nu}$ and that the Lebesgue value of $\widehat{\rho^0_\nu}$ at 0 is equal to zero. It can therefore be interpreted as a weak form of the neutrality condition (2.27), also valid when $\rho^0_\nu \notin L^1(\mathbb{R}^3)$. Besides, the fact that the Coulomb potential Φ^0_ν belongs to $L^2(\mathbb{R}^3)$ implies in particular that $\Phi^0_\nu = \rho^0_\nu \star |\cdot|^{-1}$ cannot decay as $Q|\cdot|^{-1}$ with $Q \neq 0$ at infinity, which, in some sense, constitutes another weak form of the neutrality condition (2.27).

It is interesting to compare our result on the neutrality of local defects in crystals with the result by Solovej about the asymptotic negative ionization of a molecular system when the nuclear charge goes to infinity [170, Theorem 2]. In the latter setting, the number K and the locations $(R_1, \cdots, R_K)$ of the nuclei are fixed. The asymptotic limit considered is obtained by letting the nuclear charges $\underline{z}' = (z_1, \cdots, z_L)$ of the first L nuclei $(1 \leq L \leq K)$ go to infinity, the remaining $K - L$ nuclear charges $\underline{z}'' = (z_{L+1}, \cdots, z_K)$ being kept fixed. Denoting by $Q_c(\underline{z}', \underline{z}'')$ the charge of the maximally ionized molecule with nuclear charge $(\underline{z}', \underline{z}'')$ (i.e. of the lowest energy stable molecular system with nuclear charges $(\underline{z}', \underline{z}'')$), it is proved in [170] that

$$\lim_{\underline{z}' \to \infty} Q_c(\underline{z}', \underline{z}'') := Q_\infty(\underline{z}'') < 0.$$

In the case of a crystal with a local defect, the total nuclear charge also goes to infinity, but in a different way. The charge does not accumulate at some points $(R_1, \cdots, R_L)$ as in [170]; it is spread in the whole physical space. We will elaborate further on the fundamental differences between maximally ionized molecular systems and crystals with defects in Section 2.3.4.

The third statement of Theorem 2.3.1 implies that there is no way to model a charged defect within the TFW theory: loosely speaking, if we try to put too many (or not enough) electrons in the system, the electronic density will relax to $(u^0_{\rm per} + v_\nu)^2$ and the remaining (or missing) $q - \int_{\mathbb{R}^3} \nu$ electrons will escape to (or come from) infinity with an energy $\epsilon^0_{\rm F}$.

2.3.3 Thermodynamic limit

The purpose of this section is to provide a mathematical justification of the model (2.16). We consider a crystal with a local defect characterized by the nuclear charge distribution $\rho^{\rm nuc} = \rho^{\rm nuc}_{\rm per} + \nu$ and, in order to avoid additional technical difficulties, we assume that $\nu \in L^1(\mathbb{R}^3) \cap L^2(\mathbb{R}^3)$.

In numerical simulations, the TFW ground state electronic density of such a system is usually computed with the supercell method. For a given $L \in \mathbb{N}$ large enough, the supercell model of size L is the periodic TFW model (2.2) with

$$\mathcal{R} = \mathcal{R}_L := L\mathcal{R}_1, \quad \Gamma = \Gamma_L := L\Gamma_1, \quad \rho^{\rm nuc} = \rho^{\rm nuc}_{\rm per} + \nu_L, \quad Q = Z\,L^3 + q, \tag{2.28}$$

where

$$\nu_L(x) = \sum_{z \in \mathcal{R}_L} (\chi_{\Gamma_L}\nu)(x-z),$$

$\chi_{\Gamma_L} : \mathbb{R}^3 \to \mathbb{R}$ denoting the characteristic function of the supercell Γ_L. Note that ν_L is the unique $\mathcal{R}_L$-periodic function such that $\nu_L|_{\Gamma_L} = \nu|_{\Gamma_L}$. In practice, L is chosen as large as possible (given the computational means available) to limit the error originating from the artificial periodic boundary conditions.

It is important to note that $u^0_{\rm per}$ is the unique minimizer (up to the sign) of the supercell model of size L for $\rho^{\rm nuc} = \rho^{\rm nuc}_{\rm per}$ and $Q = ZL^3$, whatever $L \in \mathbb{N}^*$. Reasoning as in the previous section, we introduce the energy functional

$$\begin{aligned}
\mathcal{E}^\nu_L(v_L) \;:=\; & \langle (H^0_{\rm per} - \epsilon^0_{\rm F})v_L, v_L\rangle_{H^{-1}_{\rm per}(\Gamma_L), H^1_{\rm per}(\Gamma_L)} \\
& + C_{\rm TF}\int_{\Gamma_L}\left(|u^0_{\rm per}+v_L|^{10/3} - |u^0_{\rm per}|^{10/3} - \frac{5}{3}|u^0_{\rm per}|^{4/3}(2u^0_{\rm per}v_L + v_L^2)\right) \\
& + \frac{1}{2}D_{\mathcal{R}_L}\left(2u^0_{\rm per}v_L + v_L^2 - \nu_L, 2u^0_{\rm per}v_L + v_L^2 - \nu_L\right),
\end{aligned} \tag{2.29}$$

which is such that

$$E^{\rm TFW}_{\mathcal{R}_L}(\rho^{\rm nuc}_{\rm per} + \nu_L, u^0_{\rm per} + v_L) - E^{\rm TFW}_{\mathcal{R}_L}(\rho^{\rm nuc}_{\rm per}, u^0_{\rm per}) = \mathcal{E}^\nu_L(v_L) - \int_{\Gamma_L}\nu_L V^0_{\rm per} + \epsilon^0_{\rm F} q, \tag{2.30}$$

with

$$q = \int_{\Gamma_L}\left(|u^0_{\rm per}+v_L|^2 - |u^0_{\rm per}|^2\right) = \int_{\Gamma_L}(2u^0_{\rm per}v_L + v_L^2). \tag{2.31}$$

While (2.14) and (2.15) are formal expressions, (2.30) and (2.31) are well-defined mathematical expressions. The ground state electronic density of the supercell model for the data defined by (2.28) is therefore obtained as

$$\rho^{0,\nu,q}_L = (u^0_{\rm per} + v_{\nu,q,L})^2$$

where $v_{\nu,q,L}$ is a minimizer of

$$I^{\nu,q}_L = \inf\left\{\mathcal{E}^\nu_L(v_L),\ v_L \in \mathcal{Q}_{+,L},\ \int_{\Gamma_L}(2u^0_{\rm per}v_L + v_L^2) = q\right\}, \tag{2.32}$$

$\mathcal{Q}_{+,L}$ denoting the convex set

$$\mathcal{Q}_{+,L} = \left\{v_L \in H^1_{\rm per}(\Gamma_L) \mid v_L \geq -u^0_{\rm per}\right\}.$$

We also introduce the minimization problem

$$I^\nu_L = \inf\left\{\mathcal{E}^\nu_L(v_L),\ v_L \in \mathcal{Q}_{+,L}\right\}, \tag{2.33}$$

in which we do not *a priori* impose the electronic charge in the supercell.

Theorem 2.3.2. *Let $\nu \in L^1(\mathbb{R}^3) \cap L^2(\mathbb{R}^3)$.*

1. Thermodynamic limit with charge constraint. *For each $q \in \mathbb{R}$ and each $L \in \mathbb{N}^*$, the minimization problem (2.32) has a unique minimizer $v_{\nu,q,L}$. For each $q \in \mathbb{R}$, the sequence $(v_{\nu,q,L})_{L\in\mathbb{N}^*}$ converges, weakly in $H^1_{\rm loc}(\mathbb{R}^3)$, and strongly in $L^p_{\rm loc}(\mathbb{R}^3)$ for all $1 \le p < 6$, towards v_ν, the unique minimizer of problem (2.16). For each $q \in \mathbb{R}$ and each $L \in \mathbb{N}^*$, $v_{\nu,q,L}$ satisfies the Euler equation*

$$\begin{aligned}&(H^0_{\rm per} - \epsilon^0_F)v_{\nu,q,L} + \frac{5}{3}C_{\rm TF}\left(|u^0_{\rm per} + v_{\nu,q,L}|^{7/3} - |u^0_{\rm per}|^{7/3} - |u^0_{\rm per}|^{4/3}v_{\nu,q,L}\right)\\ &+ \left((2u^0_{\rm per}v_{\nu,q,L} + v^2_{\nu,q,L} - \nu_L) \star_{\mathcal{R}_L} G_{\mathcal{R}_L}\right)(u^0_{\rm per} + v_{\nu,q,L}) = \mu_{\nu,q,L}(u^0_{\rm per} + v_{\nu,q,L}),\end{aligned} \tag{2.34}$$

where $\mu_{\nu,q,L} \in \mathbb{R}$ is the Lagrange multiplier of the constraint $\int_{\Gamma_L}(2u^0_{\rm per}v_{\nu,q,L} + v^2_{\nu,q,L}) = q$, and it holds $\lim_{L\to\infty} \mu_{\nu,q,L} = 0$ for each $q \in \mathbb{R}$.

2. Thermodynamic limit without charge constraint. *For each $L \in \mathbb{N}^*$, the minimization problem (2.33) has a unique minimizer $v_{\nu,L}$. It holds*

$$\begin{aligned}&(H^0_{\rm per} - \epsilon^0_F)v_{\nu,L} + \frac{5}{3}C_{\rm TF}\left(|u^0_{\rm per} + v_{\nu,L}|^{7/3} - |u^0_{\rm per}|^{7/3} - |u^0_{\rm per}|^{4/3}v_{\nu,L}\right)\\ &+ \left((2u^0_{\rm per}v_{\nu,L} + v^2_{\nu,L} - \nu_L) \star_{\mathcal{R}_L} G_{\mathcal{R}_L}\right)(u^0_{\rm per} + v_{\nu,L}) = 0.\end{aligned} \tag{2.35}$$

The sequence $(v_{\nu,L})_{L\in\mathbb{N}^*}$ also converges to v_ν, weakly in $H^1_{\rm loc}(\mathbb{R}^3)$, and strongly in $L^p_{\rm loc}(\mathbb{R}^3)$ for all $1 \le p < 6$. Besides,

$$\int_{\Gamma_L}\left(\nu_L - (2u^0_{\rm per}v_{\nu,L} + v^2_{\nu,L})\right) \underset{L\to\infty}{\longrightarrow} 0.$$

2.3.4 The special case of homogeneous host crystals

In this section, we address the special case when the host crystal is a homogeneous medium completely characterized by the positive real number α such that

$$\forall x \in \mathbb{R}^3, \quad \rho^{\rm nuc}_{\rm per}(x) = \rho^0_{\rm per}(x) = \alpha^2 \quad \text{and} \quad u^0_{\rm per}(x) = \alpha. \tag{2.36}$$

In this case, analytical expressions for the linear response can be derived, leading to the following result.

Theorem 2.3.3. *Assume that (2.36) holds for some $\alpha > 0$. For each $\nu \in \mathcal{C}$, the unique minimizer v_ν of (2.16) can be expanded as*

$$v_\nu = g \star \nu + \widetilde{r}_2(\nu) \tag{2.37}$$

where $g \in L^1(\mathbb{R}^3)$ is characterized by its Fourier transform

$$\widehat{g}(k) = \frac{1}{(2\pi)^{3/2}}\,\frac{4\pi\alpha}{C_{\rm W}|k|^4 + \frac{20}{9}C_{\rm TF}\alpha^{4/3}|k|^2 + 8\pi\alpha^2},$$

and where $\widetilde{r}_2(\nu) \in L^1(\mathbb{R}^3)$. In addition, $(g\star\nu) \in H^3(\mathbb{R}^3)\cap\mathcal{C}$, $\widetilde{r}_2(\nu) \in H^3(\mathbb{R}^3)\cap\mathcal{C}\cap L^1(\mathbb{R}^3)$, and there exists a constant $C \in \mathbb{R}_+$ such that

$$\forall \nu \in \mathcal{C}, \qquad \|g\star\nu\|_{H^3(\mathbb{R}^3)\cap\mathcal{C}} \le C\|\nu\|_{\mathcal{C}}, \tag{2.38}$$

$$\|\widetilde{r}_2(\nu)\|_{H^2(\mathbb{R}^3)\cap\mathcal{C}} \le C\left(\|\nu\|^2_{\mathcal{C}} + \|\nu\|^3_{\mathcal{C}}\right), \tag{2.39}$$

$$\|\widetilde{r}_2(\nu)\|_{L^1(\mathbb{R}^3)} \le C\left(\|\nu\|^2_{\mathcal{C}} + \|\nu\|^4_{\mathcal{C}}\right), \tag{2.40}$$

$$\|\widetilde{r}_2(\nu)\|_{H^3(\mathbb{R}^3)} \le C\left(\|\nu\|^2_{\mathcal{C}} + \|\nu\|^5_{\mathcal{C}}\right). \tag{2.41}$$

If $\nu \in L^1(\mathbb{R}^3) \cap \mathcal{C}$, then $(g \star \nu) \in L^1(\mathbb{R}^3)$,

$$\|g \star \nu\|_{L^1(\mathbb{R}^3)} \leq \|g\|_{L^1(\mathbb{R}^3)} \|\nu\|_{L^1(\mathbb{R}^3)}, \tag{2.42}$$

$v_\nu \in L^1(\mathbb{R}^3) \cap L^2(\mathbb{R}^3)$ and

$$\int_{\mathbb{R}^3} (\nu - (2u^0_{\rm per} v_\nu + v_\nu^2)) = 0. \tag{2.43}$$

Note that for homogeneous host crystals, $\mathcal{Q} = H^1(\mathbb{R}^3) \cap \mathcal{C}$ since $u^0_{\rm per}$ is a constant function. The estimates (2.38) and (2.39) therefore provide bounds for the $\mathcal{Q}$-norm. It also follows from (2.38)-(2.42) that the first term of the right hand side of (2.37) is in fact the linear component of the application $\nu \mapsto v_\nu$. The second term collects the contributions of higher orders.

Proof. In the special case under consideration, the Euler equation (2.18) also reads

$$-C_{\rm W}\Delta v_\nu + \frac{20}{9} C_{\rm TF} \alpha^{4/3} v_\nu + 2\alpha^2 \left(v_\nu \star |\cdot|^{-1}\right) = \alpha \left(\nu \star |\cdot|^{-1}\right) - \alpha \left(v_\nu^2 \star |\cdot|^{-1}\right) + \kappa_\nu, \tag{2.44}$$

where

$$\kappa_\nu = -\frac{5}{3} C_{\rm TF} \left(|\alpha + v_\nu|^{7/3} - \alpha^{7/3} - \frac{7}{3}\alpha^{4/3} v_\nu \right) + \Phi^0_\nu v_\nu$$

and

$$\Phi^0_\nu = (\nu - 2\alpha v_\nu - v_\nu^2) \star |\cdot|^{-1}. \tag{2.45}$$

We therefore obtain (2.37) with

$$\widetilde{r}_2(\nu) = -g \star v_\nu^2 + h \star \kappa_\nu, \tag{2.46}$$

the convolution kernel h being defined through its Fourier transform as

$$\widehat{h}(k) = \frac{1}{(2\pi)^{3/2}} \frac{|k|^2}{C_{\rm W}|k|^4 + \frac{20}{9} C_{\rm TF} \alpha^{4/3} |k|^2 + 8\pi\alpha^2}.$$

Let $\alpha_0 = \left(\frac{162\pi C_{\rm W}}{25 C_{\rm TF}^2}\right)^{3/2}$, $\beta_0 = \left(\frac{10 C_{\rm TF}}{9 C_{\rm W}}\right)^{1/2}$, and

$$\zeta_\pm(\alpha) := \begin{cases} \beta_0 \alpha^{2/3} \left(1 \pm \left(1 - (\alpha_0/\alpha)^{2/3}\right)^{1/2}\right)^{1/2} & \text{if } \alpha > \alpha_0, \\ \beta_0 \alpha^{2/3} \left(\left(\frac{1}{2}\left((\alpha_0/\alpha)^{1/3} + 1\right)\right)^{1/2} \pm i \left(\frac{1}{2}\left((\alpha_0/\alpha)^{1/3} - 1\right)\right)^{1/2} \right) & \text{if } 0 < \alpha < \alpha_0. \end{cases}$$

For all $\zeta \in \mathbb{C}$ such that $\Re(\zeta) > 0$, we denote by $Y_\zeta(x) = \frac{e^{-\zeta|x|}}{|x|}$ the Yukawa potential solution to

$$-\Delta Y_\zeta + \zeta^2 Y_\zeta = 4\pi\delta_0.$$

Noticing that

$$\widehat{g}(k) = \frac{4\pi\alpha}{(2\pi)^{3/2} C_{\rm W}} \frac{1}{|k|^2 + \zeta_+(\alpha)^2} \frac{1}{|k|^2 + \zeta_-(\alpha)^2} = \frac{\alpha}{4\pi C_{\rm W}} (2\pi)^{3/2} \widehat{Y}_{\zeta_+(\alpha)}(k)\, \widehat{Y}_{\zeta_-(\alpha)}(k),$$

we obtain

$$g = \frac{\alpha}{4\pi C_{\rm W}} Y_{\zeta_+(\alpha)} \star Y_{\zeta_-(\alpha)} \in L^1(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3).$$

Besides, $\widehat{h}(k) = \frac{|k|^2}{4\pi\alpha}\widehat{g}(k)$, from which we infer

$$h = -\frac{1}{4\pi\alpha}\Delta g = \frac{1}{8\pi C_{\rm W}}\left(Y_{\zeta_-(\alpha)} + Y_{\zeta_+(\alpha)}\right) - \frac{\beta_0^2\alpha^{1/3}}{4\pi}g \in L^1(\mathbb{R}^3).$$

Note that g and h are real valued (even if $0 < \alpha < \alpha_0$), and decay exponentially at infinity. It then follows from Theorem 2.3.1, Lemma 2.4.1, and Lemma 2.4.2 below that $\kappa_\nu \in L^1(\mathbb{R}^3) \cap H^1(\mathbb{R}^3)$ with

$$\|\kappa_\nu\|_{L^2(\mathbb{R}^3)} \leq C\left(\|\nu\|_{\mathcal{C}}^2 + \|\nu\|_{\mathcal{C}}^3\right), \tag{2.47}$$

$$\|\kappa_\nu\|_{L^1(\mathbb{R}^3)} \leq C\left(\|\nu\|_{\mathcal{C}}^2 + \|\nu\|_{\mathcal{C}}^4\right), \tag{2.48}$$

$$\|\kappa_\nu\|_{H^1(\mathbb{R}^3)} \leq C\left(\|\nu\|_{\mathcal{C}}^2 + \|\nu\|_{\mathcal{C}}^5\right), \tag{2.49}$$

for some constant $C \in \mathbb{R}_+$ independent of ν. For each $\nu \in \mathcal{C}$,

$$\begin{aligned}((1+|k|^2)^3 + |k|^{-2})|\widehat{(g\star\nu)}(k)|^2 &= (2\pi)^3((1+|k|^2)^3 + |k|^{-2})|\widehat{g}(k)|^2|\widehat{\nu}(k)|^2 \\ &= (2\pi)^3((1+|k|^2)^3|k|^2 + 1)|\widehat{g}(k)|^2\frac{|\widehat{\nu}(k)|^2}{|k|^2}.\end{aligned}$$

The function $k \mapsto ((1+|k|^2)^3|k|^2+1)|\widehat{g}(k)|^2$ being bounded, $(g\star\nu) \in H^3(\mathbb{R}^3)\cap\mathcal{C}$ and

$$\|g\star\nu\|_{H^3(\mathbb{R}^3)\cap\mathcal{C}} \leq C\|\nu\|_{\mathcal{C}}.$$

The function v_ν^2 being in $L^1(\mathbb{R}^3)\cap L^3(\mathbb{R}^3) \hookrightarrow L^{6/5}(\mathbb{R}^3) \hookrightarrow \mathcal{C}$, we deduce from the above inequality, Sobolev embeddings and (2.20) that

$$\|g\star v_\nu^2\|_{H^3(\mathbb{R}^3)\cap\mathcal{C}} \leq C\|\nu\|_{\mathcal{C}}^2.$$

On the other hand,

$$\begin{aligned}|k|^{-2}|\widehat{(h\star\kappa_\nu)}(k)|^2 &= (2\pi)^3|k|^{-2}|\widehat{h}(k)|^2|\widehat{\kappa}_\nu(k)|^2 \\ (1+|k|^2)^2|\widehat{(h\star\kappa_\nu)}(k)|^2 &= (2\pi)^3(1+|k|^2)^2|\widehat{h}(k)|^2|\widehat{\kappa}_\nu(k)|^2 \\ (1+|k|^2)^3|\widehat{(h\star\kappa_\nu)}(k)|^2 &= (2\pi)^3(1+|k|^2)^2|\widehat{h}(k)|^2(1+|k|^2)|\widehat{\kappa}_\nu(k)|^2.\end{aligned}$$

The functions $k \mapsto |k|^{-2}|\widehat{h}(k)|^2$ and $k \mapsto (1+|k|^2)^2|\widehat{h}(k)|^2$ being bounded, we infer from (2.47) and (2.49) that $(h\star\kappa_\nu) \in H^3(\mathbb{R}^3)\cap\mathcal{C}$ and that

$$\begin{aligned}\|h\star\kappa_\nu\|_{H^2(\mathbb{R}^3)\cap\mathcal{C}} &\leq C\|\kappa_\nu\|_{L^2(\mathbb{R}^3)} \leq C'\left(\|\nu\|_{\mathcal{C}}^2 + \|\nu\|_{\mathcal{C}}^3\right), \\ \|h\star\kappa_\nu\|_{H^3(\mathbb{R}^3)} &\leq C\|\kappa_\nu\|_{H^1(\mathbb{R}^3)} \leq C\left(\|\nu\|_{\mathcal{C}}^2 + \|\nu\|_{\mathcal{C}}^5\right).\end{aligned}$$

The L^1-bound (2.40) is a straightforward consequence of (2.20), (2.46), (2.48) and Young inequality. Lastly, if $\nu \in L^1(\mathbb{R}^3)$, then $\rho_\nu^0 = \nu - (2\alpha v_\nu + v_\nu^2) \in L^1(\mathbb{R}^3)$, so that $\widehat{\rho_\nu^0}$ is a continuous function. It follows from (2.23) that $\widehat{\rho_\nu^0}(0) = 0$, which readily leads to (2.43). □

Remark 2.3.1. *For a generic $\nu \in \mathcal{C}$, the function v_ν, hence the density $2\alpha v_\nu + v_\nu^2$, are* not *in $L^1(\mathbb{R}^3)$. This follows from the fact that the nonlinear contribution $\widetilde{r}_2(\nu)$ is always in $L^1(\mathbb{R}^3)$, while the linear contribution $g\star\nu$ is not necessarily in $L^1(\mathbb{R}^3)$ since its Fourier transform*

$$\widehat{(g\star\nu)}(k) = \frac{4\pi\alpha}{C_{\rm W}|k|^4 + \frac{20}{9}C_{\rm TF}\alpha^{4/3}|k|^2 + 8\pi\alpha^2}\widehat{\nu}(k)$$

is not necessarily in $L^\infty(\mathbb{R}^3)$.

Let us finally return to the fundamental difference between finite molecular systems and crystals as far as maximal ionization is concerned. Let us fix $\nu \in C_c^\infty(\mathbb{R}^3)$, $\nu \geq 0$, $\nu \neq 0$, and focus on the dependence of the solution to equation (2.44) with respect to α. For this purpose, we add an index α to v_ν and Φ^0_ν, and rewrite (2.44) and (2.45) as

$$\begin{cases} -C_{\mathrm{W}}\Delta v_{\nu,\alpha} + \dfrac{5}{3}C_{\mathrm{TF}}\left(|\alpha + v_{\nu,\alpha}|^{7/3} - \alpha^{7/3} - \alpha^{4/3}v_{\nu,\alpha}\right) - \Phi^0_{\nu,\alpha}(\alpha + v_{\nu,\alpha}) = 0 \\ -\Delta\Phi^0_{\nu,\alpha} = 4\pi\left(\nu - 2\alpha v_{\nu,\alpha} - v^2_{\nu,\alpha}\right) \\ v_{\nu,\alpha} > -\alpha. \end{cases} \tag{2.50}$$

We know from Theorem 2.3.3 that for each $\alpha > 0$,

$$Q_{\nu,\alpha} := \int_{\mathbb{R}^3} \nu - 2\alpha v_{\nu,\alpha} - v^2_{\nu,\alpha} = 0.$$

The TFW electronic ground state of the maximally ionized molecule with nuclear distribution ν is obtained by solving equation (2.50) for $\alpha = 0$, which also reads as

$$\begin{cases} -C_{\mathrm{W}}\Delta v_{\nu,0} + W v_{\nu,0} = 0 \\ W = \dfrac{5}{3}C_{\mathrm{TF}}v^{4/3}_{\nu,0} + \left(v^2_{\nu,0} - \nu\right) \star |\cdot|^{-1} \\ v_{\nu,0} > 0. \end{cases} \tag{2.51}$$

Let

$$Q_{\nu,0} := \int_{\mathbb{R}^3} \nu - v^2_{\nu,0}.$$

Denoting by $[W]_+ = \max(0, [W])$ the positive part of the spherical average $[W]$ of W, it results from Gauss Theorem that, for $|x|$ large enough,

$$[W]_+(x) \leq \frac{5}{3}C_{\mathrm{TF}}\left[v^{4/3}_{\nu,0}\right](x) + \left[\left(v^2_{\nu,0} - \nu\right) \star |\cdot|^{-1}\right]_+(x) \leq \frac{5}{3}C_{\mathrm{TF}}\left[v^{4/3}_{\nu,0}\right](x) + \frac{\max(0, -Q_{\nu,0})}{|x|}.$$

If $Q_{\nu,0}$ were non-negative, $[W]_+$ would be in $L^{3/2}(\mathbb{R}^3)$ and the solution $v_{\nu,0}$ to (2.51) would not be in $L^2(\mathbb{R}^3)$ [135, Lemma 7.18]. Therefore

$$0 = \lim_{\alpha \to 0^+} Q_{\nu,\alpha} > Q_{\nu,0}.$$

This simple argument allows to better understand why the result on the maximal ionization of molecules discussed in Section 2.3.2 does not extend to crystals with local defects.

2.4 Proofs

This section is devoted to the proofs of Proposition 2.2.1, Theorem 2.3.1 and Theorem 2.3.2.

In the sequel, we set

$$C_{\mathrm{TF}} = 1 \quad \text{and} \quad C_{\mathrm{W}} = 1 \qquad \text{(in order to simplify the notation)}.$$

2.4.1 Preliminary results

We first state and prove a few useful lemmas. Some of these results are simple, or well-known, but we nevertheless prove them here for the sake of self-containment.

Lemma 2.4.1. *For all $0 < m \leq M < \infty$ and all $\gamma \geq 2$, there exists $C \in \mathbb{R}_+$ such that for all $m \leq a \leq M$ and all $b \geq -a$,*

$$(\gamma - 1)a^{\gamma-2}b^2 \leq (a+b)^\gamma - a^\gamma - \gamma a^{\gamma-1}b \leq C\left(1 + |b|^{\gamma-2}\right)b^2. \tag{2.52}$$

Proof. Let $\phi(t) = (a+tb)^\gamma$. It holds for all $t \in (0,1)$, $\phi'(t) = \gamma(a+tb)^{\gamma-1}b$ and $\phi''(t) = \gamma(\gamma-1)(a+tb)^{\gamma-2}b^2$. Using the identity

$$\phi(1) - \phi(0) - \phi'(0) = \int_0^1 (1-t)\phi''(t)\,dt,$$

we get

$$(a+b)^\gamma - a^\gamma - \gamma a^{\gamma-1}b \;=\; \gamma(\gamma-1)b^2 \int_0^1 (1-t)(a+tb)^{\gamma-2}\,dt.$$

We obtain (2.52) using the fact that for all $t \in [0,1]$, $a(1-t) \leq a + tb \leq M + |b|$. □

Lemma 2.4.2. *There exists a constant $C \in \mathbb{R}_+$ such that for all $V_{\rm per} \in L^3_{\rm per}(\Gamma)$ and all $v \in H^1(\mathbb{R}^3)$, $V_{\rm per}v \in L^2(\mathbb{R}^3)$ and*

$$\|V_{\rm per}v\|_{L^2(\mathbb{R}^3)} \leq C\|V_{\rm per}\|_{L^3_{\rm per}(\Gamma)}\|v\|_{H^1(\mathbb{R}^3)}.$$

Proof. We proceed as [163, Theorem XIII-96]. Let $K = x_0 + 3\Gamma$ where $x_0 \in \mathbb{R}^3$ is such that Γ and K have the same center, and $\eta \in C_c^\infty(\mathbb{R}^3)$ supported in K and such that $\eta \equiv 1$ on Γ. It holds

$$\begin{aligned}
\int_{\mathbb{R}^3} |V_{\rm per}v|^2 &= \sum_{R\in\mathcal{R}} \|V_{\rm per}v\|^2_{L^2(R+\Gamma)} \leq \sum_{R\in\mathcal{R}} \|V_{\rm per}\|^2_{L^3_{\rm per}(\Gamma)}\|v\|^2_{L^6(R+\Gamma)} \\
&\leq \|V_{\rm per}\|^2_{L^3_{\rm per}(\Gamma)} \sum_{R\in\mathcal{R}} \|\eta(\cdot - R)v\|^2_{L^6(\mathbb{R}^3)} \\
&\leq C_0^2\|V_{\rm per}\|^2_{L^3_{\rm per}(\Gamma)} \sum_{R\in\mathcal{R}} \|\nabla(\eta(\cdot - R)v)\|^2_{L^2(\mathbb{R}^3)} \\
&= C_0^2\|V_{\rm per}\|^2_{L^3_{\rm per}(\Gamma)} \sum_{R\in\mathcal{R}} \|\nabla(\eta(\cdot - R)v)\|^2_{L^2(R+K)} \\
&\leq C_0^2\|\eta\|^2_{W^{1,\infty}(\mathbb{R}^3)}\|V_{\rm per}\|^2_{L^3_{\rm per}(\Gamma)} \sum_{R\in\mathcal{R}} \|v\|^2_{H^1(R+K)} \\
&= 27C_0^2\|\eta\|^2_{W^{1,\infty}(\mathbb{R}^3)}\|V_{\rm per}\|^2_{L^3_{\rm per}(\Gamma)}\|v\|_{H^1(\mathbb{R}^3)},
\end{aligned}$$

where C_0 is the Sobolev constant such that $\|\phi\|_{L^6(\mathbb{R}^3)} \leq C_0\|\nabla\phi\|_{L^2(\mathbb{R}^3)}$ for all $\phi \in H^1(\mathbb{R}^3)$ and where $\|\eta\|_{W^{1,\infty}(\mathbb{R}^3)} = \left(\|\eta\|^2_{L^\infty(\mathbb{R}^3)} + \|\nabla\eta\|^2_{L^\infty(\mathbb{R}^3)}\right)^{1/2}$. □

Lemma 2.4.3. *Let $\nu \in \mathcal{C}$ and $v \in \mathcal{Q}_+ \cap H^2(\mathbb{R}^3)$ such that $v > -u^0_{\rm per}$ in $\mathbb{R}^3$. For all $\epsilon > 0$ and $q \in \mathbb{R}$, there exists $\widetilde{v}_\epsilon \in \mathcal{Q}_+ \cap C^1_c(\mathbb{R}^3)$ such that*

$$\int_{\mathbb{R}^3} (2u^0_{\rm per}\widetilde{v}_\epsilon + \widetilde{v}_\epsilon^2) = q \quad \textit{and} \quad |\mathcal{E}^\nu(\widetilde{v}_\epsilon) - \mathcal{E}^\nu(v)| \leq \epsilon.$$

Proof. Let $\epsilon > 0$. As the functions of $H^2(\mathbb{R}^3)$ are continuous and decay to zero at infinity, there exists $\delta > 0$ such that

$$\forall x \in \mathbb{R}^3, \quad v(x) \geq -u^0_{\rm per}(x) + \delta. \tag{2.53}$$

For all $R > 0$, let B_R be the ball of $\mathbb{R}^3$ centered at zero and of radius R. For $\eta > 0$, we define

$$v^\eta = (u^0_{\rm per})^{-1}\mathcal{F}^{-1}\left(\chi_{\overline{B}_{1/\eta}\setminus B_\eta}\mathcal{F}(u^0_{\rm per}v)\right),$$

where $\mathcal{F}$ is the Fourier transform and $\mathcal{F}^{-1}$ the inverse Fourier transform. Clearly, $v^\eta \in H^3(\mathbb{R}^3) \hookrightarrow C^1(\mathbb{R}^3)$ and $u^0_{\rm per}v^\eta \in \mathcal{C}$. In addition, when η goes to zero, $(v^\eta)_{\eta>0}$ converges to v in $H^2(\mathbb{R}^3)$, hence in $L^\infty(\mathbb{R}^3)$, and $(u^0_{\rm per}v^\eta)_{\eta>0}$ converges to $u^0_{\rm per}v$ in $\mathcal{C}$. The function $\mathcal{E}^\nu$ being continuous on $\mathcal{Q}$, this implies that there exists some $\eta_0 > 0$ such that

$$v^{\eta_0} \in \mathcal{Q}_+ \cap C^1(\mathbb{R}^3) \quad \text{and} \quad |\mathcal{E}^\nu(v^{\eta_0}) - \mathcal{E}^\nu(v)| \leq \epsilon/4.$$

Let χ be a function of $C^\infty_c(\mathbb{R}^3)$ supported in B_2, such that $0 \leq \chi(\cdot) \leq 1$ and $\chi = 1$ in B_1. For $n \in \mathbb{N}^*$, we denote by $\chi_n(\cdot) = \chi(n^{-1}\cdot)$ and by $v^{\eta_0,n} = \chi_n v^{\eta_0}$. For each $n \in \mathbb{N}^*$, $v^{\eta_0,n} \in \mathcal{Q}_+ \cap C^1_c(\mathbb{R}^3)$ and the sequence $(v^{\eta_0,n})_{n\in\mathbb{N}^*}$ converges to v^{η_0} in $\mathcal{Q}$ when n goes to infinity. Hence, we can find some $n_0 > 0$ such that

$$v^{\eta_0,n_0} \in \mathcal{Q}_+ \cap C^1_c(\mathbb{R}^3) \quad \text{and} \quad |\mathcal{E}^\nu(v^{\eta_0,n_0}) - \mathcal{E}^\nu(v^{\eta_0})| \leq \epsilon/4.$$

Let

$$q_0 = \int_{\mathbb{R}^3}(2u^0_{\rm per}v^{\eta_0,n_0} + (v^{\eta_0,n_0})^2) \quad \text{and} \quad q_1 = q - q_0.$$

If $q_1 = 0$, $\widetilde{v}_\epsilon = v^{\eta_0,n_0}$ fulfills the conditions of Lemma 2.4.3. Otherwise, we introduce for m large enough the function w_m defined as $w_m = t_m\chi_m u^0_{\rm per}$ where t_m is the larger of the two real numbers such that

$$\int_{\mathbb{R}^3}(2u^0_{\rm per}w_m + w_m^2) = 2t_m\int_{\mathbb{R}^3}\chi_m\rho^0_{\rm per} + t_m^2\int_{\mathbb{R}^3}\chi_m^2\rho^0_{\rm per} = q_1.$$

A simple calculation shows that $t_m \underset{m\to\infty}{\sim} \frac{1}{2}q_1|\Gamma_1|Z^{-1}\left(\int_{\mathbb{R}^3}\chi\right)^{-1}m^{-3}$, and that

$$\lim_{m\to\infty}\mathcal{E}^0(w_m) = 0,$$

so that there exists $m_0 \in \mathbb{N}^*$ such that $w_{m_0} \in \mathcal{Q}_+ \cap C^1_c(\mathbb{R}^3)$ and $0 \leq \mathcal{E}^0(w_{m_0}) \leq \epsilon/4$. Let us finally choose some $R_1 \in \mathcal{R}_1 \setminus \{0\}$ and introduce the sequence of functions $(v^{\eta_0,n_0}_{m_0,p})_{p\in\mathbb{N}}$ defined by

$$v^{\eta_0,n_0}_{m_0,p}(\cdot) = v^{\eta_0,n_0}(\cdot) + w_{m_0}(\cdot - pR_1).$$

For p large enough, $v^{\eta_0,n_0}_{m_0,p}$ belongs to $\mathcal{Q}_+ \cap C^1_{\rm c}(\mathbb{R}^3)$ and satisfies

$$\int_{\mathbb{R}^3}(2u^0_{\rm per}v^{\eta_0,n_0}_{m_0,p} + (v^{\eta_0,n_0}_{m_0,p})^2) = q.$$

Besides,

$$\begin{aligned}
&|\mathcal{E}^\nu(v^{\eta_0,n_0}_{m_0,p}) - \mathcal{E}^\nu(v^{\eta_0,n_0})| \\
&\quad = \left|\mathcal{E}^0(v_{m_0}) + D(2u^0_{\rm per}v^{\eta_0,n_0} + (v^{\eta_0,n_0})^2 - \nu, (2u^0_{\rm per}w_{m_0} + w^2_{m_0})(\cdot - pR_1))\right| \\
&\quad \leq \epsilon/4 + \left|D(2u^0_{\rm per}v^{\eta_0,n_0} + (v^{\eta_0,n_0})^2 - \nu, (2u^0_{\rm per}w_{m_0} + w^2_{m_0})(\cdot - pR_1))\right|.
\end{aligned}$$

As

$$\lim_{p\to\infty} D(2u^0_{\rm per}v^{\eta_0,n_0} + (v^{\eta_0,n_0})^2 - \nu, (2u^0_{\rm per}w_{m_0} + w^2_{m_0})(\cdot - pR_1)) = 0,$$

there exists some $p_0 \in \mathbb{N}$ such that

$$\left|D(2u^0_{\rm per}v^{\eta_0,n_0} + (v^{\eta_0,n_0})^2 - \nu, (2u^0_{\rm per}w_{m_0} + w^2_{m_0})(\cdot - pR_1))\right| \leq \epsilon/4.$$

Setting $\widetilde{v}_\epsilon = v^{\eta_0,n_0}_{m_0,p_0}$, we get the desired result. □

The next four lemmas are useful to pass to the thermodynamic limit in the Coulomb term (Lemmas 2.4.4, 2.4.5 and 2.4.6) and in the kinetic energy term (Lemma 2.4.7).

Lemma 2.4.4. *There exists a constant $C \in \mathbb{R}_+$ such that for all $L \in \mathbb{N}^*$,*

$$\begin{aligned}\forall \rho_L \in L^1_{\rm per}(\Gamma_L) \cap L^{6/5}_{\rm per}(\Gamma_L), \qquad & D_{\mathcal{R}_L}(\rho_L,\rho_L) \leq C\left(\|\rho_L\|^2_{L^1_{\rm per}(\Gamma_L)} + \|\rho_L\|^2_{L^{6/5}_{\rm per}(\Gamma_L)}\right),\\ \forall v_L \in H^1_{\rm per}(\Gamma_L), \qquad & D_{\mathcal{R}_L}(v_L^2,v_L^2) \leq C\|v_L\|^4_{H^1_{\rm per}(\Gamma_L)}.\end{aligned}$$

Proof. It is well-known (see e.g. [50]) that

$$\forall x \in \Gamma_1, \quad G_{\mathcal{R}_1}(x) = |x|^{-1} + g(x),$$

with $g \in L^\infty(\Gamma_1)$, and that for all $L \in \mathbb{N}^*$,

$$\forall x \in \mathbb{R}^3, \quad G_{\mathcal{R}_L}(x) = L^{-1}G_{\mathcal{R}_1}(L^{-1}x).$$

Let $\mathcal{I} = \left\{R \in \mathcal{R}_1 \,|\, \exists (x,y) \in \overline{\Gamma}_1 \times \overline{\Gamma}_1 \text{ s.t. } x - y = R\right\}$. It holds

$$\forall (x,y) \in \Gamma_L \times \Gamma_L, \quad 0 \leq G_{\mathcal{R}_L}(x-y) \leq \sum_{R\in\mathcal{I}} |x-y-LR|^{-1} + L^{-1}\|g\|_{L^\infty}.$$

Therefore, for all $L \in \mathbb{N}^*$,

$$\begin{aligned} D_{\mathcal{R}_L}(\rho_L,\rho_L) &= \int_{\Gamma_L}\int_{\Gamma_L} G_{\mathcal{R}_L}(x-y)\rho_L(x)\rho_L(y)\,dx\,dy\\ &\leq \sum_{R\in\mathcal{I}}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\chi_{\Gamma_L}(x)|\rho_L(x)|\,\chi_{\Gamma_L}(y)|\rho_L(y)|}{|x-y-LR|}\,dx\,dy + L^{-1}\|g\|_{L^\infty}\|\rho_L\|^2_{L^1_{\rm per}(\Gamma_L)}\\ &\leq C'\|\chi_{\Gamma_L}\rho_L\|^2_{L^{6/5}(\mathbb{R}^3)} + \|g\|_{L^\infty}\|\rho_L\|^2_{L^1_{\rm per}(\Gamma_L)}\\ &= C'\|\rho_L\|^2_{L^{6/5}_{\rm per}(\Gamma_L)} + \|g\|_{L^\infty}\|\rho_L\|^2_{L^1_{\rm per}(\Gamma_L)}, \end{aligned}$$

where C' is a constant independent of L and ρ_L. Let C_1 be the Sobolev constant such that

$$\forall v_1 \in H^1_{\rm per}(\Gamma_1), \quad \|v_1\|_{L^6_{\rm per}(\Gamma_1)} \leq C_1\|v_1\|_{H^1_{\rm per}(\Gamma_1)}.$$

By an elementary scaling argument, it is easy to check that the inequality

$$\forall v_L \in H^1_{\rm per}(\Gamma_L), \quad \|v_L\|_{L^6_{\rm per}(\Gamma_L)} \leq C_1\|v_L\|_{H^1_{\rm per}(\Gamma_L)}$$

holds for all $L \in \mathbb{N}^*$. Thus, for all $v_L \in H^1_{\rm per}(\Gamma_L)$, we obtain

$$\|v_L^2\|^2_{L^{6/5}_{\rm per}(\Gamma_L)} = \|v_L\|^4_{L^{12/5}_{\rm per}(\Gamma_L)} \leq \|v_L\|^3_{L^2_{\rm per}(\Gamma_L)}\|v_L\|_{L^6_{\rm per}(\Gamma_L)} \leq C_1\|v_L\|^4_{H^1_{\rm per}(\Gamma_L)},$$

which completes the proof of Lemma 2.4.4. □

Lemma 2.4.5. *Let $\nu \in L^1(\mathbb{R}^3) \cap L^2(\mathbb{R}^3)$ and $\nu_L \in L^2_{\rm per}(\Gamma_L)$ defined by $\nu_L|_{\Gamma_L} = \nu|_{\Gamma_L}$ for all $L \in \mathbb{N}^*$. Then*

$$\lim_{L\to\infty} D_{\mathcal{R}_L}(\nu_L, \nu_L) = D(\nu, \nu). \tag{2.54}$$

Proof. Let $g_1 := |\Gamma_1|^{-1} \displaystyle\int_{\Gamma_1} G_1$ and Γ_L^* be the first Brillouin zone of the lattice $\mathcal{R}_L$. Note that $\mathcal{R}_L^* = L^{-1}\mathcal{R}_1^*$ and $\Gamma_L^* = L^{-1}\Gamma_1^*$. Let $K > 0$. We have

$$\begin{aligned} D_{\mathcal{R}_L}(\nu_L, \nu_L) &= g_1 L^{-1} \left(\int_{\Gamma_L} \nu\right)^2 + \sum_{k\in L^{-1}\mathcal{R}_1^*\setminus\{0\}} \frac{4\pi}{|k|^2} |c_{k,L}(\nu_L)|^2 \\ &= g_1 L^{-1} \left(\int_{\Gamma_L} \nu\right)^2 + 4\pi \sum_{k\in B_K\cap L^{-1}\mathcal{R}_1^*\setminus\{0\}} |\Gamma_L^*| \frac{|\widetilde{c}_{k,L}(\nu_L)|^2}{|k|^2} \\ &\quad + 4\pi \sum_{k\in B_K^c\cap L^{-1}\mathcal{R}_1^*\setminus\{0\}} \frac{|c_{k,L}(\nu_L)|^2}{|k|^2}, \end{aligned} \tag{2.55}$$

where B_K is the ball of radius K centered at 0, $B_K^c = \mathbb{R}^3 \setminus \overline{B}_K$,

$$c_{k,L}(\nu_L) = |\Gamma_L|^{-1/2} \int_{\Gamma_L} \nu_L(x) e^{-ik\cdot x}\, dx,$$

and

$$\widetilde{c}_{k,L}(\nu_L) = |\Gamma_L^*|^{-1/2} c_{k,L}(\nu_L) = \frac{1}{(2\pi)^{3/2}} \int_{\Gamma_L} \nu(x) e^{-ik\cdot x}\, dx.$$

As $\nu \in L^1(\mathbb{R}^3)$, $|\widetilde{c}_{k,L}(\nu_L)| \le (2\pi)^{-3/2} \|\nu\|_{L^1(\mathbb{R}^3)}$ for all k and L, $\widehat{\nu} \in L^\infty(\mathbb{R}^3)$, and

$$\forall k \in \mathbb{R}^3, \quad \widetilde{c}_{k,L}(\nu_L) \underset{L\to\infty}{\longrightarrow} \widehat{\nu}(k).$$

Clearly the first term in the right hand side of (2.55) goes to zero when L goes to infinity. Besides,

$$\sum_{k\in B_K\cap L^{-1}\mathcal{R}_1^*\setminus\{0\}} |\Gamma_L^*| \frac{|\widetilde{c}_{k,L}(\nu_L)|^2}{|k|^2} \underset{L\to\infty}{\longrightarrow} \int_{B_K} \frac{|\widehat{\nu}(k)|^2}{|k|^2}\, dk.$$

Lastly,

$$\begin{aligned} \sum_{k\in B_K^c\cap L^{-1}\mathcal{R}_1^*\setminus\{0\}} \frac{|c_{k,L}(\nu_L)|^2}{|k|^2} &\le \left(\sum_{k\in B_K^c\cap L^{-1}\mathcal{R}_1^*\setminus\{0\}} \frac{|c_{k,L}(\nu_L)|^2}{|k|^4}\right)^{1/2} \left(\sum_{k\in B_K^c\cap L^{-1}\mathcal{R}_1^*\setminus\{0\}} |c_{k,L}(\nu_L)|^2\right)^{1/2} \\ &\le \frac{1}{(2\pi)^{3/2}} \left(\sum_{k\in B_K^c\cap L^{-1}\mathcal{R}_1^*\setminus\{0\}} |\Gamma_L^*| \frac{1}{|k|^4}\right)^{1/2} \|\nu\|_{L^1(\mathbb{R}^3)} \|\nu\|_{L^2(\mathbb{R}^3)} \\ &\underset{L\to\infty}{\longrightarrow} \frac{1}{(2\pi^2 K)^{1/2}} \|\nu\|_{L^1(\mathbb{R}^3)} \|\nu\|_{L^2(\mathbb{R}^3)}. \end{aligned}$$

It is then easy to conclude that (2.54) holds true. □

Lemma 2.4.6. *Let $(\rho_L)_{L\in\mathbb{N}^*}$ be a sequence of functions of $L^2_{\rm loc}(\mathbb{R}^3)$ such that*

1. *for each* $L \in \mathbb{N}^*$, $\rho_L \in L^2_{\rm per}(\Gamma_L)$;

2. *there exists* $C \in \mathbb{R}_+$ *such that for all* $L \in \mathbb{N}^*$,

$$\left| \int_{\Gamma_L} \rho_L \right| \leq C \quad \textit{and} \quad D_{\mathcal{R}_L}(\rho_L, \rho_L) \leq C;$$

3. *there exists* $\rho \in \mathcal{D}'(\mathbb{R}^3)$ *such that* $(\rho_L)_{L\in\mathbb{N}^*}$ *converges to* ρ *in* $\mathcal{D}'(\mathbb{R}^3)$.

Then $\rho \in \mathcal{C}$ *and*

$$D(\rho,\rho) \leq \liminf_{L\to\infty} D_{\mathcal{R}_L}(\rho_L,\rho_L). \tag{2.56}$$

In addition, for any $p > 6/5$ *and any sequence* $(v_L)_{L\in\mathbb{N}^*}$ *of functions of* $L^p_{\rm loc}(\mathbb{R}^3)$ *such that* $v_L \in L^p_{\rm per}(\Gamma_L)$ *for all* $L \in \mathbb{N}^*$, *which weakly converges to some* $v \in L^p_{\rm loc}(\mathbb{R}^3)$ *in* $L^p_{\rm loc}(\mathbb{R}^3)$, *it holds*

$$\forall \phi \in C^\infty_c(\mathbb{R}^3), \quad \lim_{L\to\infty} D_{\mathcal{R}_L}(\rho_L, v_L\phi) = D(\rho, v\phi). \tag{2.57}$$

Proof. Let W_L the unique solution in $H^2_{\rm per}(\Gamma_L)$ to

$$\begin{cases} -\Delta W_L = 4\pi \left(\rho_L - |\Gamma_L|^{-1} \displaystyle\int_{\Gamma_L} \rho_L \right) \\ W_L \ \mathcal{R}_L\text{-periodic}, \quad \displaystyle\int_{\Gamma_L} W_L = 0. \end{cases} \tag{2.58}$$

It holds

$$\frac{1}{4\pi} \int_{\Gamma_L} |\nabla W_L|^2 = D_{\mathcal{R}_L}(\rho_L,\rho_L) - g_1 L^{-1} \left(\int_{\Gamma_L} \rho_L \right)^2 \leq C, \tag{2.59}$$

where $g_1 := |\Gamma_1|^{-1} \displaystyle\int_{\Gamma_1} G_{\mathcal{R}_1} \geq 0$. Hence the sequence $(\|\nabla W_L\|_{L^2_{\rm per}(\Gamma_L)})_{L\in\mathbb{N}^*}$ is bounded.

By Sobolev and PoincarÃ©-Wirtinger inequalities, we have

$$\forall V_1 \in H^1_{\rm per}(\Gamma_1) \text{ s.t. } \int_{\Gamma_1} V_1 = 0, \quad \|V_1\|_{L^6_{\rm per}(\Gamma_1)} \leq C_1 \|V_1\|_{H^1_{\rm per}(\Gamma_1)} \leq C_1' \|\nabla V_1\|_{L^2_{\rm per}(\Gamma_1)},$$

and by a scaling argument, we obtain that for all $L \in \mathbb{N}^*$,

$$\forall V_L \in H^1_{\rm per}(\Gamma_L) \text{ s.t. } \int_{\Gamma_L} V_L = 0, \quad \|V_L\|_{L^6_{\rm per}(\Gamma_L)} \leq C_1' \|\nabla V_L\|_{L^2_{\rm per}(\Gamma_L)},$$

where the constant C_1' does not depend on L. Thus, the sequence $(\|W_L\|_{L^6_{\rm per}(\Gamma_L)})_{L\in\mathbb{N}^*}$ is bounded. Let $\widetilde{C} \in \mathbb{R}_+$ such that

$$\forall L \in \mathbb{N}^*, \quad \|W_L\|_{L^6_{\rm per}(\Gamma_L)} \leq \widetilde{C} \quad \text{and} \quad \|\nabla W_L\|_{L^2_{\rm per}(\Gamma_L)} \leq \widetilde{C},$$

and let $(R_n)_{n\in\mathbb{N}}$ be an increasing sequence of positive real numbers such that $\lim_{n\to\infty} R_n = \infty$. Let $R > 0$. For $L > 2R$,

$$\|W_L\|_{L^6(B_R)} \leq \|W_L\|_{L^6_{\rm per}(\Gamma_L)} \leq \widetilde{C} \quad \text{and} \quad \|\nabla W_L\|_{L^2(B_R)} \leq \|\nabla W_L\|_{L^2_{\rm per}(\Gamma_L)} \leq \widetilde{C}.$$

We can therefore extract from $(W_L)_{L\in\mathbb{N}^*}$ a subsequence $(W_{L^0_n})_{n\in\mathbb{N}}$ such that $(W_{L^0_n}|_{B_{R_0}})_{n\in\mathbb{N}}$ converges weakly in $H^1(B_{R_0})$, strongly in $L^p(B_{R_0})$ for all $1 \leq p < 6$, and almost everywhere in B_{R_0} to some $W^0 \in H^1(B_{R_0})$, for which

$$\|W^0\|_{L^6(B_{R_0})} \leq \widetilde{C} \quad \text{and} \quad \|\nabla W^0\|_{L^2(B_{R_0})} \leq \widetilde{C}.$$

By recursion, we then extract from $(W_{L_n^k})_{n\in\mathbb{N}}$ a subsequence $(W_{L_n^{k+1}})_{n\in\mathbb{N}}$ such that $(W_{L_n^{k+1}}|_{B_{R_{k+1}}})_{n\in\mathbb{N}}$ converges weakly in $H^1(B_{R_{k+1}})$, strongly in $L^p(B_{R_{k+1}})$ for all $1 \le p < 6$, and almost everywhere in $B_{R_{k+1}}$ to some $W^{k+1} \in H^1(B_{R_{k+1}})$, for which

$$\|W^{k+1}\|_{L^6(B_{R_{k+1}})} \le \widetilde{C} \quad \text{and} \quad \|\nabla W^{k+1}\|_{L^2(B_{R_{k+1}})} \le \widetilde{C}. \tag{2.60}$$

Necessarily, $W^{k+1}|_{B_{R_k}} = W^k$. Let $L_n = L_n^n$ and let W be the function of $H^1_{\rm loc}(\mathbb{R}^3)$ defined by $W|_{B_{R_k}} = W^k$ for all $k \in \mathbb{N}$ (this definition is consistent since $W^{k+1}|_{B_{R_k}} = W^k$). The sequence $(W_{L_n})_{n\in\mathbb{N}}$ converges to W weakly in $H^1_{\rm loc}(\mathbb{R}^3)$, strongly in $L^p_{\rm loc}(\mathbb{R}^3)$ for all $1 \le p < 6$ and almost everywhere in $\mathbb{R}^3$. Besides, as (2.60) holds for all k, we also have

$$\|W\|_{L^6(\mathbb{R}^3)} \le \widetilde{C} \quad \text{and} \quad \|\nabla W\|_{L^2(\mathbb{R}^3)} \le \widetilde{C}.$$

Letting n go to infinity in (2.58) with $L = L_n$, we get

$$-\Delta W = 4\pi\rho.$$

We can reformulate the above results as $W \in \mathcal{C}'$ and $-\Delta W = 4\pi\rho$. As $-\Delta$ is an isomorphism from $\mathcal{C}'$ to $\mathcal{C}$, we necessarily have $\rho \in \mathcal{C}$. From (2.59), we infer that for each $R > 0$,

$$\frac{1}{4\pi}\|\nabla W\|_{L^2(B_R)} \le \liminf_{L\to\infty} D_{\mathcal{R}_L}(\rho_L, \rho_L).$$

Letting R go to infinity, we end up with (2.56). By uniqueness of the limit, the whole sequence $(W_L)_{L\in\mathbb{N}^*}$ converges to W weakly in $H^1_{\rm loc}(\mathbb{R}^3)$, and strongly in $L^p_{\rm loc}(\mathbb{R}^3)$ for all $1 \le p < 6$.

Let $p > 6/5$, $(v_L)_{L\in\mathbb{N}}$ be a sequence of functions on $L^p_{\rm loc}(\mathbb{R}^3)$ such that $v_L \in L^p_{\rm per}(\Gamma_L)$ for all $L \in \mathbb{N}^*$, and converging to some $v \in L^p_{\rm loc}(\mathbb{R}^3)$ weakly in $L^p_{\rm loc}(\mathbb{R}^3)$, and $\phi \in C^\infty_c(\mathbb{R}^3)$. We have, for L large enough,

$$\begin{aligned} D_{\mathcal{R}_L}(\rho_L, v_L\phi) &= \int_{\mathbb{R}^3} W_L v_L \phi - g_1 L^{-1} \left(\int_{\Gamma_L} \rho_L\right)\left(\int_{\Gamma_L} v_L\phi\right) \\ &= \int_{\mathrm{Supp}(\phi)} (W_L\phi) v_L - g_1 L^{-1} \left(\int_{\Gamma_L} \rho_L\right)\left(\int_{\mathrm{Supp}(\phi)} v_L\phi\right) \\ &\underset{L\to\infty}{\longrightarrow} \int_{\mathrm{Supp}(\phi)} W\phi v = D(\rho, v\phi), \end{aligned}$$

which proves (2.57). □

Let us introduce for each $L \in \mathbb{N}^*$ the bounded linear operator

$$\begin{aligned} i_L \,:\, L^2(\mathbb{R}^3) &\to L^2_{\rm per}(\Gamma_L) \\ v &\mapsto \sum_{R\in\mathcal{R}_L} (\chi_{\Gamma_L} v)(\cdot - R) \end{aligned} \tag{2.61}$$

and its adjoint $i_L^* \in \mathcal{L}(L^2_{\rm per}(\Gamma_L), L^2(\mathbb{R}^3))$. Note that for all $v_L \in L^2_{\rm per}(\Gamma_L)$, $i_L^* v_L = \chi_{\Gamma_L} v_L$ and $i_L i_L^* = 1_{L^2_{\rm per}(\Gamma_L)}$. As $C^\infty_c(\mathbb{R}^3) \subset H^1(\mathbb{R}^3)$, the domain of the self-adjoint operator $(H^0_{\rm per} - \epsilon^0_{\rm F})^{1/2}$, the function $(H^0_{\rm per} - \epsilon^0_{\rm F})^{1/2}\phi$ is in $L^2(\mathbb{R}^3)$. Using the same abuse of notation as above, we can also consider $H^0_{\rm per}$ as a self-adjoint operator on $L^2_{\rm per}(\Gamma_L)$ with domain $H^2_{\rm per}(\Gamma_L)$ and, for each $\phi \in C^\infty_c(\mathbb{R}^3)$, introduce the function $i_L^*(H^0_{\rm per} - \epsilon^0_{\rm F})^{1/2} i_L\phi$, which is well-defined in $L^2(\mathbb{R}^3)$.

Lemma 2.4.7. *Let $\phi \in C_c^\infty(\mathbb{R}^3)$. The sequence $(i_L^*(H^0_{\rm per} - \epsilon_{\rm F}^0)^{1/2} i_L\phi)_{L\in\mathbb{N}^*}$ converges to $(H^0_{\rm per} - \epsilon_{\rm F}^0)^{1/2}\phi$ in $L^2(\mathbb{R}^3)$.*

Proof. According to Bloch-Floquet theory [163], each $f \in L^2(\mathbb{R}^3)$ can be decomposed as

$$f(x) = \frac{1}{|\Gamma_1^*|}\int_{\Gamma_1^*} f_k(x)\, e^{ik\cdot x}\, dk$$

where f_k is the function of $L^2_{\rm per}(\Gamma_1)$ defined for almost all $k \in \mathbb{R}^3$ by

$$f_k(x) = \sum_{R\in\mathcal{R}_1} f(x+R)e^{-ik\cdot(x+R)}.$$

Recall that

$$\forall (f,g) \in L^2(\mathbb{R}^3)\times L^2(\mathbb{R}^3), \quad (f,g)_{L^2(\mathbb{R}^3)} = \frac{1}{|\Gamma_1^*|}\int_{\Gamma_1^*} (f_k, g_k)_{L^2_{\rm per}(\Gamma_1)}\, dk.$$

The operator $H^0_{\rm per}$, considered as a self-adjoint operator on $L^2(\mathbb{R}^3)$, commutes with the translations of the lattice $\mathcal{R}_1$ and can therefore be decomposed as

$$H^0_{\rm per} = \frac{1}{|\Gamma_1^*|}\int_{\Gamma_1^*} (H^0_{\rm per})_k\, dk$$

where $(H^0_{\rm per})_k$ is the self-adjoint operator on $L^2_{\rm per}(\Gamma_1)$ with domain $H^2_{\rm per}(\Gamma_1)$ defined by

$$(H^0_{\rm per})_k = -\Delta - 2ik\cdot\nabla + |k|^2 + \frac{5}{3}(\rho^0_{\rm per})^{2/3} + V^0_{\rm per}.$$

Let ϕ and ψ be two functions of $C_c^\infty(\mathbb{R}^3)$. Simple calculations show that for L large enough

$$(i_L^*(H^0_{\rm per} - \epsilon_{\rm F}^0)^{1/2} i_L\phi, \psi)_{L^2(\mathbb{R}^3)} = \sum_{k\in\Gamma_1^*\cap\mathcal{R}_L^*} L^{-3}((H^0_{\rm per} - \epsilon_{\rm F}^0)_k^{1/2}\phi_k, \psi_k)_{L^2_{\rm per}(\Gamma_1)}, \tag{2.62}$$

and

$$\|i_L^*(H^0_{\rm per} - \epsilon_{\rm F}^0)^{1/2} i_L\phi\|^2_{L^2(\mathbb{R}^3)} = \|(H^0_{\rm per} - \epsilon_{\rm F}^0)^{1/2}\phi\|^2_{L^2(\mathbb{R}^3)}. \tag{2.63}$$

The sequence $(i_L^*(H^0_{\rm per} - \epsilon_{\rm F}^0)^{1/2} i_L\phi)_{L\in\mathbb{N}^*}$ therefore is bounded in $L^2(\mathbb{R}^3)$, hence possesses a weakly converging subsequence.

Besides, the function $k \mapsto ((H^0_{\rm per} - \epsilon_{\rm F}^0)_k^{1/2}\phi_k, \psi_k)_{L^2_{\rm per}(\Gamma_1)}$ is continuous on $\overline{\Gamma_1^*}$ since

$$((H^0_{\rm per}-\epsilon_{\rm F}^0)_k^{1/2}\phi_k, \psi_k)_{L^2_{\rm per}(\Gamma_1)} = ((H^0_{\rm per}-\epsilon_{\rm F}^0+1)_k^{-1}(H^0_{\rm per}-\epsilon_{\rm F}^0)_k^{1/2}\phi_k, (H^0_{\rm per}-\epsilon_{\rm F}^0+1)_k\psi_k)_{L^2_{\rm per}(\Gamma_1)}$$

with $k \mapsto \phi_k$ and $k \mapsto (H^0_{\rm per} - \epsilon_{\rm F}^0 + 1)_k\psi_k$ continuous from $\overline{\Gamma_1^*}$ to $L^2_{\rm per}(\Gamma_1)$ and $k \mapsto (H^0_{\rm per} - \epsilon_{\rm F}^0 + 1)_k^{-1}(H^0_{\rm per} - \epsilon_{\rm F}^0)_k^{1/2}$ continuous from $\overline{\Gamma_1^*}$ to $\mathcal{L}(L^2_{\rm per}(\Gamma_1))$. Interpreting (2.62) as a Riemann sum, we obtain

$$\lim_{L\to\infty}(i_L^*(H^0_{\rm per} - \epsilon_{\rm F}^0)^{1/2} i_L\phi, \psi)_{L^2(\mathbb{R}^3)} \;=\; ((H^0_{\rm per} - \epsilon_{\rm F}^0)^{1/2}\phi, \psi)_{L^2(\mathbb{R}^3)}.$$

The above result allows to identify $(H^0_{\rm per} - \epsilon_{\rm F}^0)^{1/2}\phi$ as the weak limit of the sequence $(i_L^*(H^0_{\rm per} - \epsilon_{\rm F}^0)^{1/2} i_L\phi)_{L\in\mathbb{N}^*}$, and (2.63) shows that the convergence actually holds strongly in $L^2(\mathbb{R}^3)$. □

2.4.2 Proof of Proposition 2.2.1

Let $(v_n)_{n\in\mathbb{N}}$ be a minimizing sequence for (2.2). As each of the three terms of $E^{\mathrm{TFW}}_{\mathcal{R}}(\rho^{\mathrm{nuc}},\cdot)$ is non-negative, the sequence $(v_n)_{n\in\mathbb{N}}$ is clearly bounded in $H^1_{\mathrm{per}}(\Gamma)$, hence converges, up to extraction, to some $u\in H^1_{\mathrm{per}}(\Gamma)$, weakly in $H^1_{\mathrm{per}}(\Gamma)$, strongly in $L^p_{\mathrm{per}}(\Gamma)$ for each $1\le p<6$ and almost everywhere in $\mathbb{R}^3$. Passing to the liminf in the energy and to the limit in the constraint, we obtain that u satisfies $E^{\mathrm{TFW}}_{\mathcal{R}}(\rho^{\mathrm{nuc}},u)\le I_{\mathcal{R}}(\rho^{\mathrm{nuc}},Q)$ and $\int_\Gamma u^2=Q$. Therefore, u is a minimizer of (2.2). As $|u|\in H^1_{\mathrm{per}}(\Gamma)$, $E^{\mathrm{TFW}}_{\mathcal{R}}(\rho^{\mathrm{nuc}},|u|)=E^{\mathrm{TFW}}_{\mathcal{R}}(\rho^{\mathrm{nuc}},u)$ and $\int_\Gamma |u|^2=\int_\Gamma u^2$, $|u|$ also is a minimizer of (2.2). Should u be replaced with $|u|$, we can therefore assume that $u\ge 0$ in $\mathbb{R}^3$. Clearly, $-u$ also is a minimizer of (2.2).

Working on the Euler equation (2.3), we obtain by elementary elliptic regularity arguments [102] that $u\in H^3_{\mathrm{per}}(\Gamma)\hookrightarrow C^1(\mathbb{R}^3)\cap L^\infty(\mathbb{R}^3)$, and it follows from Harnack's inequality [102] that $u>0$ in $\mathbb{R}^3$.

Lastly, v_0 is a minimizer of (2.2) if and only if $\rho_0=v_0^2$ is a minimizer of

$$\inf\left\{\mathcal{E}^{\mathrm{TFW}}_{\mathcal{R}}(\rho^{\mathrm{nuc}},\rho),\ \rho\in\mathcal{K}_{\mathcal{R},Q}\right\}, \tag{2.64}$$

where

$$\mathcal{E}^{\mathrm{TFW}}_{\mathcal{R}}(\rho^{\mathrm{nuc}},\rho)=C_{\mathrm{W}}\int_\Gamma|\nabla\sqrt{\rho}|^2+C_{\mathrm{TF}}\int_\Gamma\rho^{5/3}+\frac{1}{2}\mathcal{D}_{\mathcal{R}}(\rho^{\mathrm{nuc}}-\rho,\rho^{\mathrm{nuc}}-\rho),$$

and

$$\mathcal{K}_{\mathcal{R},Q}=\left\{\rho\ge 0,\ \sqrt{\rho}\in H^1_{\mathrm{per}}(\Gamma),\ \int_\Gamma\rho=Q\right\}.$$

The functional $\rho\mapsto\mathcal{E}^{\mathrm{TFW}}_{\mathcal{R}}(\rho^{\mathrm{nuc}},\rho)$ being strictly convex on the convex set $\mathcal{K}$, (2.64) has a unique minimizer ρ_0 and it holds $\rho_0=u^2>0$. Any minimizer v_0 of (2.2) satisfying $v_0^2=\rho_0>0$, the only minimizers of (2.2) are u and $-u$.

2.4.3 Existence of a minimizer of (2.16)

The existence of a minimizer of (2.16) is an obvious consequence of the following lemma.

Lemma 2.4.8. *It holds*

$$\exists\beta>0\quad s.t.\quad\forall\nu\in\mathcal{C},\quad\forall v\in\mathcal{Q}_+,\quad\beta\|v\|^2_{H^1(\mathbb{R}^3)}\le\mathcal{E}^\nu(v), \tag{2.65}$$

$$\forall\nu\in\mathcal{C},\quad\forall v\in\mathcal{Q}_+,\quad\|u^0_{\mathrm{per}}v\|^2_{\mathcal{C}}\le\mathcal{E}^\nu(v)+\|v^2\|^2_{\mathcal{C}}+\|\nu\|^2_{\mathcal{C}}, \tag{2.66}$$

and for each $\nu\in\mathcal{C}$, the functional $\mathcal{E}^\nu$ is weakly lower semicontiuous in the closed convex subset $\mathcal{Q}_+$ of $\mathcal{Q}$.

Indeed, if $(v_n)_{n\in\mathbb{N}}$ is a minimizing sequence for (2.16), we infer from (2.65) and (2.66) that $(v_n)_{n\in\mathbb{N}}$ is bounded in $\mathcal{Q}$. We can therefore extract from $(v_n)_{n\in\mathbb{N}}$ a subsequence $(v_{n_k})_{k\in\mathbb{N}}$ weakly converging in $\mathcal{Q}$ to some $v_\nu\in\mathcal{Q}$. As $\mathcal{Q}_+$ is convex and strongly closed in $\mathcal{Q}$, it is weakly closed in $\mathcal{Q}$. Hence $v_\nu\in\mathcal{Q}_+$. Besides, $\mathcal{E}^\nu$ being weakly l.s.c. in $\mathcal{Q}_+$, we obtain

$$\mathcal{E}^\nu(v_\nu)\le\liminf_{k\to\infty}\mathcal{E}^\nu(v_{n_k})=I^\nu.$$

Therefore v_ν is a minimizer of (2.16).

Proof of Lemma 2.4.8. Using (2.10), (2.12), Lemma 2.4.1, Lemma 2.4.2, and the non-negativity of D, we obtain that for all $\nu \in \mathcal{C}$ and all $v \in \mathcal{Q}_+$,

$$\mathcal{E}^\nu(v) \geq \frac{2}{3} m^{4/3} \|v\|^2_{L^2(\mathbb{R}^3)},$$

and

$$\mathcal{E}^\nu(v) \geq \|v\|^2_{H^1(\mathbb{R}^3)} - \left(\frac{5}{3} M^{4/3} + |\epsilon_{\rm F}^0| + 1\right) \|v\|^2_{L^2(\mathbb{R}^3)} - \|V^0_{\rm per}\|_{L^3_{\rm per}(\mathbb{R}^3)} \|v\|_{L^2(\mathbb{R}^3)} \|v\|_{H^1(\mathbb{R}^3)}.$$

Therefore, there exists some constant $\beta > 0$ such that

$$\forall \nu \in \mathcal{C}, \quad \forall v \in \mathcal{Q}_+, \quad \mathcal{E}^\nu(v) \geq \beta \|v\|^2_{H^1(\mathbb{R}^3)}.$$

Besides, for all $\nu \in \mathcal{C}$ and all $v \in \mathcal{Q}_+$,

$$\begin{aligned} D(u^0_{\rm per} v, u^0_{\rm per} v) &\leq \frac{1}{2} D(2u^0_{\rm per} v + v^2 - \nu, 2u^0_{\rm per} v + v^2 - \nu) + \frac{1}{2} D(v^2 - \nu, v^2 - \nu) \\ &\leq \mathcal{E}^\nu(v) + D(v^2, v^2) + D(\nu, \nu). \end{aligned}$$

Hence (2.66).

Let $v \in \mathcal{Q}_+$ and $(v_n)_{n\in\mathbb{N}}$ be a sequence of elements of $\mathcal{Q}_+$ weakly converging to v in $\mathcal{Q}$. As $(v_n)_{n\in\mathbb{N}}$ is weakly converging, it is bounded in $\mathcal{Q}$, which means that $(v_n)_{n\in\mathbb{N}}$ and $(u^0_{\rm per} v_n)_{n\in\mathbb{N}}$ are bounded in $H^1(\mathbb{R}^3)$ and $\mathcal{C}$ respectively. We also notice that $(v_n^2)_{n\in\mathbb{N}}$ is bounded in $L^1(\mathbb{R}^3) \cap L^3(\mathbb{R}^3) \hookrightarrow L^{6/5}(\mathbb{R}^3) \hookrightarrow \mathcal{C}$.

Therefore, we can extract from $(v_n)_{n\in\mathbb{N}}$ a subsequence $(v_{n_k})_{k\in\mathbb{N}}$ such that

- $(\mathcal{E}^\nu(v_{n_k}))_{k\in\mathbb{N}}$ converges to $I = \liminf_{n\to\infty} \mathcal{E}^\nu(v_n)$ in $\mathbb{R}_+$;
- $(v_{n_k})_{k\in\mathbb{N}}$ converges to some $\widetilde{v} \in H^1(\mathbb{R}^3)$ weakly in $H^1(\mathbb{R}^3)$, strongly in $L^p_{\rm loc}(\mathbb{R}^3)$ for all $1 \leq p < 6$ and almost everywhere in $\mathbb{R}^3$;
- $(u^0_{\rm per} v_{n_k})_{k\in\mathbb{N}}$ weakly converges in $\mathcal{C}$ to some $w \in \mathcal{C}$;
- $(v^2_{n_k})_{k\in\mathbb{N}}$ weakly converges in $\mathcal{C}$ to some $z \in \mathcal{C}$.

We can rewrite the last two items above as

$$\forall V \in \mathcal{C}', \quad \int_{\mathbb{R}^3} u^0_{\rm per} v_{n_k} V \underset{k\to\infty}{\longrightarrow} \int_{\mathbb{R}^3} wV, \quad \text{and} \qquad \int_{\mathbb{R}^3} v^2_{n_k} V \underset{k\to\infty}{\longrightarrow} \int_{\mathbb{R}^3} zV.$$

Together with the strong convergence of $(v_{n_k})_{k\in\mathbb{N}}$ to $\widetilde{v}$ in $L^2_{\rm loc}(\mathbb{R}^3)$, this leads to $u^0_{\rm per}\widetilde{v} = w \in \mathcal{C}$ and $z = \widetilde{v}^2$. This in turn implies that $(v_{n_k})_{k\in\mathbb{N}}$ weakly converges in $\mathcal{Q}$ to $\widetilde{v}$. Therefore $\widetilde{v} = v$. Finally, $(v_{n_k})_{k\in\mathbb{N}}$ converges to v weakly in $H^1(\mathbb{R}^3)$ and almost everywhere in $\mathbb{R}^3$ and $(2u^0_{\rm per} v_{n_k} + v^2_{n_k} - \nu)_{k\in\mathbb{N}}$ weakly converges to $2u^0_{\rm per} v + v^2 - \nu$ in $\mathcal{C}$.

It follows from (2.12) that

$$\langle (H^0_{\rm per} - \epsilon^0_{\rm F}) v, v \rangle_{H^{-1}(\mathbb{R}^3), H^1(\mathbb{R}^3)} \leq \liminf_{k\to\infty} \langle (H^0_{\rm per} - \epsilon^0_{\rm F}) v_{n_k}, v_{n_k} \rangle_{H^{-1}(\mathbb{R}^3), H^1(\mathbb{R}^3)}.$$

By Fatou's Lemma,

$$\begin{aligned} &\int_{\mathbb{R}^3} \left(|u^0_{\rm per} + v|^{10/3} - |u^0_{\rm per}|^{10/3} - \frac{5}{3} |u^0_{\rm per}|^{4/3} (2u^0_{\rm per} v + v^2) \right) \\ &\qquad \leq \liminf_{k\to\infty} \int_{\mathbb{R}^3} \left(|u^0_{\rm per} + v_{n_k}|^{10/3} - |u^0_{\rm per}|^{10/3} - \frac{5}{3} |u^0_{\rm per}|^{4/3} (2u^0_{\rm per} v_{n_k} + v^2_{n_k}) \right). \end{aligned}$$

Lastly,

$$D(2u^0_{\rm per}v+v^2-\nu,2u^0_{\rm per}v+v^2-\nu)\leq\liminf_{k\to\infty}D(2u^0_{\rm per}v_{n_k}+v^2_{n_k}-\nu,2u^0_{\rm per}v_{n_k}+v^2_{n_k}-\nu).$$

Consequently,

$$\mathcal{E}^\nu(v)\leq\liminf_{k\to\infty}\mathcal{E}^\nu(v_{n_k})=\liminf_{n\to\infty}\mathcal{E}^\nu(v_n),$$

which proves that $\mathcal{E}^\nu$ is weakly l.s.c. in $\mathcal{Q}_+$. □

Clearly, the functional $\mathcal{E}^\nu$ is C^1 in $\mathcal{Q}$ and it holds

$$\begin{aligned}\forall h\in\mathcal{Q},\quad \langle\mathcal{E}^{\nu\prime}(v),h\rangle_{\mathcal{Q}',\mathcal{Q}} &= 2\Big(\langle(H^0_{\rm per}-\epsilon^0_{\rm F})v,h\rangle_{H^{-1}(\mathbb{R}^3),H^1(\mathbb{R}^3)}\\ &\quad+\frac{5}{3}\int_{\mathbb{R}^3}\Big(|u^0_{\rm per}+v|^{7/3}-|u^0_{\rm per}|^{7/3}-|u^0_{\rm per}|^{4/3}v\Big)h\\ &\quad+D(2u^0_{\rm per}v+v^2-\nu,(u^0_{\rm per}+v)h)\Big).\end{aligned}$$

The minimization set $\mathcal{Q}_+$ being convex, v_ν satisfies the Euler equation

$$\forall v\in\mathcal{Q}_+,\quad \langle\mathcal{E}^{\nu\prime}(v_\nu),(v-v_\nu)\rangle_{\mathcal{Q}',\mathcal{Q}}\geq 0. \tag{2.67}$$

Let $u_\nu=u^0_{\rm per}+v_\nu$ and

$$V=V^0_{\rm per}-\epsilon^0_{\rm F}+\frac{5}{3}|u_\nu|^{4/3}+(2u^0_{\rm per}v_\nu+v_\nu^2-\nu)\star|\cdot|^{-1}.$$

The function u_ν satisfies $u_\nu\in H^1_{\rm loc}(\mathbb{R}^3)$, $u_\nu\geq 0$ in $\mathbb{R}^3$, and

$$\begin{aligned}\forall\phi\in C^\infty_c(\mathbb{R}^3),\quad \int_{\mathbb{R}^3}\nabla u_\nu\cdot\nabla\phi+\int_{\mathbb{R}^3}Vu_\nu\phi &= \frac{1}{2}\langle\mathcal{E}^{\nu\prime}(v_\nu),\phi\rangle_{\mathcal{Q}',\mathcal{Q}}\\ &= \frac{1}{2}\langle\mathcal{E}^{\nu\prime}(v_\nu),(v_\nu+\phi-v_\nu)\rangle_{\mathcal{Q}',\mathcal{Q}}.\end{aligned}$$

This implies that for all $\phi\in C^\infty_c(\mathbb{R}^3)$ such that $\phi\geq 0$ in $\mathbb{R}^3$,

$$\int_{\mathbb{R}^3}\nabla u_\nu\cdot\nabla\phi+\int_{\mathbb{R}^3}Vu_\nu\phi\geq 0,$$

since $v_\nu+\phi\in\mathcal{Q}_+$. Therefore, u_ν is a non-negative supersolution of $-\Delta u+Vu=0$, with $V\in L^6_{\rm loc}(\mathbb{R}^3)$. It follows from Harnack's inequality (see Theorem 5.2 of [177]) that either u_ν is identically equal to zero in $\mathbb{R}^3$, or for each bounded domain Ω of $\mathbb{R}^3$, there exists $\eta>0$ such that $v_\nu\geq -u^0_{\rm per}+\eta$ in Ω. As the first case is excluded since $-u^0_{\rm per}\notin\mathcal{Q}_+$, (2.67) implies $\mathcal{E}^{\nu\prime}(v_\nu)=0$, which means that v_ν is a solution in $\mathcal{Q}_+$ to the elliptic equation (2.18).

Remarking that

$$\mathcal{E}^\nu(v_\nu)\leq\mathcal{E}^\nu(0)=\frac{1}{2}D(\nu,\nu)=\frac{1}{2}\|\nu\|^2_{\mathcal{C}},$$

and using (2.65) and (2.66), we finally get the estimates (2.19) and (2.20).

2.4.4 Uniqueness of the minimizer of (2.16)

Noticing that

$$\mathcal{Q}_+ = \left\{ v \in H^1(\mathbb{R}^3) \mid (u^0_{\rm per} + v)^2 - \rho^0_{\rm per} \in \mathcal{C},\ u^0_{\rm per} + v \geq 0 \right\},$$

we obtain that $v_\star$ is a minimizer of (2.16) if and only if $\rho_\star = (u^0_{\rm per} + v_\star)^2$ is a minimizer of

$$\inf \left\{ \mathcal{G}(\rho),\ \rho \in \mathcal{K} \right\} \tag{2.68}$$

where

$$\begin{aligned} \mathcal{G}(\rho) &= J(\rho) + \int_{\mathbb{R}^3} \left(\rho^{5/3} - (\rho^0_{\rm per})^{5/3} - \frac{5}{3}(\rho^0_{\rm per})^{2/3}(\rho - \rho^0_{\rm per}) \right) + \frac{1}{2} D(\rho - \rho^0_{\rm per} - \nu, \rho - \rho^0_{\rm per} - \nu), \\ J(\rho) &= \langle (H^0_{\rm per} - \epsilon^0_{\rm F})(\sqrt{\rho} - u^0_{\rm per}), (\sqrt{\rho} - u^0_{\rm per}) \rangle_{H^1(\mathbb{R}^3), H^{-1}(\mathbb{R}^3)}. \end{aligned}$$

and

$$\mathcal{K} = \left\{ \rho \geq 0 \mid \sqrt{\rho} - u^0_{\rm per} \in H^1(\mathbb{R}^3),\ \rho - \rho^0_{\rm per} \in \mathcal{C} \right\}.$$

To see that $\mathcal{K}$ is convex and that $\mathcal{G}$ is strictly convex on $\mathcal{K}$, we first introduce the set

$$\widetilde{\mathcal{K}} = \left\{ \rho \geq 0 \mid \sqrt{\rho} - u^0_{\rm per} \in H^1(\mathbb{R}^3) \cap \mathcal{E}'(\mathbb{R}^3) \right\}$$

where $\mathcal{E}'(\mathbb{R}^3)$ denotes the space of the compactly supported distributions, and observe that for all $\rho \in \widetilde{\mathcal{K}}$,

$$J(\rho) = \int_{\mathbb{R}^3} \left(|\nabla \sqrt{\rho}|^2 - |\nabla u^0_{\rm per}|^2 + \left(\frac{5}{3}(\rho^0_{\rm per})^{2/3} + V^0_{\rm per} - \epsilon^0_{\rm F} \right) (\rho - \rho^0_{\rm per}) \right).$$

Reasoning as in the proof of the convexity of the functional $\rho \mapsto \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}|^2$ on the convex set $\left\{ \rho \geq 0 \mid \sqrt{\rho} \in H^1(\mathbb{R}^3) \right\}$ (see e.g. [138]), we obtain that $\widetilde{\mathcal{K}}$ is convex and that $\mathcal{J}$ is convex on $\widetilde{\mathcal{K}}$. It then follows that $\mathcal{G}$ is strictly convex on $\widetilde{\mathcal{K}}$. We finally conclude by a density argument.

As $\mathcal{G}$ is strictly convex on the convex set $\mathcal{K}$, (2.68) has at most one minimizer. Therefore, $\rho_\nu = (u^0_{\rm per} + v_\nu)^2$ is the unique minimizer of (2.68), and v_ν is the unique minimizer of (2.16).

2.4.5 Properties of the unique minimizer of (2.16)

Throughout this section, C denotes constants independent of ν (but possibly dependent on $\rho^{\rm nuc}_{\rm per}$). The Euler equation (2.18) can be rewritten as

$$-\Delta v_\nu + V_\nu u^0_{\rm per} = f_\nu + (\nu \star |\cdot|^{-1}) u^0_{\rm per}, \tag{2.69}$$

where

$$f_\nu = (\epsilon^0_{\rm F} - V^0_{\rm per}) v_\nu - \frac{5}{3} \left(|u^0_{\rm per} + v_\nu|^{7/3} - |u^0_{\rm per}|^{7/3} \right) + \Phi^0_\nu v_\nu$$

(recall that $\Phi^0_\nu = (\nu - 2u^0_{\rm per} v_\nu - v_\nu^2) \star |\cdot|^{-1}$), and where $V_\nu = (2u^0_{\rm per} v_\nu + v_\nu^2) \star |\cdot|^{-1}$ satisfies

$$-\Delta V_\nu = 4\pi (2u^0_{\rm per} v_\nu + v_\nu^2). \tag{2.70}$$

In addition, we infer from (2.20) the following estimates

$$\begin{aligned}
\|v_\nu\|_{\mathcal{C}'} &\leq C\|v_\nu\|_{H^1(\mathbb{R}^3)} \leq C\|\nu\|_{\mathcal{C}},\\
\|v_\nu^2\|_{\mathcal{C}} &\leq C\|v_\nu^2\|_{L^{6/5}(\mathbb{R}^3)} = C\|v_\nu\|^2_{L^{12/5}(\mathbb{R}^3)} \leq C\|v_\nu\|^2_{H^1(\mathbb{R}^3)} \leq C\|\nu\|^2_{\mathcal{C}},\\
\|V_\nu\|_{\mathcal{C}'} &= \|2u^0_{\rm per}v_\nu + v_\nu^2\|_{\mathcal{C}} \leq C\left(\|v_\nu\|_{\mathcal{Q}} + \|v_\nu^2\|_{\mathcal{C}}\right) \leq C\left(\|\nu\|_{\mathcal{C}} + \|\nu\|^2_{\mathcal{C}}\right)\\
\|\Phi^0_\nu\|_{\mathcal{C}'} &\leq C\left(\|\nu\|_{\mathcal{C}} + \|V_\nu\|_{\mathcal{C}'}\right) \leq C\left(\|\nu\|_{\mathcal{C}} + \|\nu\|^2_{\mathcal{C}}\right),\\
\|f_\nu\|_{L^2(\mathbb{R}^3)} &\leq C\left(\|v_\nu\|_{H^1(\mathbb{R}^3)} + \|v_\nu\|_{L^2(\mathbb{R}^3)} + \|v_\nu\|^{7/3}_{L^{14/3}(\mathbb{R}^3)} + \|\Phi^0_\nu\|_{L^6(\mathbb{R}^3)}\|v_\nu\|_{L^3(\mathbb{R}^3)}\right)\\
&\leq C\left(\|\nu\|_{\mathcal{C}} + \|\nu\|^3_{\mathcal{C}}\right).
\end{aligned}$$

To obtain the bound on f_ν, we have used Lemma 2.4.1 and Lemma 2.4.2. Adding up (2.69) and (2.70), we obtain that $W_\nu = v_\nu + V_\nu$ is a solution in $\mathcal{C}'$ to

$$-\Delta W_\nu + u^0_{\rm per}W_\nu = \widetilde{f}_\nu + (\nu \star |\cdot|^{-1})u^0_{\rm per}, \tag{2.71}$$

where $\widetilde{f}_\nu = f_\nu + (8\pi+1)u^0_{\rm per}v_\nu + 4\pi v_\nu^2 \in L^2(\mathbb{R}^3)$, with

$$\|\widetilde{f}_\nu\|_{L^2(\mathbb{R}^3)} \leq C\left(\|f_\nu\|_{L^2(\mathbb{R}^3)} + (8\pi+1)\|v_\nu\|_{\mathcal{Q}} + 4\pi\|v_\nu\|^2_{L^2(\mathbb{R}^3)}\right) \leq C\left(\|\nu\|_{\mathcal{C}} + \|\nu\|^3_{\mathcal{C}}\right)$$

In addition, it follows from (2.20) and (2.70) that

$$\|W_\nu\|_{\mathcal{C}'} \leq \|v_\nu\|_{\mathcal{C}'} + \|V_\nu\|_{\mathcal{C}'} \leq C\left(\|\nu\|_{\mathcal{C}} + \|\nu\|^2_{\mathcal{C}}\right).$$

Since $u^0_{\rm per}$ satisfies (2.10), the elliptic equation

$$-\Delta w + u^0_{\rm per}w = \widetilde{f}_\nu$$

has a unique variational solution in $H^1(\mathbb{R}^3)$, which we denote by w_ν. It holds

$$\|w_\nu\|_{\mathcal{C}'} \leq C\|w_\nu\|_{H^1(\mathbb{R}^3)} \leq C\|\widetilde{f}_\nu\|_{L^2(\mathbb{R}^3)} \leq C\left(\|\nu\|_{\mathcal{C}} + \|\nu\|^3_{\mathcal{C}}\right).$$

The function $\widetilde{w}_\nu = W_\nu - w_\nu \in \mathcal{C}'$ then is solution to

$$-\Delta\widetilde{w}_\nu + u^0_{\rm per}\widetilde{w}_\nu = (\nu \star |\cdot|^{-1})u^0_{\rm per}, \tag{2.72}$$

and such that

$$\|\widetilde{w}_\nu\|_{\mathcal{C}'} \leq \|W_\nu\|_{\mathcal{C}'} + \|w_\nu\|_{\mathcal{C}'} \leq C\left(\|\nu\|_{\mathcal{C}} + \|\nu\|^3_{\mathcal{C}}\right).$$

Introducing $\widetilde{\rho}_\nu = -(4\pi)^{-1}\Delta\widetilde{w}_\nu \in \mathcal{C}$, (2.72) also reads

$$4\pi\frac{\widetilde{\rho}_\nu}{u^0_{\rm per}} = (\nu - \widetilde{\rho}_\nu) \star |\cdot|^{-1}.$$

Therefore,

$$4\pi\int_{\mathbb{R}^3}\frac{\widetilde{\rho}_\nu^2}{u^0_{\rm per}} = D(\nu - \widetilde{\rho}_\nu, \widetilde{\rho}_\nu) < \infty,$$

which proves that $\widetilde{\rho}_\nu \in L^2(\mathbb{R}^3)$, hence that $(\nu - \widetilde{\rho}_\nu) \star |\cdot|^{-1} \in L^2(\mathbb{R}^3)$. We also get the estimates

$$\begin{aligned}
\|(\nu - \widetilde{\rho}_\nu) \star |\cdot|^{-1}\|^2_{L^2(\mathbb{R}^3)} &\leq C\|\widetilde{\rho}_\nu\|^2_{L^2(\mathbb{R}^3)} \leq C\, D(\nu - \widetilde{\rho}_\nu, \widetilde{\rho}_\nu)\\
&\leq C\left(\|\nu\|_{\mathcal{C}}\|\widetilde{\rho}_\nu\|_{\mathcal{C}} + \|\widetilde{\rho}_\nu\|^2_{\mathcal{C}}\right) \leq C\left(\|\nu\|_{\mathcal{C}}\|\widetilde{w}_\nu\|_{\mathcal{C}'} + \|\widetilde{w}_\nu\|^2_{\mathcal{C}'}\right).
\end{aligned}$$

Thus,

$$\|(\nu-\widetilde{\rho}_\nu)\star|\cdot|^{-1}\|_{L^2(\mathbb{R}^3)}\leq C\left(\|\nu\|_{\mathcal{C}}+\|\nu\|_{\mathcal{C}}^3\right).$$

As

$$\Phi_\nu^0=\nu\star|\cdot|^{-1}-V_\nu=(\nu-\widetilde{\rho}_\nu)\star|\cdot|^{-1}+\widetilde{w}_\nu-V_\nu=(\nu-\widetilde{\rho}_\nu)\star|\cdot|^{-1}+v_\nu-w_\nu,$$

we obtain $\Phi_\nu^0\in L^2(\mathbb{R}^3)$ and

$$\|\Phi_\nu^0\|_{L^2(\mathbb{R}^3)}\leq C\left(\|\nu\|_{\mathcal{C}}+\|\nu\|_{\mathcal{C}}^3\right).$$

Hence (2.25). Introducing $\rho_\nu^0=\nu-(2u_{\rm per}^0v_\nu+v_\nu^2)$, the above statement reads

$$\left(\int_{\mathbb{R}^3}\frac{|\widehat{\rho_\nu^0}(k)|^2}{|k|^4}\,dk\right)^{1/2}\leq C\left(\|\nu\|_{\mathcal{C}}+\|\nu\|_{\mathcal{C}}^3\right)<\infty.$$

Therefore,

$$\begin{aligned}\frac{1}{|B_r|}\int_{B_r}|\widehat{\rho_\nu^0}(k)|\,dk &\leq \frac{1}{|B_r|}\left(\int_{B_r}|k|^4\,dk\right)^{1/2}\left(\int_{B_r}\frac{|\widehat{\rho_\nu^0}(k)|^2}{|k|^4}\,dk\right)^{1/2}\\ &= 3\left(\frac{r}{28\pi}\right)^{1/2}\left(\int_{\mathbb{R}^3}\frac{|\widehat{\rho_\nu^0}(k)|^2}{|k|^4}\,dk\right)^{1/2}\underset{r\to0}{\longrightarrow}0.\end{aligned}$$

Rewritting (2.69) as

$$-\Delta v_\nu=f_\nu+\Phi_\nu^0u_{\rm per}^0,$$

we conclude that $v_\nu\in H^2(\mathbb{R}^3)$ and that

$$\begin{aligned}\|v_\nu\|_{H^2(\mathbb{R}^3)} &\leq C\left(\|v_\nu\|_{H^1(\mathbb{R}^3)}+\|\Delta v_\nu\|_{L^2(\mathbb{R}^3)}\right)\\ &\leq C\left(\|v_\nu\|_{H^1(\mathbb{R}^3)}+\|f_\nu\|_{L^2(\mathbb{R}^3)}+\|\Phi_\nu^0\|_{L^2(\mathbb{R}^3)}\right)\leq C\left(\|\nu\|_{\mathcal{C}}+\|\nu\|_{\mathcal{C}}^3\right).\end{aligned}$$

Hence (2.21). Lastly, for each $1\leq j\leq 3$, the function $v_\nu^j:=\frac{\partial v_\nu}{\partial x_j}$ is a solution in $H^1(\mathbb{R}^3)$ to the elliptic equation

$$\begin{aligned}-\Delta v_\nu^j &= -\frac{\partial V_{\rm per}^0}{\partial x_j}v_\nu+(\epsilon_{\rm F}^0-V_{\rm per}^0)v_\nu^j+\frac{35}{9}\left(|u_{\rm per}^0+v_\nu|^{4/3}-|u_{\rm per}^0|^{4/3}\right)\frac{\partial u_{\rm per}^0}{\partial x_j}\\ &\quad -\frac{35}{9}|u_{\rm per}^0+v_\nu|^{4/3}v_\nu^j+\frac{\partial\Phi_\nu^0}{\partial x_j}u_{\rm per}^0+\Phi_\nu^0\frac{\partial u_{\rm per}^0}{\partial x_j}+\frac{\partial\Phi_\nu^0}{\partial x_j}v_\nu+\Phi_\nu^0v_\nu^j.\end{aligned}$$

Using (2.21) and (2.25), we obtain (2.22).

2.4.6 End of the proof of Theorem 2.3.1

We have proven in the previous two sections that:

1. (2.16) has a unique minimizer v_ν;
2. if $(v_n)_{n\in\mathbb{N}}$ is a minimizing sequence for (2.16), we can extract from $(v_n)_{n\in\mathbb{N}}$ a subsequence $(v_{n_k})_{k\in\mathbb{N}}$ which converges to v_ν, weakly in $H^1(\mathbb{R}^3)$, and strongly in $L^p_{\rm loc}(\mathbb{R}^3)$ for all $1\leq p<6$, and such that $(u_{\rm per}^0v_{n_k})_{k\in\mathbb{N}}$ converges to $u_{\rm per}^0v_\nu$ weakly in $\mathcal{C}$.

By uniqueness of the limit, this implies that any minimizing sequence $(v_n)_{n\in\mathbb{N}}$ for (2.16) converges to v_ν, weakly in $H^1(\mathbb{R}^3)$, and strongly in $L^p_{\rm loc}(\mathbb{R}^3)$ for all $1\leq p<6$, and that $(u_{\rm per}^0v_n)_{n\in\mathbb{N}}$ converges weakly to $u_{\rm per}^0v_\nu$ in $\mathcal{C}$. Lastly, the existence of a minimizing sequence for (2.16) satisfying (2.26) is a straightforward consequence of Lemma 2.4.3.

2.4.7 Thermodynamic limit with a charge constraint

Let $\nu \in L^1(\mathbb{R}^3) \cap L^2(\mathbb{R}^3)$. Clearly, $v_{\nu,q,L}$ is a minimizer of (2.32) if and only if $u^0_{\rm per} + v_{\nu,q,L}$ is a minimizer of (2.2) with $\mathcal{R} = \mathcal{R}_L$, $\rho^{\rm nuc} = \rho^{\rm nuc}_{\rm per} + \nu_L$ and $Q = ZL^3 + q$ such that $u^0_{\rm per} + v_{\nu,q,L} \geq 0$ in $\mathbb{R}^3$. It follows from Proposition 2.2.1 that (2.32) has a unique minimizer $v_{\nu,q,L}$, which satisfies $v_{\nu,q,L} \in H^3_{\rm per}(\Gamma_L) \hookrightarrow C^1(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3)$ and $u^0_{\rm per} + v_{\nu,q,L} > 0$ in $\mathbb{R}^3$, and the Euler equation (2.34) for some $\mu_{\nu,q,L} \in \mathbb{R}$.

Let $\alpha = |\Gamma_1|^{-1} \int_{\Gamma_1} u^0_{\rm per}$. For L large enough, $\alpha^2 + q/|\Gamma_L| \geq 0$ and the constant function $z_L = -\alpha + \sqrt{\alpha^2 + q/|\Gamma_L|}$ satisfies $z_L \geq -u^0_{\rm per}$ everywhere in $\mathbb{R}^3$ and

$$\int_{\Gamma_L} (2u^0_{\rm per} z_L + z_L^2) = q.$$

Using Lemma 2.4.1, Lemma 2.4.5, and the fact that $|z_L| \leq CL^{-3}$ for some constant C independent of L, we obtain

$$\begin{aligned} \mathcal{E}^\nu_L(v_{\nu,q,L}) &\leq \mathcal{E}^\nu_L(z_L) \\ &= \int_{\Gamma_L} \left(|u^0_{\rm per} + z_L|^{10/3} - |u^0_{\rm per}|^{10/3} - \frac{10}{3} |u^0_{\rm per}|^{7/3} z_L \right) + \int_{\Gamma_L} (V^0_{\rm per} - \epsilon^0_{\rm F}) z_L^2 \\ &\quad + \frac{1}{2} D_{\mathcal{R}_L} \left(2u^0_{\rm per} z_L + z_L^2 - \nu_L, 2u^0_{\rm per} z_L + z_L^2 - \nu_L \right) \underset{L\to\infty}{\longrightarrow} D(\nu,\nu). \end{aligned} \tag{2.73}$$

Besides, reasoning as in Section 2.4.3, we obtain

$$\forall v_L \in \mathcal{Q}_{+,L}, \quad \mathcal{E}^\nu_L(v_L) \geq \beta \|v_L\|^2_{H^1_{\rm per}(\Gamma_L)}, \tag{2.74}$$

where the constant $\beta > 0$ is the same as in (2.65), and

$$\begin{aligned} \forall v_L \in \mathcal{Q}_{+,L}, \quad D_{\mathcal{R}_L}(u^0_{\rm per} v_L, u^0_{\rm per} v_L) &\leq \mathcal{E}^\nu_L(v_L) + \frac{1}{2} D_{\mathcal{R}_L}(v_L^2 - \nu_L, v_L^2 - \nu_L) \\ &\leq \mathcal{E}^\nu_L(v_L) + D_{\mathcal{R}_L}(v_L^2, v_L^2) + D_{\mathcal{R}_L}(\nu_L, \nu_L) \end{aligned} \tag{2.75}$$

We infer from (2.73) and (2.74) that for each $q \in \mathbb{R}$, there exists $C_q \in \mathbb{R}_+$ such that

$$\forall L \in \mathbb{N}^*, \quad \|v_{\nu,q,L}\|_{H^1_{\rm per}(\Gamma_L)} \leq C_q. \tag{2.76}$$

By a diagonal extraction process similar to the one used in the proof of Lemma 2.4.6, we can extract from $(v_{\nu,q,L})_{L\in\mathbb{N}^*}$ a subsequence $(v_{\nu,q,L_k})_{k\in\mathbb{N}}$ which converges to some $\widetilde{v}_\nu \in H^1(\mathbb{R}^3)$, weakly in $H^1_{\rm loc}(\mathbb{R}^3)$, strongly in $L^p_{\rm loc}(\mathbb{R}^3)$ for all $1 \leq p < 6$ and almost everywhere in $\mathbb{R}^3$ and such that

$$\lim_{k\to\infty} \mathcal{E}^\nu_{L_k}(v_{\nu,q,L_k}) = \liminf_{L\to\infty} \mathcal{E}^\nu_L(v_{\nu,q,L}).$$

In particular $\widetilde{v}_\nu \geq -u^0_{\rm per}$ almost everywhere in $\mathbb{R}^3$.

Let us now prove that $u^0_{\rm per}\widetilde{v}_\nu \in \mathcal{C}$. First, we notice that it follows from (2.73), (2.75) and Lemma 2.4.4 that there exists a constant $\widetilde{C}_q$ such that

$$D_{\mathcal{R}_L}(u^0_{\rm per} v_{\nu,q,L}, u^0_{\rm per} v_{\nu,q,L}) \leq \widetilde{C}_q. \tag{2.77}$$

Besides,

$$\left| \int_{\Gamma_L} u^0_{\rm per} v_{\nu,q,L} \right| = \left| \frac{1}{2} \left(q - \int_{\Gamma_L} v^2_{\nu,q,L} \right) \right| \leq \frac{1}{2} \left(|q| + C_q^2 \right),$$

and $(u^0_{\rm per} v_{\nu,q,L_k})_{k\in\mathbb{N}}$ converges to $u^0_{\rm per}\widetilde{v}_\nu$ strongly in $L^2_{\rm loc}(\mathbb{R}^3)$, hence in the distributional sense. It therefore follows from Lemma 2.4.6 that $u^0_{\rm per}\widetilde{v}_\nu \in \mathcal{C}$. Thus, $\widetilde{v}_\nu \in \mathcal{Q}_+$.

As (2.34) holds in $H^{-1}_{\rm per}(\Gamma_L)$, we can take $u^0_{\rm per}$ as a test function. Using (2.11), we obtain

$$\begin{aligned}\mu_{\nu,q,L}\left(ZL^3+\int_{\Gamma_L} v_{\nu,q,L}u^0_{\rm per}\right) &= \int_{\Gamma_L}\frac{5}{3}\left(|u^0_{\rm per}+v_{\nu,q,L}|^{7/3}-|u^0_{\rm per}|^{7/3}-|u^0_{\rm per}|^{4/3}v_{\nu,q,L}\right)u^0_{\rm per}\\ &\quad +D_{\mathcal{R}_L}\left((2u^0_{\rm per}v_{\nu,q,L}+v^2_{\nu,q,L}-\nu_L),(u^0_{\rm per}+v_{\nu,q,L})u^0_{\rm per}\right).\end{aligned}$$

Using (2.76), (2.77) and Lemma 2.4.1, we obtain

$$\begin{aligned}&\left|\int_{\Gamma_L}\frac{5}{3}\left(|u^0_{\rm per}+v_{\nu,q,L}|^{7/3}-|u^0_{\rm per}|^{7/3}-|u^0_{\rm per}|^{4/3}v_{\nu,q,L}\right)u^0_{\rm per}\right|\le C'_qL^{3/2},\\ &\left|D_{\mathcal{R}_L}\left((2u^0_{\rm per}v_{\nu,q,L}+v^2_{\nu,q,L}-\nu_L),(u^0_{\rm per}+v_{\nu,q,L})u^0_{\rm per}\right)\right|\le C'_qL^{5/2},\\ &\left|\int_{\Gamma_L}v_{\nu,q,L}u^0_{\rm per}\right|\le\frac{1}{2}\left(|q|+C_q^2\right),\end{aligned}$$

for some constant C'_q independent of L, which allows us to conclude that $(\mu_{\nu,q,L})_{L\in\mathbb{N}^*}$ goes to zero when L goes to infinity.

Note that using Lemma 2.4.6, we can pass to the limit in the Euler equation (2.34) in the distributional sense, and prove that $\widetilde{v}_\nu$ satisfies

$$\begin{aligned}&(H^0_{\rm per}-\epsilon^0_F)\widetilde{v}_\nu+\frac{5}{3}\left(|u^0_{\rm per}+\widetilde{v}_\nu|^{7/3}-|u^0_{\rm per}|^{7/3}-|u^0_{\rm per}|^{4/3}\widetilde{v}_\nu\right)\\ &\qquad+\left((2u^0_{\rm per}\widetilde{v}_\nu+\widetilde{v}^2_\nu-\nu)\star|\cdot|^{-1}\right)(u^0_{\rm per}+\widetilde{v}_\nu)=0.\end{aligned}\tag{2.78}$$

We are now going to prove that $\mathcal{E}^\nu(\widetilde{v}_\nu)\le\mathcal{E}^\nu(v_\nu)$, which implies that $\widetilde{v}_\nu=v_\nu$ and, by uniqueness of the limit, that the whole sequence $(v_{\nu,q,L})_{L\in\mathbb{N}^*}$ converges to v_ν weakly in $H^1_{\rm loc}(\mathbb{R}^3)$, and strongly in $L^p_{\rm loc}(\mathbb{R}^3)$ for all $1\le p<6$.

Let $\epsilon>0$. From Lemma 2.4.3, there exists $v^\epsilon_{\nu,q}\in\mathcal{Q}_+\cap\mathcal{C}^2_c(\mathbb{R}^3)$ such that

$$\int_{\Gamma_L}(2u^0_{\rm per}v^\epsilon_{\nu,q}+(v^\epsilon_{\nu,q})^2)=q$$

and

$$\mathcal{E}^\nu(v_\nu)\le\mathcal{E}^\nu(v^\epsilon_{\nu,q})\le\mathcal{E}^\nu(v_{\nu,q})+\epsilon.$$

For L large enough, the $\mathcal{R}_L$-periodic function $v^\epsilon_{\nu,q,L}$ defined by $v^\epsilon_{\nu,q,L}|_{\Gamma_L}=v^\epsilon_{\nu,q}|_{\Gamma_L}$ is in the minimization set of (2.32). Using Lemma 2.4.5 and the fact that $v^\epsilon_{\nu,q}$ is compactly supported, we have for L large enough $v^\epsilon_{\nu,q,L}\in\mathcal{Q}_{+,L}$ and

$$\begin{aligned}\mathcal{E}^\nu_L(v_{\nu,q,L})\le\mathcal{E}^\nu_L(v^\epsilon_{\nu,q,L}) &= \langle(H^0_{\rm per}-\epsilon^0_{\rm F})v^\epsilon_{\nu,q},v^\epsilon_{\nu,q}\rangle_{H^{-1}(\mathbb{R}^3),H^1(\mathbb{R}^3)}\\ &\quad+\int_{\mathbb{R}^3}\left(|u^0_{\rm per}+v^\epsilon_{\nu,q}|^{10/3}-|u^0_{\rm per}|^{10/3}-\frac{5}{3}|u^0_{\rm per}|^{4/3}(2u^0_{\rm per}v^\epsilon_{\nu,q}+(v^\epsilon_{\nu,q})^2)\right)\\ &\quad+\frac{1}{2}D_{\mathcal{R}_L}\left(2u^0_{\rm per}v^\epsilon_{\nu,q,L}+(v^\epsilon_{\nu,q,L})^2-\nu_L,2u^0_{\rm per}v^\epsilon_{\nu,q,L}+(v^\epsilon_{\nu,q,L})^2-\nu_L\right)\\ &\quad\underset{L\to\infty}{\longrightarrow}\mathcal{E}^\nu(v^\epsilon_{\nu,q}).\end{aligned}$$

Therefore, for each $\epsilon>0$,

$$\mathcal{E}^\nu_L(v_{\nu,q,L})\le\mathcal{E}^\nu(v_\nu)+2\epsilon,$$

for L large enough, so that

$$\limsup_{L\to\infty}\mathcal{E}^\nu_L(v_{\nu,q,L})\le\mathcal{E}^\nu(v_\nu).\tag{2.79}$$

We are now going to prove that

$$\mathcal{E}^\nu(\widetilde{v}_\nu) \leq \liminf_{L\to\infty} \mathcal{E}^\nu_L(v_{\nu,q,L}). \tag{2.80}$$

For each $k \in \mathbb{N}$, we denote by

$$\widetilde{v}_k := i^*_{L_k} v_{\nu,q,L_k} \quad \text{and} \quad w_k := i^*_{L_k}(H^0_{\rm per} - \epsilon^0_{\rm F})^{1/2} v_{\nu,q,L_k},$$

where the operator i_{L_k} is defined by (2.61). As $\|\widetilde{v}_k\|_{L^2(\mathbb{R}^3)} = \|v_{\nu,q,L_k}\|_{L^2_{\rm per}(\Gamma_{L_k})}$ and

$$\|w_k\|^2_{L^2(\mathbb{R}^3)} = \langle (H^0_{\rm per} - \epsilon^0_{\rm F}) v_{\nu,q,L_k}, v_{\nu,q,L_k} \rangle_{H^{-1}_{\rm per}(\Gamma_{L_k}), H^{-1}_{\rm per}(\Gamma_{L_k})},$$

we can extract from $(\widetilde{v}_k)_{k\in\mathbb{N}}$ and $(w_k)_{k\in\mathbb{N}}$ subsequences $(\widetilde{v}_{k_n})_{n\in\mathbb{N}}$ and $(w_{k_n})_{n\in\mathbb{N}}$ which weakly converge in $L^2(\mathbb{R}^3)$ to some $\widetilde{v} \in L^2(\mathbb{R}^3)$ and $w \in L^2(\mathbb{R}^3)$ respectively, and such that

$$\lim_{n\to\infty} \mathcal{E}^\nu(v_{\nu,q,L_{k_n}}) = \liminf_{L\to\infty} \mathcal{E}^\nu(v_{\nu,q,L}).$$

As $(v_{\nu,q,L_k})_{k\in\mathbb{N}}$ converges to $\widetilde{v}_\nu$ strongly in $L^2_{\rm loc}(\mathbb{R}^3)$, we have $\widetilde{v} = \widetilde{v}_\nu$. Let us now prove that $w = (H^0_{\rm per} - \epsilon^0_{\rm F})^{1/2}\widetilde{v}_\nu$. For each $\phi \in C^\infty_c(\mathbb{R}^3)$, we infer from Lemma 2.4.7 that

$$\begin{aligned}
(w,\phi)_{L^2(\mathbb{R}^3)} &= \lim_{n\to\infty} (i^*_{L_{k_n}}(H^0_{\rm per} - \epsilon^0_{\rm F})^{1/2} v_{\nu,q,L_{k_n}}, \phi)_{L^2(\mathbb{R}^3)} \\
&= \lim_{n\to\infty} (i^*_{L_{k_n}}(H^0_{\rm per} - \epsilon^0_{\rm F})^{1/2} i_{L_{k_n}} \widetilde{v}_{k_n}, \phi)_{L^2(\mathbb{R}^3)} \\
&= \lim_{n\to\infty} (\widetilde{v}_{k_n}, i^*_{L_{k_n}}(H^0_{\rm per} - \epsilon^0_{\rm F})^{1/2} i_{L_{k_n}} \phi)_{L^2(\mathbb{R}^3)} \\
&= (\widetilde{v}_\nu, (H^0_{\rm per} - \epsilon^0_{\rm F})^{1/2}\phi)_{L^2(\mathbb{R}^3)} = ((H^0_{\rm per} - \epsilon^0_{\rm F})^{1/2}\widetilde{v}_\nu, \phi)_{L^2(\mathbb{R}^3)}.
\end{aligned}$$

As a consequence, $w = (H^0_{\rm per} - \epsilon^0_{\rm F})^{1/2}\widetilde{v}_\nu$.

Using the weak convergence of w_{k_n} to $w = (H^0_{\rm per} - \epsilon^0_{\rm F})^{1/2}\widetilde{v}_\nu$, Fatou's Lemma and Lemma 2.4.6, we thus obtain

$$\begin{aligned}
\mathcal{E}^\nu(\widetilde{v}_\nu) &= \|(H^0_{\rm per} - \epsilon^0_{\rm F})^{1/2}\widetilde{v}_\nu\|^2_{L^2(\mathbb{R}^3)} \\
&\quad + \int_{\mathbb{R}^3} \left(|u^0_{\rm per} + \widetilde{v}_\nu|^{10/3} - |u^0_{\rm per}|^{10/3} - \frac{5}{3}|u^0_{\rm per}|^{4/3}(2u^0_{\rm per}\widetilde{v}_\nu + \widetilde{v}_\nu^2) \right) \\
&\quad + \frac{1}{2} D\left(2u^0_{\rm per}\widetilde{v}_\nu + \widetilde{v}_\nu^2 - \nu, 2u^0_{\rm per}\widetilde{v}_\nu + \widetilde{v}_\nu^2 - \nu\right) \\
&\leq \liminf_{n\to\infty} \mathcal{E}^\nu(v_{\nu,q,L_{k_n}}) = \liminf_{L\to\infty} \mathcal{E}^\nu(v_{\nu,q,L}).
\end{aligned}$$

Hence (2.80). Collecting (2.79) and (2.80), we obtain that $\mathcal{E}^\nu(\widetilde{v}_\nu) \leq \mathcal{E}^\nu(v_\nu)$ and therefore that $\widetilde{v}_\nu = v_\nu$ since $\widetilde{v}_\nu \in \mathcal{Q}_+$ and (2.16) has a unique minimizer.

2.4.8 Thermodynamic limit without a charge constraint

Let $(v_n)_{n\in\mathbb{N}}$ be a minimizing sequence for (2.33). For all $\eta > 0$, for n large enough,

$$\beta\|v_n\|^2_{H^1_{\rm per}(\Gamma_L)} \leq \mathcal{E}^\nu_L(v_n) \leq \mathcal{E}^\nu_L(0) + \eta = \frac{1}{2} D_{\mathcal{R}_L}(\nu_L, \nu_L) + \eta.$$

Thus, $(v_n)_{n\in\mathbb{N}}$ is bounded in $H^1_{\rm per}(\Gamma_L)$. Extracting a converging subsequence and passing to the liminf in the energy, we obtain a minimizer $v_{\nu,L}$ of (2.33), such that

$$\beta\|v_{\nu,L}\|^2_{H^1_{\rm per}(\Gamma_L)} \leq \frac{1}{2} D_{\mathcal{R}_L}(\nu_L, \nu_L). \tag{2.81}$$

We also get

$$D_{\mathcal{R}_L}(u^0_{\rm per}v_{\nu,L}, u^0_{\rm per}v_{\nu,L}) \leq \widetilde{C}, \tag{2.82}$$

for some constant $\widetilde{C}$ independent of L.

Clearly, $u^0_{\rm per} + v_{\nu,L}$ is a non-negative minimizer of

$$\inf\left\{E^{\rm TFW}_{\mathcal{R}_L}(\rho^{\rm nuc}_{\rm per} + \nu_L, w_L),\ w_L \in H^1_{\rm per}(\Gamma_L)\right\}.$$

Reasoning as in the proof of Proposition 2.2.1, we obtain that $u^0_{\rm per} + v_{\nu,L}$ is the only non-negative minimizer of the above problem, and therefore that $v_{\nu,L}$ is the unique minimizer of (2.33). Besides, $v_{\nu,L} \in H^3_{\rm per}(\Gamma_L)$, $u^0_{\rm per} + v_{\nu,L} > 0$ in $\mathbb{R}^3$, and $v_{\nu,L}$ is solution to the Euler equation (2.35), which holds in $H^{-1}_{\rm per}(\Gamma_L)$. Taking $u^0_{\rm per}$ as a test function, we get

$$\begin{aligned}\int_{\Gamma_L} \frac{5}{3}\left(|u^0_{\rm per} + v_{\nu,L}|^{7/3} - |u^0_{\rm per}|^{7/3} - |u^0_{\rm per}|^{4/3}v_{\nu,L}\right)u^0_{\rm per} \\ + D_{\mathcal{R}_L}\left((2u^0_{\rm per}v_{\nu,L} + v^2_{\nu,L} - \nu_L), v_{\nu,L}u^0_{\rm per}\right) + D_{\mathcal{R}_L}\left((2u^0_{\rm per}v_{\nu,L} + v^2_{\nu,L} - \nu_L), (u^0_{\rm per})^2\right) = 0.\end{aligned}$$

We now remark that the third term can be rewritten as

$$\begin{aligned}D_{\mathcal{R}_L}\left((2u^0_{\rm per}v_{\nu,L} + v^2_{\nu,L} - \nu_L), (u^0_{\rm per})^2\right) &= g_1 Z L^2\left(\int_{\Gamma_L}(2u^0_{\rm per}v_{\nu,L} + v^2_{\nu,L} - \nu_L)\right) \\ &\quad + \int_{\Gamma_L}(2u^0_{\rm per}v_{\nu,L} + v^2_{\nu,L} - \nu_L)W^0_{\rm per}, \end{aligned}\tag{2.83}$$

where, as above, $g_1 = |\Gamma_1|^{-1}\int_{\Gamma_1} G_{\mathcal{R}_1}$ and where $W^0_{\rm per}$ is the unique solution in $H^2_{\rm per}(\Gamma_1)$ to

$$\left\{\begin{array}{l} -\Delta W^0_{\rm per} = 4\pi\left(\rho^0_{\rm per} - |\Gamma_1|^{-1}Z\right) \\ W^0_{\rm per}\ \mathcal{R}_1\text{-periodic}, \quad \displaystyle\int_{\Gamma_1} W^0_{\rm per} = 0. \end{array}\right.$$

We finally obtain

$$\begin{aligned}g_1 Z L^2\left(\int_{\Gamma_L}(\nu - (2u^0_{\rm per}v_{\nu,L} + v^2_{\nu,L}))\right) &= \int_{\Gamma_L}\frac{5}{3}\left(|u^0_{\rm per} + v_{\nu,L}|^{7/3} - |u^0_{\rm per}|^{7/3} - |u^0_{\rm per}|^{4/3}v_{\nu,L}\right)u^0_{\rm per} \\ &\quad + D_{\mathcal{R}_L}\left((2u^0_{\rm per}v_{\nu,L} + v^2_{\nu,L} - \nu_L), v_{\nu,L}u^0_{\rm per}\right) \\ &\quad + \int_{\Gamma_L}(2u^0_{\rm per}v_{\nu,L} + v^2_{\nu,L} - \nu_L)W^0_{\rm per}.\end{aligned}$$

As the right hand side is bounded by $CL^{3/2}$ for a constant C independent of L, it holds

$$\lim_{L\to\infty}\int_{\Gamma_L}(\nu - (2u^0_{\rm per}v_{\nu,L} + v^2_{\nu,L})) = 0.$$

Proceeding *mutatis mutandis* as in the previous section, it can be shown that the sequence $(v_{\nu,L})_{L\in\mathbb{N}^*}$ converges weakly in $H^1_{\rm loc}(\mathbb{R}^3)$ and strongly in $L^p_{\rm loc}(\mathbb{R}^3)$ for all $1 \leq p < 6$, towards the unique minimizer v_ν of (2.16).

Acknowledgements

The authors are grateful to Claude Le Bris, Mathieu Lewin and Gabriel Stoltz, as well as to the anonymous referees, for useful comments and suggestions.

Chapter 3

Periodic Schrödinger operators with local defects and spectral pollution

The results of this chapter are gathered in an article, written with Eric Cancès and Yvon Maday, which was submitted to *SIAM Journal of Numerical Analysis* . In this article, we prove that standard Galerkin methods for the discretization of perturbed periodic Schrödinger operators are prone to spectral pollution and that the corresponding spurious states can be interpreted as surface states. In a one-dimensional setting, we can exactly characterize the set of spurious eigenvalues. This is proved in Section 3.5.1 of this chapter.

We also prove that if periodic boundary conditions (instead of Dirichlet boundary conditions) are used, if the simulation domain contains an integer number of unit cells and a Fourier discretization is used, the supercell method does not produce any pollution. However, when the simulation domain does not contain an integer number of unit cells, spectral pollution occurs. Numerical simulations, whose results are presented in Section 3.5.2, were performed to illustrate this situation.

Lastly, another method to circumvent the problem of spectral pollution is proposed. It relies on the use of augmented finite elements discretization using the so-called Wannier functions associated with the unperturbed periodic Schrödinger operator. Details on how this method can be implemented in practice are given in Section 3.5.3.

Periodic Schrödinger operators with local defects and spectral pollution[1]

Eric Cancès[2] Virginie Ehrlacher[2] Yvon Maday[3]

Abstract

This article deals with the numerical calculation of eigenvalues of perturbed periodic Schrödinger operators located in spectral gaps. Such operators are encountered in the modeling of the electronic structure of crystals with local defects, and of photonic crystals. The usual finite element Galerkin approximation is known to give rise to spectral pollution. In this article, we give a precise description of the corresponding spurious states. We then prove that the supercell model does not produce spectral pollution. Lastly, we extend results by Lewin and Séré on some no-pollution criteria. In particular, we prove that using approximate spectral projectors enables one to eliminate spectral pollution in a given spectral gap of the reference periodic Schrödinger operator.

3.1 Introduction

Periodic Schrödinger operators are encountered in the modeling of the electronic structure of crystals, as well as the study of photonic crystals. They are self-adjoint operators on $L^2(\mathbb{R}^d)$ with domain $H^2(\mathbb{R}^d)$ of the form

$$H^0_{\rm per} = -\Delta + V_{\rm per},$$

where Δ is the Laplace operator and $V_{\rm per}$ a $\mathcal{R}$-periodic function of $L^p_{\rm loc}(\mathbb{R}^d)$ ($\mathcal{R}$ being a periodic lattice of $\mathbb{R}^d$), with $p = 2$ if $d \leq 3$, $p > 2$ for $d = 4$ and $p = d/2$ for $d \geq 5$.

Such operators describe perfect crystals, by contrast with real crystals, in which the underlying periodic structure is perturbed by the presence of local or extended defects. In solid state physics, local defects are due to impurities, vacancies, or interstitial atoms, while extended defects correspond to dislocations or grain boundaries. The properties of the crystal can be dramatically affected by the presence of defects. In this article, we consider the case of a d-dimensional crystal with a single local defect, whose properties are encoded in the perturbed periodic Schrödinger operator

$$H = H^0_{\rm per} + W = -\Delta + V_{\rm per} + W, \qquad W \in L^\infty(\mathbb{R}^d), \qquad W(x) \underset{|x|\to\infty}{\to} 0. \tag{3.1}$$

[1]This work was financially supported by the ANR grant MANIF.

[2]Université Paris Est, CERMICS, Projet MICMAC, Ecole des Ponts ParisTech - INRIA, 6 & 8 avenue Blaise Pascal, 77455 Marne-la-Vallée Cedex 2, France, (cances@cermics.enpc.fr, ehrlachv@cermics.enpc.fr)

[3]Université Pierre et Marie Curie-Paris 6, UMR 7598, Laboratoire J.-L. Lions, Paris, F-75005 France, and Division of Applied Mathematics, Brown University, 182 George Street, Providence, RI 02912, USA, (maday@ann.jussieu.fr)

Note that we do not assume here that W is compactly supported. This allows us in particular to handle the mean-field model considered in [37]. In the latter model, $d = 3$ and the self-consistent potential W generated by the defect is of the form $W = \rho \star |\cdot|^{-1}$ with $\rho \in L^2(\mathbb{R}^3) \cap \mathcal{C}$, $\mathcal{C}$ denoting the Coulomb space. Such potentials are continuous and vanish at infinity, but are not compactly supported in general.

Computing the spectrum of the operator H is a key step to understand the properties of the system. It is well known that the self-adjoint operator $H^0_{\rm per}$ is bounded from below on $L^2(\mathbb{R}^d)$, and that the spectrum $\sigma(H^0_{\rm per})$ of $H^0_{\rm per}$ is purely absolutely continuous, and composed of a finite or countable number of closed intervals of $\mathbb{R}$ [163]. The open interval laying between two such closed intervals is called a spectral gap. The multiplication operator W being a compact perturbation of $H^0_{\rm per}$, it follows from Weyl's theorem [163] that H is self-adjoint on $L^2(\mathbb{R}^d)$ with domain $H^2(\mathbb{R}^d)$, and that H and $H^0_{\rm per}$ have the same essential spectrum:

$$\sigma_{\rm ess}(H) = \sigma_{\rm ess}(H^0_{\rm per}) = \sigma(H^0_{\rm per}).$$

Contrarily to $H^0_{\rm per}$, which has no discrete spectrum, H may possess discrete eigenvalues. While the discrete eigenvalues located below the minimum of $\sigma_{\rm ess}(H)$ are easily obtained by standard variational approximations (in virtue of the Rayleigh-Ritz theorem [163]), it is more difficult to compute numerically the discrete eigenvalues located in spectral gaps, for spectral pollution may occur [24].

In Section 3.2, we recall that the usual finite element Galerkin approximation may give rise to spectral pollution [24], and give a precise description of the corresponding spurious states. In Section 3.3, we show that the supercell model does not produce spectral pollution. Lastly, we extend in Section 3.4 results by Lewin and Séré [134] on some no-pollution criteria, which guarantee in particular that the numerical method introduced in [37], involving approximate spectral projectors, and is spectral pollution free.

3.2 Galerkin approximation

The discrete eigenvalues of H and the associated eigenvectors can be obtained by solving the variational problem

$$\left\{ \begin{array}{l} \text{find } (\psi, \lambda) \in H^1(\mathbb{R}^d) \times \mathbb{R} \text{ such that} \\ \forall \phi \in H^1(\mathbb{R}^d),\ a(\psi, \phi) = \lambda \langle \psi, \phi \rangle_{L^2}, \end{array} \right.$$

where $\langle \cdot, \cdot \rangle_{L^2}$ is the scalar product of $L^2(\mathbb{R}^d)$ and a the bilinear form associated with H:

$$a(\psi, \phi) = \int_{\mathbb{R}^d} \nabla\psi \cdot \nabla\phi + \int_{\mathbb{R}^d} (V_{\rm per} + W)\psi\phi.$$

A sequence $(X_n)_{n\in\mathbb{N}}$ of finite dimensional subspaces of $H^1(\mathbb{R}^d)$ being given, we consider for all $n \in \mathbb{N}$, the self-adjoint operator $H|_{X_n} : X_n \to X_n$ defined by

$$\forall (\psi_n, \phi_n) \in X_n \times X_n,\ \langle H|_{X_n}\psi_n, \phi_n \rangle_{L^2} = a(\psi_n, \phi_n).$$

The so-called Galerkin method consists in approximating the spectrum of the operator H by the eigenvalues of the discretized operators $H|_{X_n}$ for n large enough, the latter being obtained by solving the variational problem

$$\left\{ \begin{array}{l} \text{find } (\psi_n, \lambda_n) \in X_n \times \mathbb{R} \text{ such that} \\ \forall \phi_n \in X_n,\ a(\psi_n, \phi_n) = \lambda_n \langle \psi_n, \phi_n \rangle_{L^2}. \end{array} \right. \tag{3.2}$$

According to the Rayleigh-Ritz theorem [163], under the natural assumption that the sequence $(X_n)_{n\in\mathbb{N}}$ satisfies

$$\forall \phi \in H^1(\mathbb{R}^d),\ \inf_{\phi_n \in X_n} \|\phi - \phi_n\|_{H^1} \underset{n\to\infty}{\longrightarrow} 0, \tag{3.3}$$

the Galerkin method allows to compute the eigenmodes of H associated with the discrete eigenvalues located below the bottom of the essential spectrum. It is also known (see e.g. [56] for details) that, as H is bounded below, (3.3) implies

$$\sigma(H) \subset \liminf_{n\to\infty} \sigma\left(H|_{X_n}\right), \tag{3.4}$$

where the right-hand side is the limit inferior of the sets $\sigma\left(H|_{X_n}\right)$, that is the set of the complex numbers λ such that there exists a sequence $(\lambda_n)_{n\in\mathbb{N}}$, with $\lambda_n \in \sigma(H|_{X_n})$ for each $n \in \mathbb{N}$, converging toward λ. In particular, any discrete eigenvalue λ of the operator H is well-approximated by a sequence of eigenvalues of the discretized operators $H|_{X_n}$. On the other hand, (3.3) is not strong enough an assumption to prevent spectral pollution. Some sequences of eigenvalues of $\sigma(H|_{X_n})$ may indeed converge to a real number which does not belong to the spectrum of H:

$$\limsup_{n\to\infty} \sigma\left(H|_{X_n}\right) \nsubseteq \sigma(H) \quad \text{in general,} \tag{3.5}$$

where the limit superior of the sets $\sigma\left(H|_{X_n}\right)$ is the set of the complex numbers λ such that there exists a subsequence $(\sigma(H|_{X_{n_k}}))_{k\in\mathbb{N}}$ of $(\sigma(H|_{X_n}))_{n\in\mathbb{N}}$ for which

$$\forall k \in \mathbb{N}, \quad \exists \lambda_{n_k} \in \sigma(H|_{X_{n_k}}) \quad \text{and} \quad \lim_{k\to\infty} \lambda_{n_k} = \lambda.$$

Spectral pollution has been observed in many situations in physics and mechanics, and this phenomenon is now well-documented (see e.g. [66] and references therein). In [24], Boulton and Levitin report numerical simulations on perturbed periodic Schrödinger operators showing that "*the natural approach of truncating $\mathbb{R}^d$ to a large compact domain and applying the projection method to the corresponding Dirichlet problem is prone to spectral pollution*". Truncating $\mathbb{R}^d$ indeed seems reasonable since it is known that the bound states of H decay exponentially fast at infinity [147, 168]. The following result provides details on the behavior of the spurious modes when the approximation space is constructed using the finite element method.

Proposition 3.2.1. *Let $(\mathcal{T}_n^\infty)_{n\in\mathbb{N}}$ be a sequence of uniformly regular meshes of $\mathbb{R}^d$, invariant with respect to the translations of the lattice $\mathcal{R}$, and such that $h_n := \max_{K\in\mathcal{T}_n^\infty} \mathrm{diam}(K) \to_{n\to\infty} 0$. Let $(\Omega_n)_{n\in\mathbb{N}}$ be an increasing sequence of closed convex sets of $\mathbb{R}^d$ converging to $\mathbb{R}^d$, $\mathcal{T}_n := \{K \in \mathcal{T}_n^\infty \,|\, K \subset \Omega_n\}$ and X_n the finite-dimensional approximation space of $H_0^1(\Omega_n) \hookrightarrow H^1(\mathbb{R}^d)$ obtained with $\mathcal{T}_n$ and $\mathbb{P}_m$ finite elements ($m \in \mathbb{N}^*$). Let $\lambda \in \limsup_{n\to\infty} \sigma\left(H|_{X_n}\right) \setminus \sigma(H)$ and $(\psi_{n_k}, \lambda_{n_k}) \in X_{n_k} \times \mathbb{R}$ be such that $H|_{X_{n_k}}\psi_{n_k} = \lambda_{n_k}\psi_{n_k}$, $\|\psi_{n_k}\|_{L^2} = 1$ and $\lim_{k\to\infty} \lambda_{n_k} = \lambda$. Then, the sequence $(\psi_{n_k})_{k\in\mathbb{N}}$, considered as a sequence of functions of $H^1(\mathbb{R}^d)$, converges to 0 weakly in $H^1(\mathbb{R}^d)$ and strongly in $L^q_{\mathrm{loc}}(\mathbb{R}^d)$, with $q = \infty$ if $d = 1$, $q < \infty$ if $d = 2$ and $q < 2d/(d-2)$ if $d \geq 3$, in the sense that*

$$\forall K \subset \mathbb{R}^d, \quad K \text{ compact}, \quad \int_K |\psi_{n_k}|^q \underset{k\to\infty}{\longrightarrow} 0,$$

and it holds

$$\forall \epsilon > 0, \quad \exists R > 0 \quad s.\ t. \quad \liminf_{k\to\infty} \int_{\partial\Omega_{n_k} + B(0,R)} |\psi_{n_k}|^2 \geq 1 - \epsilon. \tag{3.6}$$

The latter result shows that the mass of the spurious states concentrates on the boundary of the simulation domain Ω_{n_k}.

This phenomenon is clearly observed on the two dimensional numerical simulations reported below, which have been performed with the finite element software FreeFem++ [1], with $V_{\rm per}(x,y) = \cos(x) + 3\sin(2(x+y)+1)$ and $W(x,y) = -(x+2)^2(2y-1)^2\exp(-(x^2+y^2))$. We have checked numerically, using the Bloch decomposition method, that there is a gap (α,β), with $\alpha \simeq -0.341$ and $\beta \simeq 0.016$, between the first and second bands of $H^0_{\rm per} = -\Delta + V_{\rm per}$. We have also checked numerically, using the pollution free supercell method (see Theorem 3.3.1 below), that $H = H^0_{\rm per} + W$ has exactly one eigenvalue in the gap (α,β) approximatively equal to -0.105. Our simulations have been performed with a sequence of $\mathbb{P}_1$-finite element approximation spaces $(X_n)_{40\leq n\leq 100}$, where for each $40 \leq n \leq 100$,

- $\Omega_n = \left[-4\pi\dfrac{m_n}{n}, 4\pi\dfrac{m_n}{n}\right]^2$, with $m_n = \left[n\left(\dfrac{n-40}{20}+5\right)\right]$;
- $\mathcal{T}_n^\infty$ is a uniform $2\pi\mathbb{Z}^2$-periodic mesh of $\mathbb{R}^2$ consisting of $2n^2$ isometrical isoceles rectangular triangles per unit cell.

The spectra of $H|_{X_n}$ in the gap (α,β) for $40 \leq n \leq 100$ are displayed on Fig. 3.1. We clearly see that all these operators have an eigenvalue close to -0.1, which is an approximation of a true eigenvalue of H. The corresponding eigenfunction for $n = 88$ (blue circle on Fig. 3.1) is displayed on Fig. 3.2 (top); as expected, it is localized in the vicinity of the defect. On the other hand, most of these discretized operators have several eigenvalues in the range (α,β), which cannot be associated with an eigenvalue of H, and can be interpreted as spurious modes. The eigenfunction of $H|_{X_n}$ close to -0.290, obtained for $n = 88$ (blue square on Fig. 3.1), is displayed on Fig. 3.2 (bottom); in agreement with the analysis carried out in Proposition 3.2.1, it is localized in the vicinity of the boundary of the computational domain.

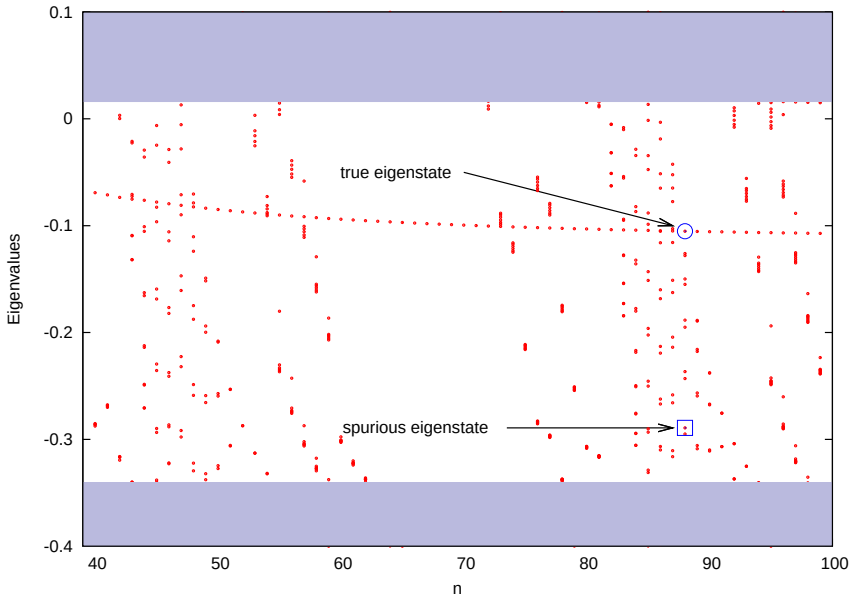

Figure 3.1: Spectrum of $H|_{X_n}$ in the gap (α,β) for $40 \leq n \leq 100$

Proposition 3.2.1 provides a characterization of the spurious eigenstates. Note that for any $\lambda \in \mathrm{Conv}(\sigma_{\mathrm{ess}}(H)) \setminus \sigma(H)$, there exist sequences of simulation domains $(\Omega_n)_{n\in\mathbb{N}}$ and meshes $(\mathcal{T}_n^\infty)_{n\in\mathbb{N}}$ satisfying the assumptions of Proposition 3.2.1 and such that, if X_n is the finite-dimensional approximation space of $H_0^1(\Omega_n) \hookrightarrow H^1(\mathbb{R}^d)$ obtained with Ω_n, $\mathcal{T}_n^\infty$ and $\mathbb{P}_m$ finite elements, then $\lambda \in \liminf_{n\to\infty} \sigma\left(H|_{X_n}\right)$.

To see this, let us consider for each $t > 0$ the domain $\Omega_t := B(0,t)$ and the self-adjoint operator $H_t := -\Delta + V_{\mathrm{per}} + W$ on $L^2(\Omega_t)$ with domain $H^2(\Omega_t) \cap H_0^1(\Omega_t)$. The operator H_t being bounded below, with compact resolvent, its spectrum is purely discrete. We denote by $\epsilon_1(t) \le \epsilon_2(t) \le \epsilon_3(t) \le \cdots$ its eigenvalues counted with their multiplicities. It readily follows from the min-max principle (see e.g. [163]) that

- for each $t > 0$, $\epsilon_j(t) \underset{j\to+\infty}{\uparrow} +\infty$;
- for each $j \in \mathbb{N}^*$, the function $t \mapsto \epsilon_j(t)$ is non-increasing and continuous, and satisfies

$$\epsilon_j(t) \underset{t\to+\infty}{\downarrow} \lambda_j(H) := \inf_{Y_j\in\mathcal{Y}_j} \sup_{v\in Y_j\setminus\{0\}} \frac{a(v,v)}{\|v\|_{L^2}^2} \le \min \sigma_{\mathrm{ess}}(H_{\mathrm{per}}^0),$$

  where $\mathcal{Y}_j$ is the set of the vector subspaces of $H^1(\mathbb{R}^d)$ of dimension j.

This implies that for all $\lambda \in \mathrm{Conv}(\sigma_{\mathrm{ess}}(H)) \setminus \sigma(H)$, there exists a sequence $(t_n)_{n\in\mathbb{N}}$ such that $t_n \underset{n\to\infty}{\longrightarrow} +\infty$ and $\mathrm{dist}(\lambda, \sigma(H_{t_n})) \underset{n\to\infty}{\longrightarrow} 0$. Besides, for all $t > 0$, if $(\mathcal{T}_n^\infty)_{n\in\mathbb{N}}$ is a sequence of uniformly regular meshes such that $h_n := \max_{K\in\mathcal{T}_n^\infty} \mathrm{diam}(K)$ goes to 0 as n goes to infinity, it holds that $\lim_{n\to\infty} H|_{X_n^t} = \sigma(H_t)$, where X_n^t is the $\mathbb{P}_m$ finite-element discretization space of $H_0^1(\Omega_t) \hookrightarrow H^1(\mathbb{R}^d)$ built from $\mathcal{T}_n^\infty$. Thus, for all $\lambda \in \mathrm{Conv}(\sigma_{\mathrm{ess}}(H)) \setminus \sigma(H)$, there exists an increasing sequence $(t_n)_{n\in\mathbb{N}}$ of positive real numbers going to infinity, and an increasing mapping $\phi : \mathbb{N}^* \to \mathbb{N}^*$ such that

$$\lambda \in \liminf_{n\to\infty} \sigma(H|_{X_{\phi(n)}^{t_n}}).$$

A natural question is of course whether spectral pollution occurs for a given sequence of simulation domains $(\Omega_n)_{n\in\mathbb{N}}$ and meshes $(\mathcal{T}_n^\infty)_{n\in\mathbb{N}}$ satisfying the assumptions of Proposition 3.2.1. However, it seems difficult to answer this question with a pen and paper analysis.

Remark 3.2.1. *Using the results in [185], it is possible to characterize the spurious states generated by finite element discretizations of one-dimensional perturbed Schrödinger operators: for $\mathcal{R} = b\mathbb{Z}$ and $\Omega_n = [-(n+t)b, (n+t)b]$, the spurious eigenvalues are the discrete eigenvalues in $[\min(\sigma(H_{\mathrm{per}}^0)), +\infty) \setminus \sigma(H)$ of the operators $H^+(t)$ on $L^2(\mathbb{R}_+)$ with domain $H^2(\mathbb{R}_+) \cap H_0^1(\mathbb{R}_+)$ and $H^-(t)$ on $L^2(\mathbb{R}_-)$ with domain $H^2(\mathbb{R}_-) \cap H_0^1(\mathbb{R}_-)$, respectively defined by $H^\pm(t) = -\dfrac{d^2}{dx^2} + V_{\mathrm{per}}(x \mp tb)$. Besides, the spurious eigenvectors of $H|_{X_n}$ converge (in some sense, and up to translation) to the discrete eigenvectors of $H^\pm(t)$. As*

$$\left(\bigcup_{t\in[0,b)} \sigma(H^\pm(t)) \right) \cap [\min(\sigma(H_{\mathrm{per}}^0)), +\infty) = [\min(\sigma(H_{\mathrm{per}}^0)), +\infty),$$

any $\lambda \in [\min(\sigma(H_{\mathrm{per}}^0)), +\infty) \setminus \sigma(H)$ is a spurious eigenvalue, in the sense that there exists an increasing sequence $(\Omega_n)_{n\in\mathbb{N}}$ of closed intervals of $\mathbb{R}$ converging to $\mathbb{R}$ such that

$$\lambda \in \liminf_{n\to\infty} \sigma(H|_{X_n}).$$

We refer to Section 3.5.1 of Appendix 3.5 for a proof and a numerical illustration of this result. The proof of similar results for $d \ge 2$ is work in progress.

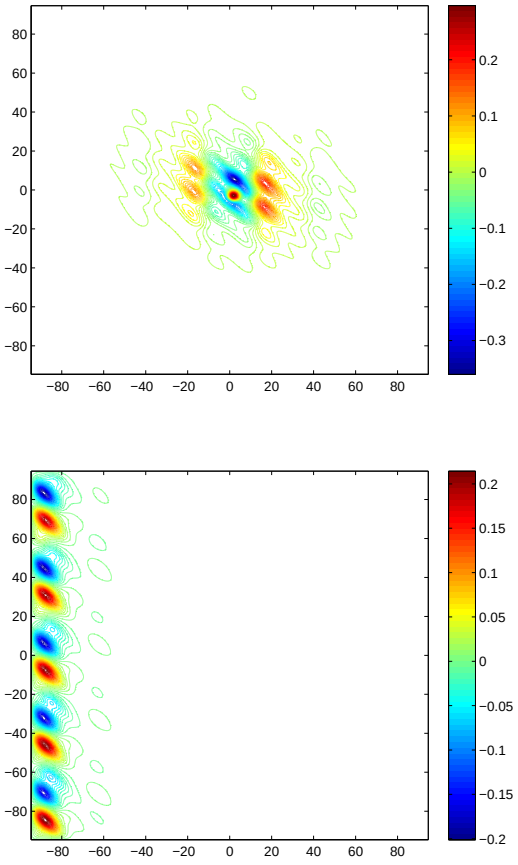

Figure 3.2: A true eigenfunction, localized close to the defect (top), and a "spurious" eigenfunction, localized close to the boundary (bottom).

Proof of Proposition 3.2.1.

Convergence of $(\psi_{n_k})_{k\in\mathbb{N}}$. We first notice that, since $H = -\frac{1}{2}\Delta + \frac{1}{2}(-\Delta + 2V_{\rm per}) + W$, with W bounded in $L^\infty(\mathbb{R}^d)$ and $-\Delta + 2V_{\rm per}$ bounded below, there exists a constant $C \in \mathbb{R}_+$ such that

$$\forall \psi \in H^1(\mathbb{R}^d), \quad a(\psi,\psi) \geq \frac{1}{2}\|\nabla\psi\|_{L^2}^2 - C\|\psi\|_{L^2}^2. \tag{3.7}$$

As

$$\forall k \in \mathbb{N}, \quad \|\psi_{n_k}\|_{L^2} = 1 \quad \text{and} \quad a(\psi_{n_k},\psi_{n_k}) = \lambda_{n_k} \underset{k\to\infty}{\longrightarrow} \lambda,$$

we infer from (3.7) that the sequence $(\psi_{n_k})_{k\in\mathbb{N}}$ is bounded in $H^1(\mathbb{R}^d)$. It therefore converges, up to extraction, to some function $\phi \in H^1(\mathbb{R}^d)$, weakly in $H^1(\mathbb{R}^d)$, and strongly in $L^q_{\rm loc}(\mathbb{R}^d)$ with $q = \infty$ if $d = 1$, $q < \infty$ if $d = 2$ and $q < 2d/(d-2)$ if $d \geq 3$. It is easy to deduce from (3.3) and the continuity of a on $H^1(\mathbb{R}^d) \times H^1(\mathbb{R}^d)$ that ϕ satisfies $H\phi = \lambda\phi$ and therefore that $\phi = 0$ since $\lambda \notin \sigma(H)$ by assumption. Consequently, the whole sequence $(\psi_{n_k})_{k\in\mathbb{N}}$ converges to zero weakly in $H^1(\mathbb{R}^d)$ and strongly in $L^q_{\rm loc}(\mathbb{R}^d)$.

Proof of (3.6). We argue by contradiction. Assume that there exists $\epsilon > 0$ such that

$$\forall R > 0, \quad \liminf_{k\to\infty} \int_{\partial\Omega_{n_k}+B(0,R)} |\psi_{n_k}|^2 < 1 - \epsilon.$$

As $\|\psi_{n_k}\|_{L^2} = 1$ for all k, the above inequality also reads

$$\forall R > 0, \quad \limsup_{k\to\infty} \int_{\Omega_{n_k}^R} |\psi_{n_k}|^2 > \epsilon,$$

where $\Omega_{n_k}^R = \{x \in \Omega_{n_k} \,|\, d(x, \partial\Omega_{n_k}) \geq R\}$. We could then extract from $(\psi_{n_k})_{k\in\mathbb{N}}$ a subsequence, still denoted by $(\psi_{n_k})_{k\in\mathbb{N}}$, such that there exists an increasing sequence $(R_{n_k})_{k\in\mathbb{N}}$ of real numbers going to infinity for which

$$\forall k \in \mathbb{N}, \quad \int_{\Omega_{n_k}^{R_{n_k}}} |\psi_{n_k}|^2 \geq \epsilon.$$

In the sequel, we denote by

$$C^0(\mathcal{T}_n^\infty) = \left\{ v \in C^0(\mathbb{R}^d) \mid \forall K \in \mathcal{T}_n^\infty,\ v|_K \in \mathbb{P}_m \right\}$$

the set of continuous functions built from $\mathcal{T}_n^\infty$ and $\mathbb{P}_m$-finite elements, by P_n the interpolation operator from $C^0(\mathbb{R}^d)$ onto $C^0(\mathcal{T}_n^\infty)$, and by

$$X_n^\infty = C^0(\mathcal{T}_n^\infty) \cap H^1(\mathbb{R}^d).$$

The space X_n^∞ is an (infinite dimensional) closed subspace of $H^1(\mathbb{R}^d)$. Obviously $X_n \hookrightarrow X_n^\infty$. We then introduce a sequence $(\chi_{n_k})_{k\in\mathbb{N}}$ of functions of $C_c^\infty(\mathbb{R}^d)$ such that for all $k \in \mathbb{N}$,

$$\mathrm{Supp}(\chi_{n_k}) \subset \Omega_{n_k},\ \chi_k \equiv 1 \text{ on } \Omega_{n_k}^{R_{n_k}},\ \text{and } \forall |\alpha| \leq (m+1),\ \|\partial^\alpha \chi_{n_k}\|_{L^\infty} \leq C R_{n_k}^{-|\alpha|}, \quad (3.8)$$

for a constant $C \in \mathbb{R}_+$ independent of k.

The rest of the proof is divided into three steps.

Step 1. We first prove that for all $k \in \mathbb{N}$ such that $h_{n_k} \leq 1$,

$$\forall \phi_{n_k}^\infty \in X_{n_k}^\infty, \quad \|\chi_{n_k}\phi_{n_k}^\infty - P_{n_k}(\chi_{n_k}\phi_{n_k}^\infty)\|_{H^1} \leq C h_{n_k} R_{n_k}^{-1} \|\phi_{n_k}^\infty\|_{H^1}, \quad (3.9)$$

for some constant C independent of k and $\phi_{n_k}^\infty$.

Let us recall some classical direct and inverse inequalities used in finite element analysis (see e.g. [83]). Since the sequence of meshes $(\mathcal{T}_n^\infty)_{n\in\mathbb{N}}$ is assumed to be uniformly regular, there exists $C \in \mathbb{R}_+$ such that for all $n \in \mathbb{N}$ and $K \in \mathcal{T}_n^\infty$, we have

- for all $v \in \mathcal{C}^0(\mathbb{R}^d)$,

$$\|v|_K - (P_n v)|_K\|_{H^1(K)} \leq C h_n^m \max_{|\alpha|=m+1} \|\partial^\alpha v|_K\|_{L^2(K)}, \quad (3.10)$$

- for all $v_n^\infty \in X_n^\infty$ and $\beta \in \mathbb{N}^d$ with $2 \leq |\beta| \leq m$,

$$\|\partial^\beta v_n^\infty|_K\|_{L^2(K)} \leq C h_n^{-(|\beta|-1)} \|v_n^\infty|_K\|_{H^1(K)}. \quad (3.11)$$

To prove (3.9), we notice that for all $K \in \mathcal{T}_{n_k}$, $(\chi_{n_k}\phi^\infty_{n_k})|_K \in C^\infty(K)$, and $\partial^\beta \phi^\infty_{n_k}|_K = 0$ if $|\beta| = m+1$, so that

$$\begin{aligned}
\|\chi_{n_k}\phi^\infty_{n_k} - P_{n_k}(\chi_{n_k}\phi^\infty_{n_k})\|^2_{H^1} &= \sum_{K\in\mathcal{T}_{n_k}} \|(\chi_{n_k}\phi^\infty_{n_k})|_K - (P_{n_k}(\chi_{n_k}\phi^\infty_{n_k}))|_K\|^2_{H^1(K)} \\
&\leq Ch^{2m}_{n_k} \sum_{K\in\mathcal{T}_{n_k}} \max_{|\alpha|=m+1} \|\partial^\alpha(\chi_{n_k}\phi^\infty_{n_k})|_K\|^2_{L^2(K)} \\
&\leq Ch^{2m}_{n_k} \sum_{K\in\mathcal{T}_{n_k}} \max_{|\alpha|=m+1} \sum_{\beta\leq\alpha} \|\partial^{\alpha-\beta}\chi_{n_k}\|^2_{L^\infty} \|\partial^\beta\phi^\infty_{n_k}|_K\|^2_{L^2(K)} \\
&\leq Ch^{2m}_{n_k} R^{-2}_{n_k} \sum_{K\in\mathcal{T}_{n_k}} \max_{|\beta|\leq m} \|\partial^\beta\phi^\infty_{n_k}|_K\|^2_{L^2(K)} \\
&\leq Ch^{2m}_{n_k} R^{-2}_{n_k} \sum_{K\in\mathcal{T}_{n_k}} (1+h^{-2(m-1)}_{n_k})\|\phi^\infty_{n_k}|_K\|^2_{H^1(K)} \\
&\leq Ch^2_{n_k} R^{-2}_{n_k} \|\phi^\infty_{n_k}\|^2_{H^1},
\end{aligned}$$

where we have used (3.10), (3.11) and (3.8) to obtain the first, fourth, and third inequalities respectively.

Step 2. Let $\widetilde{\psi}_{n_k} = P_{n_k}(\chi_{n_k}\psi_{n_k})$. In this second step, we are going to prove that

$$\forall \phi^\infty_{n_k} \in X^\infty_{n_k}, \quad \left|(a^0 - \lambda_{n_k})(\widetilde{\psi}_{n_k}, \phi^\infty_{n_k})\right| \leq \eta_{n_k}\|\phi^\infty_{n_k}\|_{H^1}, \tag{3.12}$$

where $(\eta_{n_k})_{k\in\mathbb{N}}$ is a sequence of positive real numbers going to 0 at infinity.

For all $k \in \mathbb{N}$, $\|\widetilde{\psi}_{n_k}\|_{L^2} \geq \epsilon^{1/2}$ and for all $\phi^\infty_{n_k} \in X^\infty_{n_k}$,

$$\begin{aligned}
(a-\lambda_{n_k})(\widetilde{\psi}_{n_k}, \phi^\infty_{n_k}) &= (a-\lambda_{n_k})(\chi_{n_k}\psi_{n_k}, \phi^\infty_{n_k}) \\
&\quad -(a-\lambda_{n_k})(\chi_{n_k}\psi_{n_k} - P_{n_k}(\chi_{n_k}\psi_{n_k}), \phi^\infty_{n_k}) \\
&= (a-\lambda_{n_k})(\psi_{n_k}, \chi_{n_k}\phi^\infty_{n_k}) \\
&\quad -(a-\lambda_{n_k})(\chi_{n_k}\psi_{n_k} - P_{n_k}(\chi_{n_k}\psi_{n_k}), \phi^\infty_{n_k}) \\
&\quad -\int_{\mathbb{R}^d} (\Delta\chi_{n_k}\psi_{n_k}\phi^\infty_{n_k} + 2\phi^\infty_{n_k}\nabla\chi_{n_k}\cdot\nabla\psi_{n_k}) \\
&= (a-\lambda_{n_k})(\psi_{n_k}, \chi_{n_k}\phi^\infty_{n_k} - P_{n_k}(\chi_{n_k}\phi^\infty_{n_k})) \\
&\quad -(a-\lambda_{n_k})(\chi_{n_k}\psi_{n_k} - P_{n_k}(\chi_{n_k}\psi_{n_k}), \phi^\infty_{n_k}) \\
&\quad -\int_{\mathbb{R}^d} (\Delta\chi_{n_k}\psi_{n_k}\phi^\infty_{n_k} + 2\phi^\infty_{n_k}\nabla\chi_{n_k}\cdot\nabla\psi_{n_k}),
\end{aligned}$$

where we have used that $(a-\lambda_{n_k})(\psi_{n_k}, P_{n_k}(\chi_{n_k}\phi^\infty_{n_k})) = 0$ since $P_{n_k}(\chi_{n_k}\phi^\infty_{n_k}) \in X_{n_k}$. Denoting by

$$a^0(\psi,\phi) = \int_{\mathbb{R}^d} \nabla\psi\cdot\nabla\phi + \int_{\mathbb{R}^d} V_{\rm per}\psi\phi,$$

we end up with

$$\begin{aligned}
(a^0-\lambda_{n_k})(\widetilde{\psi}_{n_k}, \phi^\infty_{n_k}) &= (a-\lambda_{n_k})(\psi_{n_k}, \chi_{n_k}\phi^\infty_{n_k} - P_{n_k}(\chi_{n_k}\phi^\infty_{n_k})) \\
&\quad -(a-\lambda_{n_k})(\chi_{n_k}\psi_{n_k} - P_{n_k}(\chi_{n_k}\psi_{n_k}), \phi^\infty_{n_k}) \\
&\quad -\int_{\mathbb{R}^d} (\Delta\chi_{n_k}\psi_{n_k}\phi^\infty_{n_k} + 2\phi^\infty_{n_k}\nabla\chi_{n_k}\cdot\nabla\psi_{n_k}) \\
&\quad -\int_{\mathbb{R}^d} W\widetilde{\psi}_{n_k}\phi^\infty_{n_k}.
\end{aligned} \tag{3.13}$$

Using the boundedness of $(\psi_{n_k})_{k\in\mathbb{N}}$ in $H^1(\mathbb{R}^d)$, the properties of χ_{n_k} and W, and the fact that $(\psi_{n_k})_{k\in\mathbb{N}}$ strongly converges to 0 in $L^2_{\rm loc}(\mathbb{R}^d)$, we deduce from (3.13) and (3.9) that

$$\forall \phi^\infty_{n_k} \in X^\infty_{n_k}, \quad \left|(a^0 - \lambda_{n_k})(\widetilde{\psi}_{n_k}, \phi^\infty_{n_k})\right| \leq \eta_{n_k} \|\phi^\infty_{n_k}\|_{H^1},$$

where the sequence of positive real numbers $(\eta_{n_k})_{k\in\mathbb{N}}$ goes to zero when k goes to infinity.

Step 3. Let us finally construct a particular sequence of test functions $(\phi^\infty_{n_k})_{k\in\mathbb{N}}$ for (3.12), which will allow us to complete our argument by contradiction. We can use Bloch theory (see e.g. [163]) and expand the functions of $X^\infty_{n_k}$ as

$$\phi^\infty_{n_k}(x) = \fint_{\Gamma^*} (\phi^\infty_{n_k})_q(x)\, dq,$$

where Γ^* is the first Brillouin zone of the perfect crystal, and where for all $q \in \Gamma^*$,

$$(\phi^\infty_{n_k})_q(x) = \sum_{R\in\mathcal{R}} \phi^\infty_{n_k}(x+R) e^{-iq\cdot R}.$$

For each $q \in \Gamma^*$, the function $(\phi^\infty_{n_k})_q$ belongs to the complex Hilbert space

$$L^2_q(\Gamma) := \left\{ v(x) e^{iq\cdot x},\ v \in L^2_{\rm loc}(\mathbb{R}^d),\ v\ \mathcal{R}\text{-periodic} \right\},$$

where Γ denotes the Wigner-Seitz cell of the lattice $\mathcal{R}$ (notice that the functions $(\phi^\infty_{n_k})_q$ are complex-valued). Recall that if $\mathcal{R} = b\mathbb{Z}^d$ (cubic lattice of parameter $b > 0$), then $\Gamma = (-b/2, b/2]^d$ and $\Gamma^* = (-\pi/b, \pi/b]^d$. Besides, if for all $\phi_q \in L^2_q(\Gamma)$,

$$\|\phi_q\|^2_{L^2_q(\Gamma)} := \int_\Gamma |\phi_q(x)|^2\, dx,$$

it holds that

$$\left\|\phi^\infty_{n_k}\right\|^2_{L^2} = \fint_{\Gamma^*} \left\|(\phi^\infty_{n_k})_q\right\|^2_{L^2_q(\Gamma)}\, dq.$$

The mesh $\mathcal{T}^\infty_{n_k}$ being invariant with respect to the translations of the lattice $\mathcal{R}$, it holds in fact

$$(\phi^\infty_{n_k})_q \in C^0(\mathcal{T}^\infty_{n_k}) \cap L^2_q(\Gamma).$$

We thus have for all $\phi^\infty_{n_k} \in X^\infty_{n_k}$,

$$(a^0 - \lambda_{n_k})(\widetilde{\psi}_{n_k}, \phi^\infty_{n_k}) = \fint_{\Gamma^*} (a^0_q - \lambda_n)((\widetilde{\psi}_{n_k})_q, (\phi^\infty_{n_k})_q)\, dq,$$

where

$$a^0_q(\psi_q, \phi_q) = \int_\Gamma \nabla\psi^*_q \cdot \nabla\phi_q + \int_\Gamma V_{\rm per}\psi^*_q\phi_q. \tag{3.14}$$

Let $(e_{n,l,q}, \epsilon_{n,l,q})_{1\leq l\leq N_n}$, $\epsilon_{n,1,q} \leq \epsilon_{n,2,q} \leq \cdots \leq \epsilon_{n,N_n,q}$, be an $L^2_q(\Gamma)$-orthonormal basis of eigenmodes of a^0_q in $C^0(\mathcal{T}^\infty_n) \cap L^2_q(\Gamma)$. Expanding $(\widetilde{\psi}_{n_k})_q$ in the basis $(e_{n_k,l,q})_{1\leq l\leq N_{n_k}}$, we get

$$(\widetilde{\psi}_{n_k})_q = \sum_{j=1}^{N_{n_k}} c_{n_k,j,q} e_{n_k,j,q}.$$

Choosing $\phi^\infty_{n_k}$ such that

$$(\phi^\infty_{n_k})_q = \sum_{j=1}^{N_{n_k}} c_{n_k,j,q}(1_{\epsilon_{n_k,j,q}-\lambda_{n_k}\geq 0} - 1_{\epsilon_{n_k,j,q}-\lambda_{n_k}<0}) e_{n_k,j,q},$$

we obtain $\|\phi^\infty_{n_k}\|_{L^2} = \|\widetilde{\psi}_{n_k}\|_{L^2}$ and

$$(a^0 - \lambda_{n_k})(\widetilde{\psi}_{n_k}, \phi^\infty_{n_k}) = \fint_{\Gamma^*} \sum_{j=1}^{N_{n_k}} |\epsilon_{n_k,j,q} - \lambda_{n_k}|\, |c_{n_k,j,q}|^2.$$

It is easy to check that $\liminf_{k\to\infty} \min_{j,q} |\epsilon_{n_k,j,q} - \lambda_{n_k}| = \zeta := \mathrm{dist}(\lambda, \sigma(H^0_{\mathrm{per}})) > 0$. Hence,

$$\liminf_{k\to\infty} (a^0 - \lambda_{n_k})(\widetilde{\psi}_{n_k}, \phi^\infty_{n_k}) \geq \zeta \fint_{\Gamma^*} \sum_{j=1}^{N_{n_k}} |c_{n_k,j,q}|^2 = \zeta \|\widetilde{\psi}_{n_k}\|^2_{L^2} \geq \zeta\epsilon.$$

Besides,

$$\|\phi^\infty_{n_k}\|_{L^2} = \|\widetilde{\psi}_{n_k}\|_{L^2} \quad \text{and} \quad a^0(\phi^\infty_{n_k}, \phi^\infty_{n_k}) = a^0(\widetilde{\psi}_{n_k}, \widetilde{\psi}_{n_k}),$$

which implies that the sequence $(\phi^\infty_{n_k})_{k\in\mathbb{N}}$ is bounded in $H^1(\mathbb{R}^d)$. Consequently,

$$0 < \zeta\epsilon \leq \liminf_{k\to\infty} (a^0 - \lambda_{n_k})(\widetilde{\psi}_{n_k}, \phi^\infty_{n_k}) \leq \liminf_{k\to\infty} \eta_{n_k} \|\phi^\infty_{n_k}\|_{H^1} = 0.$$

We reach a contradiction. □

A careful look on the above proof shows that the assumptions in Proposition 3.2.1 can be weakened: in particular, the mesh $\mathcal{T}_n$ can be refined in the regions where $|W|$ is large, and coarsened in the vicinity of the boundary of Ω_n.

3.3 Supercell method

In solid state physics and materials science, the current state-of-the-art technique to compute the discrete eigenvalues of a perturbed periodic Schrödinger operator in spectral gaps is the supercell method. Let $\mathcal{R}$ be the periodic lattice of the host crystal and Γ its Wigner-Seitz cell. In the case of a cubic lattice of paramater $b > 0$, we have $\mathcal{R} = b\mathbb{Z}^d$ and $\Gamma = (-b/2, b/2]^d$. The supercell method consists in solving the spectral problem

$$\left\{ \begin{array}{l} \text{find } (\psi_{L,N}, \lambda_{L,N}) \in X_{L,N} \times \mathbb{R} \text{ such that} \\ \forall \phi_{L,N} \in X_{L,N},\ a_L(\psi_{L,N}, \phi_{L,N}) = \lambda_{L,N} \langle \psi_{L,N}, \phi_{L,N} \rangle_{L^2_{\mathrm{per}}(\Gamma_L)}, \end{array} \right. \tag{3.15}$$

where $\Gamma_L = L\Gamma$ (with $L \in \mathbb{N}^*$) is the supercell,

$$L^2_{\mathrm{per}}(\Gamma_L) = \left\{ u_L \in L^2_{\mathrm{loc}}(\mathbb{R}^d) \mid u_L \ L\mathcal{R}\text{-periodic} \right\},$$

$$a_L(u_L, v_L) = \int_{\Gamma_L} \nabla u_L \cdot \nabla v_L + \int_{\Gamma_L} (V_{\mathrm{per}} + W) u_L v_L, \quad \langle u_L, v_L \rangle_{L^2_{\mathrm{per}}(\Gamma_L)} = \int_{\Gamma_L} u_L v_L,$$

and $X_{L,N}$ is a finite dimensional subspace of

$$H^1_{\mathrm{per}}(\Gamma_L) = \left\{ u_L \in L^2_{\mathrm{per}}(\Gamma_L) \mid \nabla u_L \in \left(L^2_{\mathrm{per}}(\Gamma_L)\right)^d \right\}.$$

We denote by $H_{L,N} = H_L|_{X_{L,N}}$, where H_L is the unique self-adjoint operator on $L^2_{\mathrm{per}}(\Gamma_L)$ associated with the quadratic form a_L. It then holds that $D(H_L) = H^2_{\mathrm{per}}(\Gamma_L)$,

$$\forall \phi_L \in H^2_{\mathrm{per}}(\Gamma_L), \quad H_L \phi_L = -\Delta \phi_L + (V_{\mathrm{per}} + W_L)\phi_L,$$

and

$$\forall \phi_{L,N} \in X_{L,N}, \quad H_{L,N}\phi_{L,N} = -\Delta\phi_{L,N} + \Pi_{X_{L,N}}\left((V_{\rm per} + W_L)\phi_{L,N}\right),$$

where $W_L \in L^\infty_{\rm per}(\Gamma_L)$ denotes the $L\mathcal{R}$-periodic extension of $W|_{\Gamma_L}$ and $\Pi_{X_{L,N}}$ is the orthogonal projector of $L^2_{\rm per}(\Gamma_L)$ on $X_{L,N}$ for the $L^2_{\rm per}(\Gamma_L)$ inner product.

Again for the sake of clarity, we restrict ourselves to cubic lattices ($\mathcal{R} = b\mathbb{Z}^d$) and to the most popular discretization method for supercell model, namely the Fourier (also called planewave) method. We therefore consider approximation spaces of the form

$$X_{L,N} = \left\{ \sum_{k \in 2\pi(bL)^{-1}\mathbb{Z}^d \,|\, |k| \le 2\pi(bL)^{-1}N} c_k e_{L,k} \;\Big|\; \forall k,\, c_{-k} = c_k^* \right\},$$

where $e_{L,k}(x) = |\Gamma_L|^{-1/2} e^{ik\cdot x}$.

From the classical Jackson inequality for Fourier truncation, we deduce by scaling the following property of the discretization spaces $X_{L,N}$: for all real numbers r and s such that $0 \le r \le s$, there exists a constant $C > 0$ such that for all $L \in \mathbb{N}^*$ and all $\phi_L \in H^s_{\rm per}(\Gamma_L)$,

$$\|\phi_L - \Pi_{X_{L,N}}\phi_L\|_{H^r_{\rm per}(\Gamma_L)} \le C\left(\frac{L}{N}\right)^{s-r} \|\phi_L\|_{H^s_{\rm per}(\Gamma_L)}. \tag{3.16}$$

Our analysis of the supercell method requires some assumption on the potential $V_{\rm per}$. We define the functional space $\mathcal{M}_{\rm per}(\Gamma)$ as

$$\mathcal{M}_{\rm per}(\Gamma) = \left\{ v \in L^2_{\rm per}(\Gamma) \mid \|v\|_{\mathcal{M}_{\rm per}(\Gamma)} := \sup_{L\in\mathbb{N}^*} \sup_{w \in H^1_{\rm per}(\Gamma_L)\setminus\{0\}} \frac{\|vw\|_{L^2_{\rm per}(\Gamma_L)}}{\|w\|_{H^1_{\rm per}(\Gamma_L)}} < \infty \right\}.$$

It is quite standard to prove that $\mathcal{M}_{\rm per}(\Gamma)$ is a normed space and that the space of the $\mathcal{R}$-periodic functions of class C^∞ is dense in $\mathcal{M}_{\rm per}(\Gamma)$. We denote the $\mathcal{R}$-periodic Lorentz spaces [15] by $L^{p,q}_{\rm per}(\Gamma)$.

Proposition 3.3.1. *The following embeddings are continuous:*

$$\begin{aligned} &\text{for } d = 1, \quad L^2_{\rm per}(\Gamma) \hookrightarrow \mathcal{M}_{\rm per}(\Gamma), \\ &\text{for } d = 2, \quad L^{2,\infty}_{\rm per}(\Gamma) \hookrightarrow \mathcal{M}_{\rm per}(\Gamma), \\ &\text{for } d = 3, \quad L^{3,\infty}_{\rm per}(\Gamma) \hookrightarrow \mathcal{M}_{\rm per}(\Gamma). \end{aligned}$$

Proof. We only prove the result for $d = 3$; the other two embeddings are obtained by similar arguments. Let us first recall that the Lorentz space $L^{3,\infty}(\Gamma)$ is a L^2-multiplier of $L^{6,2}(\Gamma)$ (this can be seen by combining results on convolution multiplier spaces [8] and continuity properties of the Fourier transform on Lorentz spaces [15]), in the sense that

$$\exists C_1 \in \mathbb{R}_+ \mid \forall f \in L^{3,\infty}(\Gamma),\ \forall g \in L^{6,2}(\Gamma),\ \|fg\|_{L^2(\Gamma)} \le C_1 \|f\|_{L^{3,\infty}(\Gamma)} \|g\|_{L^{6,2}(\Gamma)}.$$

Besides, the embedding of $H^1(\Gamma)$ into $L^{6,2}(\Gamma)$ is continuous (see [4] for instance)

$$\exists C_2 \in \mathbb{R}_+ \mid \forall g \in H^1(\Gamma),\ \|g\|_{L^{6,2}(\Gamma)} \le C_2 \|g\|_{H^1(\Gamma)}. \tag{3.17}$$

Let $v \in L^{3,\infty}_{\rm per}(\Gamma)$. Denoting by $\mathcal{I}_L := \mathcal{R} \cap (-Lb/2, Lb/2]^3$, we have, for all $w \in H^1_{\rm per}(\Gamma_L)$,

$$\begin{aligned}
\|vw\|^2_{L^2_{\rm per}(\Gamma_L)} &= \int_{\Gamma_L} |vw|^2 = \sum_{R\in\mathcal{I}_L} \int_{\Gamma+R} |v(x)w(x)|^2\,dx \\
&= \sum_{R\in\mathcal{I}_L} \int_{\Gamma} |v(x)w(x+R)|^2\,dx = \sum_{R\in\mathcal{I}_L} \|vw(.+R)\|^2_{L^2(\Gamma)} \\
&\le C_1^2 \sum_{R\in\mathcal{I}_L} \|v\|^2_{L^{3,\infty}(\Gamma)} \|w(.+R)\|^2_{L^{6,2}(\Gamma)} \\
&\le C_1^2 \|v\|^2_{L^{3,\infty}(\Gamma)} \sum_{R\in\mathcal{I}_L} \|w(.+R)\|^2_{L^{6,2}(\Gamma)} \\
&\le C_1^2 C_2^2 \|v\|^2_{L^{3,\infty}(\Gamma)} \sum_{R\in\mathcal{I}_L} \|w(.+R)\|^2_{H^1(\Gamma)} \\
&\le C_1^2 C_2^2 \|v\|^2_{L^{3,\infty}(\Gamma)} \sum_{R\in\mathcal{I}_L} \int_{\Gamma} \left(|w(x+R)|^2 + |\nabla w(x+R)|^2\right) dx \\
&\le C_1^2 C_2^2 \|v\|^2_{L^{3,\infty}(\Gamma)} \int_{\Gamma_L} \left(|w(x)|^2 + |\nabla w(x)|^2\right) dx \\
&\le C_1^2 C_2^2 \|v\|^2_{L^{3,\infty}(\Gamma)} \|w\|^2_{H^1_{\rm per}(\Gamma_L)}.
\end{aligned}$$

Therefore, $v \in \mathcal{M}_{\rm per}(\Gamma)$ and $\|v\|_{\mathcal{M}_{\rm per}(\Gamma)} \le C_1 C_2 \|v\|_{L^{3,\infty}(\Gamma)}$. □

Remark 3.3.1. *In dimension 3, the $\mathcal{R}$-periodic Coulomb kernel G_1 defined by*

$$-\Delta G_1 = 4\pi \left(\sum_{R\in\mathcal{R}} \delta_R - |\Gamma|^{-1} \right), \quad \min_{x\in\mathbb{R}^3} G_1(x) = 0,$$

is in $L^{3,\infty}_{\rm per}(\Gamma)$, hence in $\mathcal{M}_{\rm per}(\Gamma)$. The functional setting we have introduced therefore allows us to deal with the electronic structure of crystals containing point-like nuclei.

Theorem 3.3.1. *Assume that $V_{\rm per} \in \mathcal{M}_{\rm per}(\Gamma)$. Then*

$$\lim_{N,L\to\infty \,|\, N/L\to\infty} \sigma(H_{L,N}) = \sigma(H).$$

Proof.

Step 1. Let us first establish that

$$\sigma(H) \subset \liminf_{N,L\to\infty \,|\, N/L\to\infty} \sigma(H_{L,N}).$$

Let $\lambda \in \sigma(H)$ and $(N_L)_{L\in\mathbb{N}^*}$ be a sequence of integers such that $\dfrac{N_L}{L} \underset{L\to\infty}{\longrightarrow} \infty$. Let $\epsilon > 0$ and $\psi \in C^\infty_{\rm c}(\mathbb{R}^d)$ be such that $\|\psi\|_{L^2} = 1$ and $\|(H-\lambda)\psi\|_{L^2} \le \epsilon$. We denote by ψ_L the $L\mathcal{R}$-periodic extension of $\psi|_{\Gamma_L}$. Since ψ is compactly supported, there exists $L_0 \in \mathbb{N}^*$ such that for all $L \ge L_0$, ${\rm Supp}(\psi) \subset \Gamma_L$. Consequently, for all $L \ge L_0$,

$$\|\psi_L\|_{L^2_{\rm per}(\Gamma_L)} = 1 \quad \text{and} \quad \|(H_L - \lambda)\psi_L\|_{L^2_{\rm per}(\Gamma_L)} \le \epsilon.$$

Let $\psi_{L,N_L} := \Pi_{X_{L,N_L}} \psi_L$. We are going to prove that

$$\|(H_L - \lambda)\psi_L - (H_{L,N_L} - \lambda)\,\psi_{L,N_L}\|_{L^2_{\rm per}(\Gamma_L)} \underset{L\to\infty}{\longrightarrow} 0. \tag{3.18}$$

First, we infer from (3.16) and the density of $H^1_0(\Omega)$ in $L^2(\Omega)$ for any bounded domain Ω of $\mathbb{R}^d$, that

$$\forall \phi \in L^2_{\rm c}(\mathbb{R}^d), \quad \|(1 - \Pi_{X_{L,N_L}})\phi_L\|_{L^2_{\rm per}(\Gamma_L)} \underset{L\to\infty}{\longrightarrow} 0,$$

where $L^2_{\rm c}(\mathbb{R}^d)$ denotes the space of the square integrable functions on $\mathbb{R}^d$ with compact supports, and where ϕ_L is the $L\mathcal{R}$-periodic extension of $\phi|_{\Gamma_L}$. As ψ, $\Delta\psi$, $V_{\rm per}\psi$ and $W\psi$ are square integrable, with compact supports, we therefore have for all $L \geq L_0$,

$$\begin{aligned}
&\|\psi_L - \psi_{L,N_L}\|_{L^2_{\rm per}(\Gamma_L)} = \left\|\left(1 - \Pi_{X_{L,N_L}}\right)\psi_L\right\|_{L^2_{\rm per}(\Gamma_L)} \underset{L\to\infty}{\longrightarrow} 0,\\
&\|-\Delta\psi_L + \Delta\psi_{L,N_L}\|_{L^2_{\rm per}(\Gamma_L)} = \left\|\left(1 - \Pi_{X_{L,N_L}}\right)(-\Delta\psi)_L\right\|_{L^2_{\rm per}(\Gamma_L)} \underset{L\to\infty}{\longrightarrow} 0,\\
&\|W_L\psi_L - \Pi_{X_{L,N_L}}(W_L\psi_L)\|_{L^2_{\rm per}(\Gamma_L)} = \left\|\left(1 - \Pi_{X_{L,N_L}}\right)(W\psi)_L\right\|_{L^2_{\rm per}(\Gamma_L)} \underset{L\to\infty}{\longrightarrow} 0,\\
&\|V_{\rm per}\psi_L - \Pi_{X_{L,N_L}}(V_{\rm per}\psi_L)\|_{L^2_{\rm per}(\Gamma_L)} = \left\|\left(1 - \Pi_{X_{L,N_L}}\right)(V_{\rm per}\psi)_L\right\|_{L^2_{\rm per}(\Gamma_L)} \underset{L\to\infty}{\longrightarrow} 0.
\end{aligned}$$

We infer from the last two convergence results that, on the one hand,

$$\begin{aligned}
&\|W_L\psi_L - \Pi_{X_{L,N_L}}(W_L\psi_{L,N_L})\|_{L^2_{\rm per}(\Gamma_L)}\\
&\quad\leq \left\|W_L\psi_L - \Pi_{X_{L,N_L}}(W_L\psi_L)\right\|_{L^2_{\rm per}(\Gamma_L)} + \left\|\Pi_{X_{L,N_L}}\left(W_L(\psi_L - \psi_{L,N_L})\right)\right\|_{L^2_{\rm per}(\Gamma_L)}\\
&\quad\leq \left\|W_L\psi_L - \Pi_{X_{L,N_L}}(W_L\psi_L)\right\|_{L^2_{\rm per}(\Gamma_L)} + \|W\|_{L^\infty}\,\|\psi_L - \psi_{L,N_L}\|_{L^2_{\rm per}(\Gamma_L)}\\
&\quad\underset{L\to\infty}{\longrightarrow} 0,
\end{aligned}$$

and that, on the other hand,

$$\begin{aligned}
&\|V_{\rm per}\psi_L - \Pi_{X_{L,N_L}}(V_{\rm per}\psi_{L,N_L})\|_{L^2_{\rm per}(\Gamma_L)}\\
&\quad\leq \left\|V_{\rm per}\psi_L - \Pi_{X_{L,N_L}}(V_{\rm per}\psi_L)\right\|_{L^2_{\rm per}(\Gamma_L)} + \left\|\Pi_{X_{L,N_L}}\left(V_{\rm per}(\psi_L - \psi_{L,N_L})\right)\right\|_{L^2_{\rm per}(\Gamma_L)}\\
&\quad\leq \left\|\left(1 - \Pi_{X_{L,N_L}}\right)V_{\rm per}\psi_L\right\|_{L^2_{\rm per}(\Gamma_L)} + \|V_{\rm per}\|_{\mathcal{M}_{\rm per}(\Gamma)}\|\psi_L - \psi_{L,N_L}\|_{H^1_{\rm per}(\Gamma_L)}\\
&\quad\underset{L\to\infty}{\longrightarrow} 0.
\end{aligned}$$

Collecting the above results, we obtain (3.18). Thus, for L large enough,

$$\|(H_{L,N_L} - \lambda)\psi_{L,N_L}\|_{L^2_{\rm per}(\Gamma_L)} \leq 2\varepsilon.$$

As $\|\psi_{L,N_L}\|_{L^2_{\rm per}(\Gamma_L)} = 1$ for all $L \geq L_0$, we infer that for L large enough, $\mathrm{dist}(\lambda, \sigma(H_{L,N_L})) \leq 2\epsilon$, so that $\lambda \in \liminf_{L\to\infty} \sigma(H_{L,N_L})$.

Step 2. Let us now prove that

$$\limsup_{N,L\to\infty \,|\, N/L\to\infty} \sigma(H_{N,L}) \subset \sigma(H).$$

We argue by contradiction, assuming that there exists $\lambda \in \mathbb{R} \setminus \sigma(H)$ and a sequence $(L_k, N_k)_{k\in\mathbb{N}}$ with $L_k \underset{k\to\infty}{\to} \infty$, $N_k \underset{k\to\infty}{\to} \infty$, $N_k/L_k \underset{k\to\infty}{\to} \infty$, such that for each k, there exists $(\psi_{L_k,N_k}, \lambda_{L_k,N_k}) \in X_{L_k,N_k} \times \mathbb{R}$ satisfying

$$\left\{ \begin{array}{l} \forall \phi_{L_k,N_k} \in X_{L_k,N_k},\ a_{L_k}(\psi_{L_k,N_k}, \phi_{L_k,N_k}) = \lambda_{L_k,N_k} \langle \psi_{L_k,N_k}, \phi_{L_k,N_k} \rangle_{L^2_{\rm per}(\Gamma_{L_k})} \\ \|\psi_{L_k,N_k}\|_{L^2(\Gamma_{L_k})} = 1, \end{array} \right.$$

and $\lim\limits_{k\to\infty} \lambda_{L_k,N_k} = \lambda$. Each function ψ_{L_k,N_k} is then solution to the PDE

$$-\frac{1}{2}\Delta\psi_{L_k,N_k} + \Pi_{X_{L_k,N_k}}\left((V_{\rm per} + W_{L_k})\psi_{L_k,N_k}\right) = \lambda_{L_k,N_k}\psi_{L_k,N_k}. \tag{3.19}$$

Reasoning as in the proof of Proposition 3.2.1, it can be checked that the sequence $(\|\psi_{L_k,N_k}\|_{H^1_{\rm per}(\Gamma_{L_k})})_{k\in\mathbb{N}}$ is bounded, and that

$$\psi_{L_k,N_k} \underset{k\to\infty}{\longrightarrow} 0 \quad \text{in } L^2_{\rm loc}(\mathbb{R}^d). \tag{3.20}$$

For all k, we consider a cut-off function $\chi_k \in C^\infty_{\rm c}(\mathbb{R}^d)$ such that $0 \le \chi_k \le 1$ on $\mathbb{R}^d$, $\chi_k \equiv 1$ on Γ_{L_k}, ${\rm Supp}(\chi_k) \subset (L_k + L_k^{1/2})\Gamma$, $\|\nabla\chi_k\|_{L^\infty} \le CL_k^{-1/2}$, and $\|\Delta\chi_k\|_{L^\infty} \le CL_k^{-1}$ for some constant $C \in \mathbb{R}_+$ independent of k. We then set $\widetilde{\psi}_k = \chi_k\psi_{L_k,N_k}$. It holds $\widetilde{\psi}_k \in H^2(\mathbb{R}^d)$, $1 \le \|\widetilde{\psi}_k\|_{L^2} \le 2^{d/2}$ and

$$\begin{array}{rcl} -\frac{1}{2}\Delta\widetilde{\psi}_k + V_{\rm per}\widetilde{\psi}_k - \lambda\widetilde{\psi}_k & = & \chi_k\left(V_{\rm per}\psi_{L_k,N_k} - \Pi_{X_{L_k,N_k}}\left(V_{\rm per}\psi_{L_k,N_k}\right)\right) \\ & & -\chi_k\Pi_{X_{L_k,N_k}}\left(W_{L_k}\psi_{L_k,N_k}\right) - \nabla\chi_k\cdot\nabla\psi_{L_k,N_k} \\ & & -\frac{1}{2}\Delta\chi_k\psi_{L_k,N_k} + (\lambda_{L_k,N_k} - \lambda)\widetilde{\psi}_k. \end{array} \tag{3.21}$$

As $(\lambda_{L_k,N_k})_{k\in\mathbb{N}}$ converges to λ in $\mathbb{R}$ and $\|\widetilde{\psi}_k\|_{L^2} \le 2^{d/2}$, we have

$$(\lambda_{L_k,N_k} - \lambda)\widetilde{\psi}_k \underset{k\to\infty}{\longrightarrow} 0 \quad \text{strongly in } L^2(\mathbb{R}^d).$$

Using the facts that ${\rm Supp}(\chi_k) \subset 2\Gamma_{L_k}$, $\|\nabla\chi_k\|_{L^\infty} \le CL_k^{-1/2}$ and $\|\Delta\chi_k\|_{L^\infty} \le CL_k^{-1}$ for a constant $C \in \mathbb{R}_+$ independent of k, and the boundedness of the sequence $(\|\psi_{L_k,N_k}\|_{H^1_{\rm per}(\Gamma_{L_k})})_{k\in\mathbb{N}}$, we get

$$-\nabla\chi_k\cdot\nabla\psi_{L_k,N_k} - \frac{1}{2}\Delta\chi_k\psi_{L_k,N_k} \underset{k\to\infty}{\longrightarrow} 0 \quad \text{strongly in } L^2(\mathbb{R}^d).$$

It also follows from (3.20) that the sequence $\|W_{L_k}\psi_{L_k,N_k}\|_{L^2_{\rm per}(\Gamma_{L_k})}$ goes to zero, leading to

$$\chi_k\Pi_{X_{L_k,N_k}}\left(W_{L_k}\psi_{L_k,N_k}\right) \underset{k\to\infty}{\longrightarrow} 0 \quad \text{strongly in } L^2(\mathbb{R}^d).$$

Lastly,

$$\chi_k\left(V_{\rm per}\psi_{L_k,N_k} - \Pi_{X_{L_k,N_k}}\left(V_{\rm per}\psi_{L_k,N_k}\right)\right) \underset{k\to\infty}{\longrightarrow} 0 \quad \text{strongly in } L^2(\mathbb{R}^d). \tag{3.22}$$

To show the above convergence result, we consider $\epsilon > 0$ and, using the density of e.g. $W^{1,\infty}_{\rm per}(\Gamma) := \{W_{\rm per} \in L^\infty_{\rm per}(\Gamma) \mid \nabla W_{\rm per} \in L^\infty_{\rm per}(\Gamma)\}$ in $\mathcal{M}_{\rm per}(\Gamma)$, we can choose some

$\widetilde{V}_{\rm per} \in W^{1,\infty}_{\rm per}(\Gamma)$ such that $\|V_{\rm per} - \widetilde{V}_{\rm per}\|_{\mathcal{M}_{\rm per}(\Gamma)} \leq \varepsilon$. We then deduce from (3.16) that, for all $k \in \mathbb{N}$,

$$\begin{aligned}
&\left\|V_{\rm per}\psi_{L_k,N_k} - \Pi_{X_{L_k,N_k}}\left(V_{\rm per}\psi_{L_k,N_k}\right)\right\|_{L^2_{\rm per}(\Gamma_{L_k})} \\
&\leq \left\|(V_{\rm per} - \widetilde{V}_{\rm per})\psi_{L_k,N_k}\right\|_{L^2_{\rm per}(\Gamma_{L_k})} + \left\|\widetilde{V}_{\rm per}\psi_{L_k,N_k} - \Pi_{X_{L_k,N_k}}\left(\widetilde{V}_{\rm per}\psi_{L_k,N_k}\right)\right\|_{L^2_{\rm per}(\Gamma_{L_k})} \\
&\leq \|V_{\rm per} - \widetilde{V}_{\rm per}\|_{\mathcal{M}_{\rm per}(\Gamma)}\|\psi_{L_k,N_k}\|_{H^1_{\rm per}(\Gamma_{L_k})} + \frac{L_k}{N_k}\|\widetilde{V}_{\rm per}\psi_{L_k,n_k}\|_{H^1_{\rm per}(\Gamma_{L_k})} \\
&\leq \varepsilon\|\psi_{L_k,N_k}\|_{H^1_{\rm per}(\Gamma_{L_k})} + \frac{L_k}{N_k}\|\psi_{L_k,n_k}\|_{H^1_{\rm per}(\Gamma_{L_k})}(\|\widetilde{V}_{\rm per}\|_{L^\infty} + \|\nabla\widetilde{V}_{\rm per}\|_{L^\infty}).
\end{aligned}$$

Since the sequence $\left(\|\psi_{L_k,N_k}\|_{H^1_{\rm per}(\Gamma_{L_k})}\right)_{k\in\mathbb{N}^*}$ is bounded, this yields

$$\left\|V_{\rm per}\psi_{L_k,N_k} - \Pi_{X_{L_k,N_k}}\left(V_{\rm per}\psi_{L_k,N_k}\right)\right\|_{L^2_{\rm per}(\Gamma_{L_k})} \underset{k\to\infty}{\longrightarrow} 0,$$

which implies (3.22).

Collecting the above convergence results, we obtain that the right-hand side of (3.21) goes to zero strongly in $L^2(\mathbb{R}^d)$. Therefore, $(\widetilde{\psi}_k/\|\widetilde{\psi}_k\|_{L^2})_{k\in\mathbb{N}}$ is a Weyl sequence for λ, which contradicts the fact that $\lambda \notin \sigma(H^0_{\rm per})$. □

A similar result was proved in [172] for compactly supported defects in 2D photonic crystals, with $V_{\rm per} \in L^\infty(\mathbb{R}^2)$ and $N = \infty$. In [41], we prove that the error made on the eigenvalues and the associated eigenvectors decays exponentially with respect to the size of the supercell. We did not consider here the error due to numerical integration. The numerical analysis of the latter will be presented in Chapter 4 of this thesis work.

Note that, if instead of supercells of the form $\Gamma_L = L\Gamma$, $L \in \mathbb{N}^*$, we had used computational domains of the form $\Gamma_{L+t} = (L+t)\Gamma$, $t \in (0,1)$, we would have observed spectral pollution. As in the case studied in the previous section, the spurious eigenvectors concentrate on the boundary $\partial\Gamma_{L+t}$. In the one-dimensional setting ($\mathcal{R} = b\mathbb{Z}$), and for a fixed value of t, the translated spurious modes weakly converge in $H^1(\mathbb{R})$, when L goes to infinity, to the eigenmodes of the dislocation operator $H(t) = -\dfrac{d^2}{dx^2} + 1_{x<0}V_{\rm per}(x + tb/2) + 1_{x>0}V_{\rm per}(x - tb/2)$ studied in [122]. We refer to Section 3.5.2 of the Appendix 3.5 for more details and a numerical illustration of this phenomenon.

3.4 A no-pollution criterion

Spectral pollution can be avoided by using e.g. the quadratic projection method, introduced in an abstract setting in [167], and applied to the case of perturbed periodic Schrödinger operators in [24]. An alternative way to prevent spectral pollution is to impose constraints on the approximation spaces $(X_n)_{n\in\mathbb{N}}$. Consider a gap $(\alpha,\beta) \subset \mathbb{R} \setminus \sigma(H^0_{\rm per})$ in the spectrum of $H^0_{\rm per}$ and denote by $P = \chi_{(-\infty,\gamma]}(H^0_{\rm per})$ where $\gamma = \frac{\alpha+\beta}{2}$ and where $\chi_{(-\infty,\gamma]}$ is the characteristic function of the interval $(-\infty,\gamma]$.

Theorem 3.4.1. *Let $(P_n)_{n\in\mathbb{N}}$ be a sequence of linear projectors on $L^2(\mathbb{R}^d)$ such that for all $n \in \mathbb{N}$, $Ran(P_n) \subset H^1(\mathbb{R}^d)$, and $\sup_{n\in\mathbb{N}}\|P_n\|_{\mathcal{L}(L^2)} < \infty$, and $(X_n)_{n\in\mathbb{N}}$ a sequence of finite dimensional discretization spaces satisfying (3.3) as well as the following two properties:*

(A1) $\forall n \in \mathbb{N},\ X_n = X_n^+ \oplus X_n^-$ *with* $X_n^- \subset \mathrm{Ran}(P_n)$ *and* $X_n^+ \subset \mathrm{Ran}(1-P_n)$*;*

(A2) $\displaystyle\sup_{\phi_n \in X_n \setminus \{0\}} \frac{\|(P-P_n)\phi_n\|_{H^1(\mathbb{R}^d)}}{\|\phi_n\|_{H^1(\mathbb{R}^d)}} \underset{n\to\infty}{\longrightarrow} 0.$

Then,

$$\lim_{n\to\infty} \sigma(H|_{X_n}) \cap (\alpha,\beta) = \sigma(H) \cap (\alpha,\beta).$$

The above result is an extension, for the specific case of perturbed periodic Schrödinger operators, to the results in [134, Theorem 2.6] in the sense that (i) the exact spectral projector P is replaced by an approximate projector P_n, and (ii) the discretization space X_n may consist of functions of $H^1(\mathbb{R}^d)$ (the form domain of H), while in [134], the basis functions are assumed to belong to $H^2(\mathbb{R}^d)$ (the domain of H).

Proof. From (3.4), we already know that $\sigma(H) \cap (\alpha,\beta) \subset \liminf_{n\to\infty} \sigma(H|_{X_n}) \cap (\alpha,\beta)$. Conversely, let $\lambda \in (\limsup_{n\to\infty} \sigma(H|_{X_n}) \cap (\alpha,\beta)) \setminus \sigma(H)$, and $(\psi_{n_k})_{k\in\mathbb{N}}$ be a sequence of functions of $H^1(\mathbb{R}^d)$ such that for all $k \in \mathbb{N}$, $\psi_{n_k} \in X_{n_k}$, $\|\psi_{n_k}\|_{L^2(\mathbb{R}^d)} = 1$ and $(H|_{X_{n_k}} - \lambda)\psi_{n_k} \longrightarrow_{k\to\infty} 0$ strongly in $L^2(\mathbb{R}^d)$. Reasoning as in the proof of Proposition 3.2.1, we obtain that the sequence $(\psi_{n_k})_{k\in\mathbb{N}}$ converges to 0, weakly in $H^1(\mathbb{R}^d)$, and strongly in $L^2_{\mathrm{loc}}(\mathbb{R}^d)$. Let us then expand ψ_{n_k} as $\psi_{n_k} = \psi_{n_k}^+ + \psi_{n_k}^-$ with $\psi_{n_k}^+ := (1-P_{n_k})\psi_{n_k} \in X_{n_k}^+$ and $\psi_{n_k}^- := P_{n_k}\psi_{n_k} \in X_{n_k}^-$ and notice that

$$(a^0-\lambda)(\psi_{n_k}^+,\psi_{n_k}^+) + (a^0-\lambda)(\psi_{n_k}^-,\psi_{n_k}^+) = (a-\lambda)(\psi_{n_k},\psi_{n_k}^+) - \int_{\mathbb{R}^d} W\psi_{n_k}\psi_{n_k}^+.$$

Since $\psi_{n_k}^+ = (1-P_n)\psi_{n_k} \in X_{n_k}$,

$$\begin{aligned} \left|(a-\lambda)(\psi_{n_k},\psi_{n_k}^+)\right| &= \left|\langle (H|_{X_{n_k}} - \lambda)\psi_{n_k}, (1-P_n)\psi_{n_k}\rangle_{L^2}\right| \\ &\le \left(1 + \sup_{k\in\mathbb{N}} \|P_{n_k}\|_{\mathcal{L}(L^2)}\right) \|(H|_{X_{n_k}} - \lambda)\psi_{n_k}\|_{L^2} \underset{k\to\infty}{\longrightarrow} 0. \end{aligned}$$

Besides, as W vanishes at infinity, $(\psi_{n_k})_{k\in\mathbb{N}}$ converges to 0 in $L^2_{\mathrm{loc}}(\mathbb{R}^d)$ and $\sup_{k\in\mathbb{N}} \|\psi_{n_k}^+\|_{L^2} \le 1 + \sup_{k\in\mathbb{N}} \|P_{n_k}\|_{\mathcal{L}(L^2)} < \infty$, we also have

$$\int_{\mathbb{R}^d} W\psi_{n_k}\psi_{n_k}^+ \underset{k\to\infty}{\longrightarrow} 0.$$

Therefore,

$$(a^0-\lambda)(\psi_{n_k}^+,\psi_{n_k}^+) + (a^0-\lambda)(\psi_{n_k}^-,\psi_{n_k}^+) \underset{k\to\infty}{\longrightarrow} 0.$$

Likewise,

$$(a^0-\lambda)(\psi_{n_k}^+,\psi_{n_k}^-) + (a^0-\lambda)(\psi_{n_k}^-,\psi_{n_k}^-) = (a-\lambda)(\psi_{n_k},\psi_{n_k}^-) - \int_{\mathbb{R}^d} W\psi_{n_k}\psi_{n_k}^- \underset{k\to\infty}{\longrightarrow} 0.$$

Substracting the second equation from the first one, we obtain

$$(a^0-\lambda)(\psi_{n_k}^+,\psi_{n_k}^+) - (a^0-\lambda)(\psi_{n_k}^-,\psi_{n_k}^-) \underset{k\to\infty}{\longrightarrow} 0.$$

Now, we notice that

$$\begin{aligned} (a^0-\lambda)(\psi_{n_k}^-,\psi_{n_k}^-) &= (a^0-\lambda)(P_{n_k}\psi_{n_k}, P_{n_k}\psi_{n_k}) \\ &= (a^0-\lambda)(P\psi_{n_k}, P\psi_{n_k}) + 2(a^0-\lambda)(P\psi_{n_k}, (P_{n_k}-P)\psi_{n_k}) \\ &\quad +(a^0-\lambda)((P_{n_k}-P)\psi_{n_k}, (P_{n_k}-P)\psi_{n_k}), \end{aligned}$$

and

$$\begin{aligned}(a^0-\lambda)(\psi_{n_k}^+,\psi_{n_k}^+) &= (a^0-\lambda)((1-P_{n_k})\psi_{n_k},(1-P_{n_k})\psi_{n_k})\\ &= (a^0-\lambda)((1-P)\psi_{n_k},(1-P)\psi_{n_k})\\ &\quad +2(a^0-\lambda)((1-P)\psi_{n_k},(P-P_{n_k})\psi_{n_k})\\ &\quad +(a^0-\lambda)((P-P_{n_k})\psi_{n_k},(P-P_{n_k})\psi_{n_k}).\end{aligned}$$

Besides, there exists $\eta_+,\eta_->0$ such that for all $\psi\in H^1(\mathbb{R}^d)$,

$$\begin{aligned}(a^0-\lambda)((1-P)\psi,(1-P)\psi) &\geq \eta_+\|(1-P)\psi\|^2_{L^2(\mathbb{R}^d)},\\ -(a^0-\lambda)(P\psi,P\psi) &\geq \eta_-\|P\psi\|^2_{L^2(\mathbb{R}^d)}.\end{aligned}$$

Thus,

$$\begin{aligned}(a^0-\lambda)(\psi_{n_k}^+,\psi_{n_k}^+)-(a^0-\lambda)(\psi_{n_k}^-,\psi_{n_k}^-) &\geq \min(\eta_+,\eta_-)\|\psi_{n_k}\|^2_{L^2(\mathbb{R}^d)}\\ &\quad +2(a^0-\lambda)(\psi_{n_k},(P-P_{n_k})\psi_{n_k}).\end{aligned}$$

From assumption $(A2)$ and the boundedness of $(\psi_{n_k})_{k\in\mathbb{N}}$ in $H^1(\mathbb{R}^d)$, we deduce that

$$(a^0-\lambda)(\psi_{n_k},(P-P_{n_k})\psi_{n_k})\underset{k\to\infty}{\longrightarrow}0,$$

which imply that $\|\psi_{n_k}\|_{L^2}\underset{k\to\infty}{\longrightarrow}0$. This contradicts the fact that $\|\psi_{n_k}\|_{L^2}=1$ for all $k\in\mathbb{N}$. □

The assumptions made in Theorem 3.4.1 allow in particular to consider approximation spaces built from approximate spectral projectors of $H^0_{\rm per}$. As a matter of illustration, let us consider the case when the approximate spectral projectors are constructed by means of the finite element method. As in Section 3.2, we consider a sequence $(\mathcal{T}_n^\infty)_{n\in\mathbb{N}}$ of uniformly regular meshes of $\mathbb{R}^d$, invariant with respect to the translations of the lattice $\mathcal{R}$, and such that $h_n:=\max_{K\in\mathcal{T}_n^\infty}\mathrm{diam}(K)\longrightarrow_{n\to\infty}0$, and denote by X_n^∞ the infinite dimensional closed vector subspace of $H^1(\mathbb{R}^d)$ built from $(\mathcal{T}_n^\infty)_{n\in\mathbb{N}}$ and $\mathbb{P}_m$-finite elements. Assume that we want to compute the eigenvalues of $H=H^0_{\rm per}+W$ located inside the gap (α,β) between the $J^{\rm th}$ and $(J+1)^{\rm st}$ bands of $H^0_{\rm per}$. Using Bloch theory [163], we obtain

$$P=\chi_{(-\infty,\gamma]}(H^0_{\rm per})=\fint_{\Gamma^*}P_q\,dq,$$

where P_q is the rank-J orthogonal projector on $L^2_q(\Gamma)$ defined by

$$P_q=\sum_{j=1}^J|e_{j,q}\rangle\langle e_{j,q}|,$$

where $(\epsilon_{j,q},e_{j,q})_{j\in\mathbb{N}^*}$, $\epsilon_{1,q}\leq\epsilon_{2,q}\leq\cdots$, is an $L^2_q(\Gamma)$-orthonormal basis of eigenmodes of the quadratic form a^0_q defined by (3.14). For n large enough, we introduce

$$P_n:=\fint_{\Gamma^*}\sum_{j=1}^J|e_{n,j,q}\rangle\langle e_{n,j,q}|\,dq, \tag{3.23}$$

where $(\epsilon_{n,j,q},e_{n,j,q})_{1\leq j\leq N_n}$, $\epsilon_{n,1,q}\leq\epsilon_{n,2,q}\leq\cdots\leq\epsilon_{n,N_n,q}$, is the $L^2_q(\Gamma)$-orthonormal basis of eigenmodes of a^0_q in $C^0(\mathcal{T}_n^\infty)\cap L^2_q(\Gamma)$ already introduced in the proof of Proposition 3.2.1.

We have seen in Section 3.2 that using approximation spaces of the form

$$X_n = \{\psi_n \in X_n^\infty \mid \mathrm{Supp}(\psi_n) \subset \Omega_n\},$$

where $(\Omega_n)_{n\in\mathbb{N}}$ is an increasing sequence of closed convex sets of $\mathbb{R}^d$ converging to $\mathbb{R}^d$, leads, in general, to spectral pollution. We now consider the approximation spaces

$$\widetilde{X}_n = X_n^+ \oplus X_n^- \quad \text{where} \quad X_n^- = P_n X_n \quad \text{and} \quad X_n^+ = (1-P_n)X_n. \tag{3.24}$$

Note that $\widetilde{X}_n = X_n + P_n X_n$, so that $\widetilde{X}_n$ can be seen as an augmentation of X_n.

Corollary 3.4.1. *The sequence of approximation spaces $(\widetilde{X}_n)_{n\in\mathbb{N}}$ defined by (3.24) satisfies (3.3) and it holds*

$$\lim_{n\to\infty} \sigma(H|_{\widetilde{X}_n}) \cap (\alpha,\beta) = \sigma(H) \cap (\alpha,\beta). \tag{3.25}$$

Proof. As $\widetilde{X}_n = X_n + P_n X_n$ with $(X_n)_{n\in\mathbb{N}}$ satisfying (3.3), it is clear that $(\widetilde{X}_n)_{n\in\mathbb{N}}$ satisfies (3.3). The sequence $(P_n)_{n\in\mathbb{N}}$ is a sequence of orthogonal projectors of $L^2(\mathbb{R}^d)$ such that $\mathrm{Ran}(P_n) \subset X_n^\infty \subset H^1(\mathbb{R}^d)$. Besides, $\|P_n\|_{\mathcal{L}(L^2)} = 1$ since the projector P_n is orthogonal. It follows from the minmax principle [163] and usual a priori error estimates for linear elliptic eigenvalue problems [9] that

$$\sup_{1\le j\le J,\, q\in\Gamma^*} \epsilon_{n,j,q} \underset{n\to\infty}{\longrightarrow} \alpha \qquad \text{and} \qquad \inf_{j\ge J+1,\, q\in\Gamma^*} \epsilon_{n,j,q} \underset{n\to\infty}{\longrightarrow} \beta,$$

and that there exists $C \in \mathbb{R}_+$ such that

$$\|P_n - P\|_{\mathcal{L}(H^1)} \le C \sup_{q\in\Gamma^*} \sup_{\substack{v_q \in \mathrm{Ran}(P_q) \\ \|v_q\|_{L^2_q(\Gamma)} = 1}} \inf_{v_q^n \in C^0(\mathcal{T}_n^\infty)\cap L^2_q(\Gamma)} \|v_q - v_q^n\|_{H^1_q(\Gamma)} \underset{n\to\infty}{\longrightarrow} 0.$$

We conclude using Theorem 3.4.1. □

Let us finally present some numerical simulations illustrating Corollary 3.4.1 in a one-dimensional setting, with $V_{\mathrm{per}}(x) = \cos(x) + 3\sin(2x+1)$ and $W(x) = -(x+2)^2 e^{-x^2}$. We focus on the spectral gap (α,β) located between the first and second bands of $H^0_{\mathrm{per}} = -\frac{d^2}{dx^2} + V_{\mathrm{per}}$ (corresponding to $J=1$). Numerical simulations done with the pollution-free supercell model show that $\alpha \simeq -1.15$ and $\beta \simeq -0.65$, and that H has exactly two discrete eigenvalues $\lambda_1 \simeq -1.04$ and $\lambda_2 \simeq -0.66$ in the gap (α,β).

The simulations below have been performed with a uniform mesh of $\mathbb{R}$ centered on 0, consisting of segments of length $h = \pi/50$, and with $\Omega = [-L, L]$, for different values of L. The sums over $\mathcal{R}$ have been truncated using very large cut-offs; likewise, the integrals on the Brillouin zone have been computed numerically on a very fine uniform integration grid, in order to eliminate the so-called k-point discretization errors. The numerical analysis of the approximations resulting from the truncation of the sums over $\mathcal{R}$ and from the numerical integration on Γ^*, is work in progress.

The spectra of the operators $H|_{X_n}$ (standard finite element discretization spaces) and $H|_{\widetilde{X}_n}$ (augmented finite element discretization spaces defined by (3.24)) are displayed in Fig. 3.3. The variational approximation of H in X_n is seen to generate spectral pollution, while, in agreement with Corollary 3.4.1, no spectral pollution is observed with the discretization spaces $\widetilde{X}_n$.

Acknowledgements

We thank François Murat for helpful discussions.

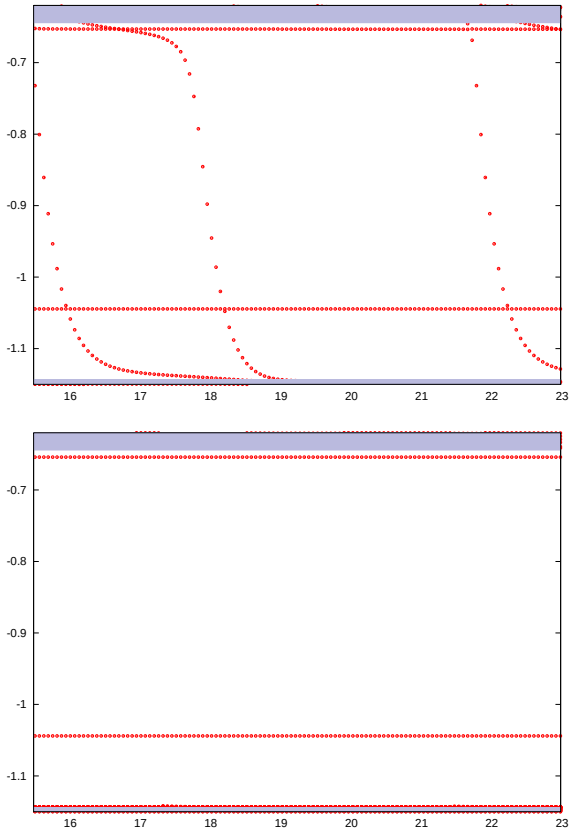

Figure 3.3: The spectra of the variational approximations of H for various sizes of the simulation domain, obtained with standard finite element discretization spaces X_n (top) and with augmented finite element discretization spaces $\widetilde{X}_n$ defined by (3.24) (bottom).

3.5 Appendix

3.5.1 One-dimensional characterization of Galerkin spurious states

In this section, we prove the results announced in Remark 1.2.1.

Proposition 3.5.1. *Let $\mathcal{R} = b\mathbb{Z}$ with $b > 0$ and $t, t' \in [0,1)$. Let $H^+(t) := -\Delta + V_{\mathrm{per}}(\cdot - tb)$ (respectively $H^-(t') := -\Delta + V_{\mathrm{per}}(\cdot + t'b)$) be the operator on $L^2(\mathbb{R}_+)$ (respectively $L^2(\mathbb{R}_-)$) with domain $H_2(\mathbb{R}_+) \cap H_0^1(\mathbb{R}_+)$ (respectively $H^2(\mathbb{R}_-) \cap H_0^1(\mathbb{R}_-)$).*

Let $(\mathcal{T}_n^\infty)_{n\in\mathbb{N}}$ be a sequence of uniformly regular meshes of $\mathbb{R}$, invariant with respect to the translations of the lattice $\mathcal{R}$ and such that $h_n := \max_{K\in\mathcal{T}_n^\infty} \mathrm{diam}(K) \underset{n\to\infty}{\longrightarrow} 0$. Let $(p_n)_{n\in\mathbb{N}}$ and $(p'_n)_{n\in\mathbb{N}}$ be two increasing sequences of integers. For all $n \in \mathbb{N}$, let $\Omega_n = [-(p_n+t)b, (p'_n+t')b]$, $\mathcal{T}_n := \{K \in \mathcal{T}_n^\infty \,|K \subset \Omega_n\}$ and X_n the finite-dimensional approximation space of $H_0^1(\Omega_n) \hookrightarrow H^1(\mathbb{R})$ obtained with $\mathcal{T}_n$ and $\mathbb{P}_m$ finite elements ($m \in \mathbb{N}^*$).

Let $\lambda \in \mathrm{limsup}_{n\to\infty} \sigma(H|_{X_n}) \setminus \sigma(H)$. Then, $\lambda \in \sigma(H^+(t)) \cup \sigma(H^-(t'))$. Besides, let $(\psi_{n_k}, \lambda_{n_k}) \in X_{n_k} \times \mathbb{R}$ be such that $H|_{X_{n_k}}\psi_{n_k} = \lambda_{n_k}\psi_{n_k}$, $\|\psi_{n_k}\|_{L^2} = 1$ and

$\lim_{k\to\infty} \lambda_{n_k} = \lambda$. *For all* $k \in \mathbb{N}$, *let*

$$\phi^+_{n_k} := \psi_{n_k}(\cdot - (p_{n_k} + t)b) \text{ and } \phi^-_{n_k} := \psi_{n_k}(\cdot + (p'_{n_k} + t')b).$$

Then, up to the extraction of a subsequence (still denoted by $(\psi_{n_k})_{k\in\mathbb{N}}$*), the sequence* $(\phi^+_{n_k})_{k\in\mathbb{N}}$ *(respectively* $(\phi^-_{n_k})_{k\in\mathbb{N}}$*) converges to* ϕ^+ *(respectively* ϕ^-*) weakly in* $H^1(\mathbb{R})$ *and strongly in* $L^\infty_{\rm loc}(\mathbb{R})$ *where* $\phi^+|_{\mathbb{R}_+} \in \mathrm{Ker}(H^+(t) - \lambda)$ *and* $\phi^+|_{\mathbb{R}_-} = 0$ *(respectively* $\phi^-|_{\mathbb{R}_-} \in \mathrm{Ker}(H^-(t') - \lambda)$ *and* $\phi^-|_{\mathbb{R}_+} = 0$*). Besides, either* ϕ^+ *or* ϕ^- *is non identically zero.*

Proof. For all $n \in \mathbb{N}$, let

$$\mathcal{T}^+_n := \{K + (p_n + t)b \mid K \in \mathcal{T}^\infty_n,\ K \subset \Omega_n\} \text{ and } \mathcal{T}^-_n := \{K - (p'_n + t')b \mid K \in \mathcal{T}^\infty_n,\ K \subset \Omega_n\}$$

and X^+_n (respectively X^-_n) the finite-dimensional approximation space of $H^1_0(\Omega_n + (p_n + t)b) \hookrightarrow H^1_0(\mathbb{R}_+)$ (respectively of $H^1_0(\Omega_n - (p'_n + t')b) \hookrightarrow H^1_0(\mathbb{R}_-)$) obtained with $\mathcal{T}^+_n$ (respectively $\mathcal{T}^-_n$) and $\mathbb{P}_m$ finite elements. Let us also introduce the self-adjoint operator $H^+(t)|_{X^+_n} : X^+_n \to X^+_n$ defined by

$$\forall \xi_n, \kappa_n \in X^+_n,\ \langle H^+(t)|_{X^+_n}\xi_n, \kappa_n\rangle_{L^2(\mathbb{R}_+)} = \int_{\mathbb{R}_+} \nabla\xi_n \cdot \nabla\kappa_n + \int_{\mathbb{R}_+} V_{\rm per}(\cdot - tb)\xi_n\kappa_n.$$

We know from Proposition 3.2.1 that $(\psi_{n_k})_{k\in\mathbb{N}}$ is bounded in $H^1(\mathbb{R})$. Then, so are $(\phi^+_{n_k})_{k\in\mathbb{N}}$ and $(\phi^-_{n_k})_{k\in\mathbb{N}}$. Thus, up to the extraction of a subsequence (still denoted by $(\psi_{n_k})_{k\in\mathbb{N}}$ for the sake of simplicity), there exists $\phi^+, \phi^- \in H^1(\mathbb{R})$ such that $(\phi^+_{n_k})_{k\in\mathbb{N}}$ (respectively $(\phi^-_{n_k})_{k\in\mathbb{N}}$) converges weakly in $H^1(\mathbb{R})$ and strongly in $L^\infty_{\rm loc}(\mathbb{R})$ towards ϕ^+ (respectively ϕ^-).

Since for all $k \in \mathbb{N}$, $\phi^+_{n_k}|_{\mathbb{R}_-} = 0$, necessarily $\phi^+|_{\mathbb{R}_-} = 0$. Besides, $\phi^+_{n_k}|_{\mathbb{R}_+} \in X^+_{n_k}$ and

$$H^+(t)|_{X^+_{n_k}}\phi^+_{n_k}|_{\mathbb{R}_+} + \Pi_{X^+_{n_k}}\left[W(\cdot - (p_{n_k} + t)b)\phi^+_{n_k}|_{\mathbb{R}_+}\right] - \lambda_{n_k}\phi^+_{n_k}|_{\mathbb{R}_+} = 0,$$

where $\Pi_{X^+_{n_k}}$ denotes the $L^2(\mathbb{R}_+)$-orthogonal projector onto the finite-dimensional space $X^+_{n_k}$. Using Proposition 3.2.1, in particular (3.6) and the fact that $W \in L^\infty(\mathbb{R})$ with $W(x) \underset{|x|\to\infty}{\longrightarrow} 0$, it holds that the following convergence holds in $\mathcal{D}'(\mathbb{R}^*_+)$,

$$H(t)^+|_{X^+_{n_k}}\phi^+_{n_k}|_{\mathbb{R}_+} + \Pi_{X^+_{n_k}}\left[W(\cdot - (p_{n_k} + t)b)\phi^+_{n_k}|_{\mathbb{R}_+}\right] - \lambda_{n_k}\phi^+_{n_k}|_{\mathbb{R}_+} \underset{k\to\infty}{\longrightarrow} (H^+(t) - \lambda)\phi^+|_{\mathbb{R}_+}.$$

This implies that $\phi^+|_{\mathbb{R}_+} \in \mathrm{Ker}(H^+(t) - \lambda)$. Similarly, it holds that $\phi^-|_{\mathbb{R}_-} \in \mathrm{Ker}(H^-(t') - \lambda)$.

Let us now assume that $\phi^+ = \phi^- = 0$. Then, using (3.6) and the fact that $(\phi^+_{n_k})_{k\in\mathbb{N}}$ (respectively $(\phi^-_{n_k})_{k\in\mathbb{N}}$) strongly converges in $L^\infty_{\rm loc}(\mathbb{R})$ towards ϕ^+ (respectively ϕ^-), this implies that $\|\psi_{n_k}\|_{L^2} \underset{k\to\infty}{\longrightarrow} 0$, which leads to a contradiction. Thus, either ϕ^+ or ϕ^- is non identically zero, and $\lambda \in \sigma(H^+(t)) \cup \sigma(H^-(t'))$. □

We present below some numerical simulations performed with the software Scilab [2] which illustrate this result. Here, $b = 2\pi$, $V_{\text{per}}(x) = \cos(x) + 3\sin(2x+1)$ and $W(x) = -(x+2)^2 e^{-x^2}$. The operator $H^0 := -\Delta + V_{\text{per}}$ has a gap (α, β) where $\alpha \approx -0.65$ and $\beta \approx 2.22$.

We consider below simulation domains $\Omega_n = [-L_n, L_n]$ where $L_n = 50 + 2\pi n$ for different values of n. A uniform grid is used and the mesh size is equal to $h_n = h = \frac{\pi}{50}$ along with a $\mathbb{P}_1$ finite element discretization. For any $n \in \mathbb{N}$, the discretized operator $H|_{X_n}$ has an eigenvalue λ_n close to $\lambda \approx -0.34$ and we can check numerically that $\sigma(H) \cap (-0.5, 0.5) = \emptyset$. Thus, λ is a spurious eigenvalue of the operator H. The eigenvectors ψ_n of $H|_{X_n}$ associated with the eigenvalue λ_n are plotted in Fig. 3.4 for the following values of n: $n = 0, 1, 2, 3, 4, 5, 10, 15$. We can see numerically that these eigenvectors, up to translation, have a limit. This limit is in fact an eigenvector of $H^-(t)$ with $t = \frac{50}{2\pi} - \left[\frac{50}{2\pi}\right]$ where $[\cdot]$ is the integer part.

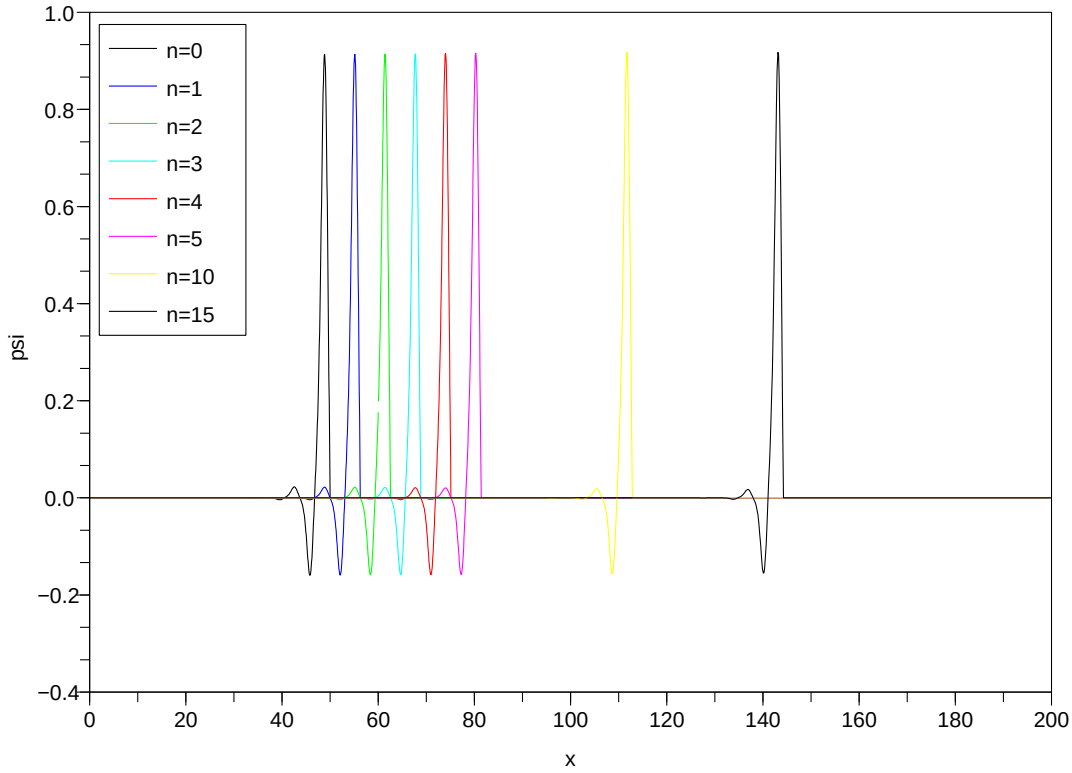

Figure 3.4: Spurious eigenvectors in a one-dimensional setting

3.5.2 Periodic boundary conditions

We present in this section some numerical results obtained when using $\mathbb{P}_1$ finite elements and periodic boundary conditions for the approximation of the spectrum of perturbed periodic Schrödinger operators $H = -\Delta + V_{\text{per}} + W$.

A first observation is the following: as in the case where the operator H was discretized with finite elements and Dirichlet boundary conditions, spurious eigenvectors localize in the vicinity of the frontier of the simulation domain. A first series of numerical tests was performed in a two-dimensional case with the finite element software FreeFem++ [1], with $V_{\rm per}(x,y) = \cos(x) + 3\sin(2(x+y)+1)$ and $W(x,y) = -(x+2)^2(2y-1)^2\exp(-(x^2+y^2))$. We consider the same gap (α,β) as in the preceding example with $\alpha \simeq -0.341$ and $\beta \simeq 0.016$, between the first and second bands of $H^0_{\rm per} = -\Delta + V_{\rm per}$. The operator $H = H^0_{\rm per} + W$ has exactly one eigenvalue in the gap (α,β) approximatively equal to -0.105. Our simulations have been performed with a sequence of $\mathbb{P}_1$-finite element approximation spaces $(X_n)_{40\leq n\leq 100}$ with periodic boundary conditions, where for each $40 \leq n \leq 100$,

- $\Omega_n = \left[-4\pi\dfrac{m_n}{n}, 4\pi\dfrac{m_n}{n}\right]^2$, with $m_n = \left[n\left(\dfrac{n-40}{20}+5\right)\right]$;
- $\mathcal{T}_n^\infty$ is a uniform $2\pi\mathbb{Z}^2$-periodic mesh of $\mathbb{R}^2$ consisting of $2n^2$ isometrical isoceles rectangular triangles per unit cell.

The spectra of the obtained discretized operators in the gap (α,β) for $40 \leq n \leq 100$ are displayed on Fig. 3.5. We clearly see that all these operators have an eigenvalue close to -0.1, which is an approximation of a true eigenvalue of H. The corresponding eigenfunction for $n = 87$ (blue circle on Fig. 3.5) is displayed on Fig. 3.6 (top); as expected, it is localized in the vicinity of the defect. On the other hand, most of these discretized operators have several eigenvalues in the range (α,β), which cannot be associated with an eigenvalue of H, and can be interpreted as spurious modes. The eigenfunction of the discretized operator close to -0.098, obtained for $n = 87$ (blue square on Fig. 3.5), is displayed on Fig. 3.6 (bottom). Similar to the case of Galerkin approximations with Dirichlet bounadry conditions (see Proposition 3.2.1), it is localized in the vicinity of the boundary of the computational domain.

A second series of numerical tests was performed on the following one-dimensional example with the software Scilab [2]: $b = 2\pi$, $V_{\rm per}(x) = \cos(x) + 3\sin(2x+1)$ and $W(x) = -(x+2)^2 e^{-x^2}$. The operator $H^0 := -\Delta + V_{\rm per}$ has a gap (α,β) where $\alpha \approx 3.85$ and $\beta \approx 4.75$.

We consider below simulation domains $\Omega_n = [-L_n, L_n]$ where $L_n = 150 + 2\pi n$ for different values of n. A uniform grid is used and the mesh size is equal to $h_n = h = \frac{\pi}{50}$ along with a $\mathbb{P}_1$ finite element discretization. For any $n \in \mathbb{N}$, the discretized operator $H|_{X_n}$ has an eigenvalue λ_n close to $\lambda \approx 4.53$ and we can check numerically that $\sigma(H) \cap (4.3, 4.7) = \emptyset$. Thus, λ is a spurious eigenvalue of the operator H. The eigenvectors $\psi_n|_{\Omega_n}$ of $H|_{X_n}$ associated with the eigenvalue λ_n are plotted in Fig. 3.7 for the following values of n: $n = -5, -4, -3, -2, -1, 0, 5, 10$.

For all $n \in \mathbb{N}$,

$$\phi_n(x) = \begin{cases} \psi_n(x - L_n) & \text{if } \ 0 \leq x \leq L_n/2, \\ \psi_n(x + L_n) & \text{if } \ -L_n/2 \leq x \leq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then, we can see numerically that the sequence $(\phi_n)_{n\in\mathbb{N}}$ has a limit, which is an eigenfunction of $H(t) = -\dfrac{d^2}{dx^2} + 1_{x<0}V_{\rm per}(x+tb) + 1_{x>0}V_{\rm per}(x-tb)$ with $t = \frac{150}{2\pi} - \left[\frac{150}{2\pi}\right]$ where $[\cdot]$ denotes the integer part.

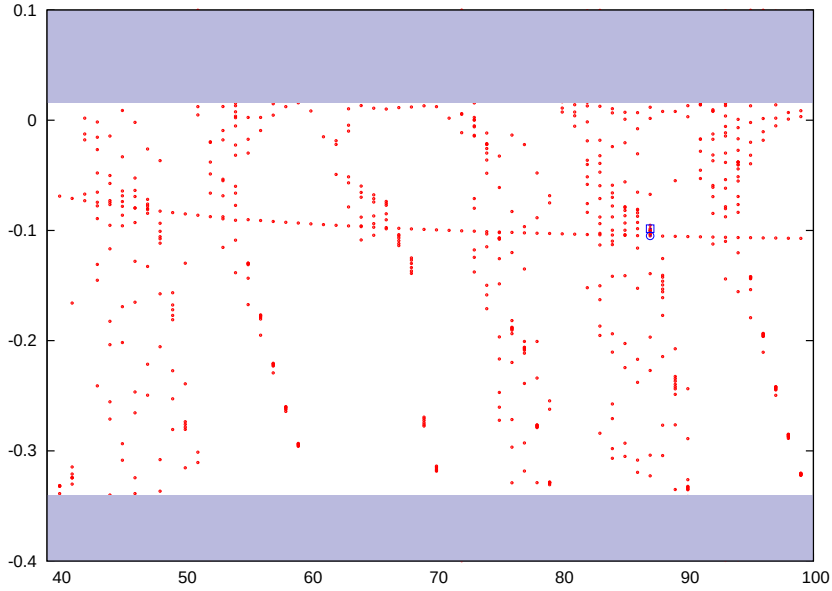

Figure 3.5: Spectrum of discretized operators with periodic boundary conditions in the gap (α, β) for $40 \leq n \leq 100$

3.5.3 Wannier function discretization

We detail here how the numerical simulations presented in Section 3.4 were obtained with the software Scilab [2]. Let us recall that in this case, $d = 1$, $V_{\mathrm{per}}(x) = \cos(x) + 3\sin(2x+1)$ and $W(x) = -(x+2)^2 e^{-x^2}$.

We want to discretize the operator

$$H = -\Delta + V_{\mathrm{per}} + W$$

with a discretized space of the form

$$\widetilde{X}_n = X_n \oplus P_n X_n$$

where X_n is a finite-dimensional subspace of $H^1(\mathbb{R})$. In our simulations, X_n is obtained with $\mathbb{P}_1$-finite elements and P_n is an approximation of the spectral projector on the lowest band of the operator $H^0_{\mathrm{per}} = -\Delta + V_{\mathrm{per}}$, i.e. P_n is defined by (3.23) with $J = 1$. The aim of this section is to precise how the finite element space X_n and the projector P_n have been constructed.

Let $n \in \mathbb{N}^*$. We introduce a discretization step h_n such that $\pi = M_n h_n$ with M_n an integer (in the code, we chose $M_n = 50$ and $h_n = \pi/50$ to be independent on $n \in \mathbb{N}$). The supercell we consider is $[-L_n, L_n]$ where $L_n = n h_n$ with n an integer. A uniform dicretization of $[-L_n, L_n]$ denoted by $(x_i)_{0 \leq i \leq 2n}$ is defined by

$$\forall -n \leq i \leq n,\ x_i = i \cdot h_n,$$

giving rise to the finite element discretization space

$$X_n = \{(\chi_i^n)_{-n \leq i \leq n}\}$$

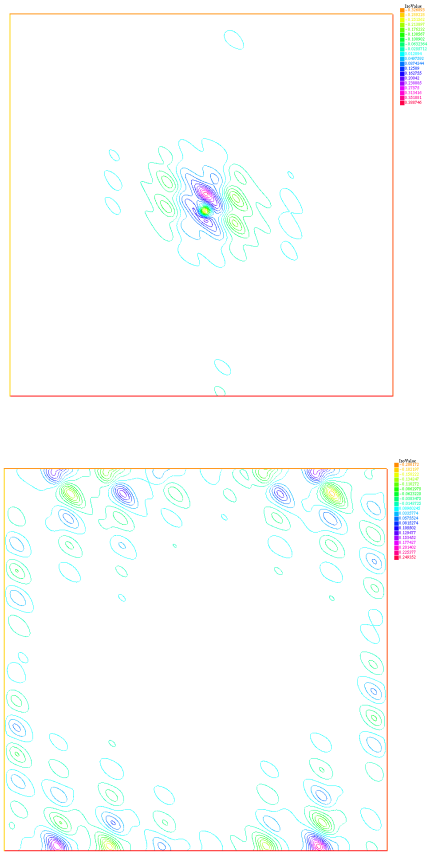

Figure 3.6: A true eigenfunction, localized close to the defect (top), and a "spurious" eigenfunction, localized close to the boundary (bottom).

where for all i, χ_i^n is the $\mathbb{P}_1$ finite element function centered in x_i.

Besides, for all $-M_n \leq k \leq M_n - 1$, we denote by

$$y_k = k \cdot h_n,$$

and by ϕ_k^n the $\mathbb{P}_1$ element function centered in y_k.

Then, for all $i \in \mathbb{N}$,

$$\chi_i^n(x) = \phi_{k_i^n}^n(x - R_i^n 2\pi),$$

with

$$k_i^n = ((i + M_n) \bmod [2M_n]) - M_n \in \{-M_n, \cdots, M_n - 1\}$$

and

$$R_i^n = \frac{(i + M_n) - ((i + M_n) \bmod [2M_n])}{2M_n}.$$

Here we consider that for all $j \in \mathbb{N}$, $j \mod [2M_n]$ is an integer between 0 and $2M_n - 1$.

Furthermore, for all $-n \leq i \leq n$, all $q \in \Gamma^* = [-1/2, 1/2)$ and $x \in \Gamma = [-\pi, \pi)$,

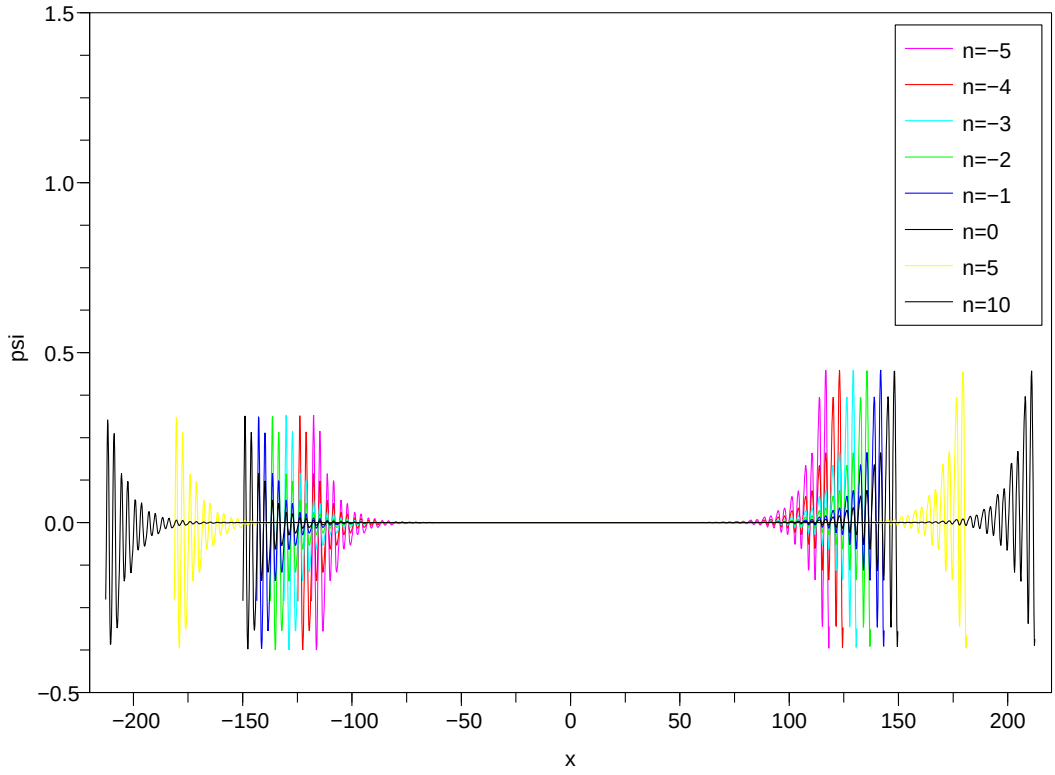

Figure 3.7: Spurious eigenvectors in a one-dimensional setting with periodic boundary conditions

$$(\chi_i^n)_q(x) = \sum_{r\in\mathbb{Z}} \chi_i^n(x + r2\pi)e^{-iqr2\pi} = \widetilde{\phi}_{k_i^n}^{q,n}(x)e^{-iqR_i^n 2\pi},$$

where the functions $\widetilde{\phi}_{k_i^n}^{q,n}$ are defined below.

For $q \in \Gamma^*$, we introduce the following discretization space

$$T_{n,q} = \text{Span}\left\{\left(\widetilde{\phi}_k^{q,n}\right)_{-M_n\le k\le(M_n-1)}\right\}$$

where for all $-M_n \le k \le M_n - 1$, $\widetilde{\phi}_k^{q,n}(x)e^{iq\cdot x}$ is 2π-periodic and for all $x \in [-\pi, \pi]$,

$$\forall -(M_n-1) \le k \le M_n - 1,\ \widetilde{\phi}_k^{q,n}(x) = \phi_k^n(x),$$

and

$$\widetilde{\phi}_{-M_n}^{q,n}(x) = \begin{cases} \phi_{-M_n}^n(x) \text{ si } x \le -\pi + h_n, \\ e^{iq2\pi}\phi_{-M_n}^n(x - 2\pi) \text{ si } x \ge \pi - h_n. \end{cases}$$

The function $e_{n,1,q}$ is an eigenfunction associated with the lowest eigenvalue $\varepsilon_{n,1,q}$ of the operator $H_q|_{T_{n,q}}$ where $H_q := -\Delta + V_{\text{per}}$.

Thus, $q \in \Gamma^*$, the function $e_{n,1,q}$ can be decomposed in the basis $\left\{\widetilde{\phi}_k^{q,n}\right\}_{-M_n \le k \le M_n - 1}$ as follows:

$$e_{n,1,q}(x) = \sum_{k=-M_n}^{M_n-1} \lambda_{k,q,n} \widetilde{\phi}_k^{q,n}(x),$$

The *Wannier function* $\psi^n \in L^2(\mathbb{R})$ associated with the first band of the operator H^0_{per} is then defined as

$$\psi^n(x) := \fint_{\Gamma^*} e_{n,1,q}(x)\, dq.$$

For all $r \in \mathbb{Z}$, we also define

$$\psi^n_r(x) := \psi^n(x - r2\pi),$$

and the family $\{\psi^n_r\}_{r\in\mathbb{Z}}$ forms an $L^2(\mathbb{R})$-orthonormal basis of $\mathrm{Ran}(P_n)$.

In our simulations, the integrals over Γ^* were computed numerically using a uniform grid $\mathcal{G}_q$ of Γ^*, with $Q = 1000$ discretization points.

It holds that

$$\begin{aligned}
\langle \widetilde{\phi}^{q,n}_{-M_n}, \widetilde{\phi}^{q,n}_{M_n} \rangle_{L^2_q(\Gamma)} &= e^{-iq2\pi} \frac{1}{6} h_n, \\
\langle \widetilde{\phi}^{q,n}_{-M_n}, \widetilde{\phi}^{q,n}_{-M_n} \rangle_{L^2_q(\Gamma)} &= \frac{2}{3} h_n, \\
\langle \widetilde{\phi}^{q,n}_{-M_n}, \widetilde{\phi}^{q,n}_{-M_n+1} \rangle_{L^2_q(\Gamma)} &= \frac{1}{6} h_n, \\
\langle \widetilde{\phi}^{q,n}_{-M_n}, \widetilde{\phi}^{q,n}_{k} \rangle_{L^2_q(\Gamma)} &= 0 \quad \text{if } k \neq M_n, -M_n, -M_n + 1, \\
\langle \widetilde{\phi}^{q,n}_{M_n}, \widetilde{\phi}^{q,n}_{-M_n} \rangle_{L^2_q(\Gamma)} &= e^{iq2\pi} \frac{1}{6} h_n, \\
\langle \widetilde{\phi}^{q,n}_{M_n}, \widetilde{\phi}^{q,n}_{M_n} \rangle_{L^2_q(\Gamma)} &= \frac{2}{3} h_n, \\
\langle \widetilde{\phi}^{q,n}_{M_n}, \widetilde{\phi}^{q,n}_{M_n-1} \rangle_{L^2_q(\Gamma)} &= \frac{1}{6} h_n, \\
\langle \widetilde{\phi}^{q,n}_{M_n}, \widetilde{\phi}^{q,n}_{k} \rangle_{L^2_q(\Gamma)} &= 0 \quad \text{if } k \neq M_n, -M_n, M_n - 1, \\
& \quad \text{if } k' \neq -M_n, M_n, \\
\langle \widetilde{\phi}^{q,n}_{k'}, \widetilde{\phi}^{q,n}_{k} \rangle_{L^2_q(\Gamma)} &= \frac{1}{6} h_n \quad \text{if } |k - k'| = 1, \\
\langle \widetilde{\phi}^{q,n}_{k'}, \widetilde{\phi}^{q,n}_{k} \rangle_{L^2_q(\Gamma)} &= \frac{2}{3} h_n \quad \text{if } k = k', \\
\langle \widetilde{\phi}^{q,n}_{k'}, \widetilde{\phi}^{q,n}_{k} \rangle_{L^2_q(\Gamma)} &= 0 \text{ otherwise.}
\end{aligned}$$

Since we have the following expansion

$$e_{n,1,q}(x) = \sum_{k=-M_n}^{M_n-1} \lambda_{k,q,n} \widetilde{\phi}_k^{q,n}(x),$$

for all $-n \leq i \leq n$ and $r, s \in \mathbb{Z}$,

$$\begin{aligned}
\langle \chi_i^n, \psi_r^n \rangle_{L^2(\mathbb{R})} &= \frac{1}{Q} \sum_{q \in \mathcal{G}_q} \sum_{k=-M_n}^{M_n-1} \lambda_{k,q,n} e^{iq\cdot(R_i^n - r)2\pi} \langle \widetilde{\phi}_{k_i^n}^n, \widetilde{\phi}_k^n \rangle_{L_q^2(\Gamma)}, \\
\langle \chi_i^n, H_{\mathrm{per}}^0 \psi_r^n \rangle_{L^2(\mathbb{R})} &= \frac{1}{Q} \sum_{q \in \mathcal{G}_q} \sum_{k=-M_n}^{M_n-1} \lambda_{k,q,n} \varepsilon_{n,1,q} e^{iq\cdot(R_i^n - r)2\pi} \langle \widetilde{\phi}_{k_i^n}^n, \widetilde{\phi}_k^n \rangle_{L_q^2(\Gamma)}, \\
\langle \psi_r^n, \psi_s^n \rangle_{L^2(\mathbb{R})} &= \frac{1}{Q} \sum_{q \in \mathcal{G}_q} \sum_{k=-M_n}^{M_n-1} \sum_{k'=-M_n}^{M_n-1} \lambda_{k,q,n}^* \lambda_{k',q,n} e^{iq\cdot(r-s)2\pi} \langle \widetilde{\phi}_k^n, \widetilde{\phi}_{k'}^n \rangle_{L_q^2(\Gamma)}, \\
\langle \psi_r^n, H_{\mathrm{per}}^0 \psi_s^n \rangle_{L^2(\mathbb{R})} &= \frac{1}{Q} \sum_{q \in \mathcal{G}_q} \sum_{k=-M_n}^{M_n-1} \sum_{k'=-M_n}^{M_n-1} \varepsilon_{n,1,q} \lambda_{k,q,n}^* \lambda_{k',q,n} e^{iq\cdot(r-s)2\pi} \langle \widetilde{\phi}_k^n, \widetilde{\phi}_{k'}^n \rangle_{L_q^2(\Gamma)}.
\end{aligned}$$

The remaining terms $\langle \chi_i^n, W\psi_r^n \rangle_{L^2(\mathbb{R})}$ and $\langle \psi_r^n, W\psi_s^n \rangle_{L^2(\mathbb{R})}$ were computed using a standard numerical integration scheme.

In practice, P_n is the orthogonal projector onto the finite-dimensional space

$$\mathrm{Span}\{\psi_r^n,\ r \in \mathbb{Z},\ |r| \leq \kappa_n\},$$

where $\kappa_n \in \mathbb{N}^*$ is a finite threshold (in the code, we chose $\kappa_n = 50$).

Now, for all $-n \leq i \leq n$,

$$P_n \chi_i^n = \sum_{s=-\kappa_n}^{\kappa_n} \mu_{i,s,n} \psi_s^n.$$

Since it holds

$$\langle \psi_r^n, P_n \chi_i^n \rangle_{L^2(\mathbb{R})} = \langle \psi_r^n, \chi_i^n \rangle_{L^2(\mathbb{R})} = \sum_{s=-\kappa_n}^{\kappa_n} \mu_{i,s,n} \langle \psi_r^n, \psi_s^n \rangle_{L^2(\mathbb{R})},$$

we can easily compute the coefficients $\mu_{i,s,n}$.

Using the preceding formulas, we can then compute all the terms $\langle \chi_i^n, P_n \chi_j^n \rangle_{L^2(\mathbb{R})}$, $\langle \chi_i^n, HP_n \chi_j^n \rangle_{L^2(\mathbb{R})}$, $\langle P_n \chi_i^n, P_n \chi_j^n \rangle_{L^2(\mathbb{R})}$ and $\langle P_n \chi_i^n, HP_n \chi_j^n \rangle_{L^2(\mathbb{R})}$ for all $-n \leq i, j \leq n$ and compute the spectral decomposition of the operator $H|_{X_n + P_n X_n}$.

Chapter 4

Non-consistent approximations of self-adjoint eigenproblems: Application to the supercell method

In this chapter, we present results which were gathered in an article submitted to *Numerische Mathematik*. We prove some a priori error estimates for the approximation of non-variational eigenvalues by means of non-consistent methods. We apply these results to the so-called supercell method to compute discrete eigenvalues located in spectral gaps of perturbed periodic Schrödinger operators, taking into account Fourier discretization an numerical integration. In particular, we prove that when no discretization method is used, the rate of convergence of the supercell method is exponential with respect to the size of the supercell.

Non-consistent approximations of self-adjoint eigenproblems: Application to the supercell method[1]

Eric Cancès[2] Virginie Ehrlacher[2] Yvon Maday[3]

Abstract

In this article, we introduce a general theoretical framework to analyze non-consistent approximations of the discrete eigenmodes of a self-adjoint operator. We focus in particular on the discrete eigenvalues laying in spectral gaps. We first provide *a priori* error estimates on the eigenvalues and eigenvectors in the absence of spectral pollution. We then show that the supercell method for perturbed periodic Schrödinger operators falls into the scope of our study. We prove that this method is spectral pollution free, and we derive optimal convergence rates for the planewave discretization method, taking numerical integration errors into account. Some numerical illustrations are provided.

4.1 Introduction

This article is concerned with the numerical analysis of the computation of the discrete eigenmodes of a self-adjoint operator A, on an infinite dimensional separable Hilbert space $\mathcal{H}$. The focus is particularly set on the eigenmodes corresponding to discrete eigenvalues located in spectral gaps.

The main application we have in mind is concerned with perturbed periodic Schrödinger operators of the form

$$A := -\Delta + V_{\rm per} + W,$$

where Δ is the Laplace operator on $L^2(\mathbb{R}^d)$, $V_{\rm per}$ a periodic function of $L^p_{\rm loc}(\mathbb{R}^d)$ with $p = 2$ if $d \leq 3$, $p > 2$ for $d = 4$ and $p = d/2$ for $d \geq 5$, and $W \in L^\infty(\mathbb{R}^d)$ a perturbation of the potential going to zero at infinity. The operator A is self-adjoint and bounded from below on $\mathcal{H} := L^2(\mathbb{R}^d)$ with domain $H^2(\mathbb{R}^d)$. Perturbed periodic Schrödinger operators are encountered in electronic structure theory, and in the study of photonic crystals. In the case of a perfectly periodic crystal ($W = 0$), the spectrum of the operator $A^0 := -\Delta + V_{\rm per}$ is purely absolutely continuous, and composed of a union of intervals of $\mathbb{R}$. It follows from

[1]This work was financially supported by the ANR grant MANIF.

[2]Université Paris Est, CERMICS, Projet MICMAC, Ecole des Ponts ParisTech - INRIA, 6 & 8 avenue Blaise Pascal, 77455 Marne-la-Vallée Cedex 2, France, (cances@cermics.enpc.fr, ehrlachv@cermics.enpc.fr)

[3]Université Pierre et Marie Curie-Paris 6, UMR 7598, Laboratoire J.-L. Lions, Paris, F-75005 France, and Division of Applied Mathematics, Brown University, 182 George Street, Providence, RI 02912, USA, (maday@ann.jussieu.fr)

Weyl's theorem [163] that the essential spectra of A and A^0 are identical. On the other hand, when $W \neq 0$, some discrete eigenvalues may appear in the band gaps of the spectrum of A. The corresponding eigenmodes, which can be interpreted as bound states trapped by local defects, are difficult to compute for numerical methods can produce spectral pollution.

In a general theoretical framework, the eigenvalues of A and the associated eigenvectors can be obtained by solving the variational problem

$$\begin{cases} \text{find } (\psi, \lambda) \in Q(A) \times \mathbb{R} \text{ such that} \\ \forall \phi \in Q(A),\ a(\psi, \phi) = \lambda m(\psi, \phi), \end{cases} \tag{4.1}$$

where $m(\cdot, \cdot)$ is the scalar product of $\mathcal{H}$, $Q(A)$ the form domain of A, and $a(\cdot, \cdot)$ the sesquilinear form associated with A (see for instance [56]).

A sequence $(X_n)_{n\in\mathbb{N}}$ of finite dimensional approximation subspaces of $Q(A)$ being given, we consider for all $n \in \mathbb{N}$, the self-adjoint operator $A|_{X_n} : X_n \to X_n$ defined by

$$\forall (\psi_n, \phi_n) \in X_n \times X_n, \quad m\left(A|_{X_n}\psi_n, \phi_n\right) = a(\psi_n, \phi_n).$$

The standard Galerkin method consists in approximating the discrete eigenvalues of the operator A by the eigenvalues of the discretized operators $A|_{X_n}$, the latter being obtained by solving the variational problem

$$\begin{cases} \text{find } (\psi_n, \lambda_n) \in X_n \times \mathbb{R} \text{ such that} \\ \forall \phi_n \in X_n,\ a(\psi_n, \phi_n) = \lambda_n m(\psi_n, \phi_n). \end{cases}$$

According to the Rayleigh-Ritz theorem [163], under the natural assumption that the sequence $(X_n)_{n\in\mathbb{N}}$ satisfies

$$\forall \phi \in Q(A), \quad \inf_{\phi_n \in X_n} \|\phi - \phi_n\|_{Q(A)} \underset{n\to\infty}{\longrightarrow} 0,$$

this method allows to compute the eigenmodes of A associated with the discrete eigenvalues located below the bottom of the essential spectrum. Indeed, if A is bounded below and possesses exactly M discrete eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_M$ (taking multiplicities into account) lower than $\min \sigma_{\text{ess}}(A)$, where $\sigma_{\text{ess}}(A)$ denotes the essential spectrum of A, and if $\left\{\lambda_j^n\right\}_{1\leq j\leq \dim\ X_n}$ are the eigenvalues of $A|_{X_n}$, it is well-known that

$$\forall 1 \leq j \leq M,\ \lambda_j^n \underset{n\to\infty}{\downarrow} \lambda_j.$$

The situation is much more delicate when one tries to approximate eigenvalues which are located in spectral gaps of A since

$$\forall M < j \leq \dim\ X_n, \quad \lambda_j^n \underset{n\to\infty}{\downarrow} \min \sigma_{\text{ess}}(A).$$

When dealing with the approximation of discrete eigenvalues of A located in spectral gaps, the standard Galerkin method may give rise to spectral pollution: some sequences $(\lambda_n)_{n\in\mathbb{N}}$, where for each n, $\lambda_n \in \sigma(A|_{X_n})$, may converge to real numbers which do not belong to the spectrum of A. Spectral pollution occurs in a broad variety of physical settings, including elasticity theory, electromagnetism, hydrodynamics and quantum physics [7, 20, 64, 162, 166], and has been extensively studied in the framework of the standard Galerkin method [21, 22, 23, 65, 71, 72, 73, 110, 131]. We refer to [24, 40, 134] for an analysis of spectral pollution for perturbed periodic Schrödinger operators.

On the other hand, few results have been published on the numerical computation of eigenmodes in spectral gaps by means of non-consistent methods, based on generalized eigenvalue problems of the form

$$\begin{cases} \text{find } (\psi_n, \lambda_n) \in X_n \times \mathbb{R} \text{ such that} \\ \forall \phi_n \in X_n, \quad a_n(\psi_n, \phi_n) = \lambda_n m_n(\psi_n, \phi_n), \end{cases}$$

where for all $n \in \mathbb{N}$, $a_n(\cdot,\cdot)$ and $m_n(\cdot,\cdot)$ are symmetric bilinear forms on $Q(A)$, *a priori* different from $a(\cdot,\cdot)$ and $m(\cdot,\cdot)$.

In this article, we consider a general theoretical framework to analyze non-consistent methods for the computation of the discrete eigenmodes of a self-adjoint operator. After introducing some notation and definitions in Sections 4.2.1 and 4.2.2, we state our main result (Theorem 4.3.1) in Section 4.3. Theorem 4.3.1 provides *a priori* error estimates on the eigenvalues and eigenvectors in the absence of spectral pollution. Its proof is given in Section 4.4.

In Section 4.5, we show that the supercell method for perturbed periodic Schrödinger operators falls into the scope of Theorem 4.3.1. We prove that this method is spectral pollution free, and we derive optimal convergence rates for the planewave discretization method, taking numerical integration errors into account. The corresponding proofs are detailed in Section 4.6, and some numerical illustrations are provided in Section 4.7.

4.2 Approximations of a self-adjoint operator

4.2.1 Some notation

Throughout this paper, $\mathcal{H}$ denotes a separable Hilbert space, endowed with the scalar product $m(\cdot,\cdot)$ and associated norm $\|\cdot\|_{\mathcal{H}}$, and A a self-adjoint operator on $\mathcal{H}$ with dense domain $D(A)$. We denote by $Q(A) := D(|A|^{1/2})$ the form domain of A and by $a(\cdot,\cdot)$ the symmetric bilinear form on $Q(A)$ associated with A. Recall that the vector space $Q(A)$, endowed with the scalar product $\langle\cdot,\cdot\rangle_{Q(A)}$, defined as

$$\forall \psi, \phi \in Q(A), \quad \langle \psi, \phi \rangle_{Q(A)} := m(\psi,\phi) + m\left(|A|^{1/2}\psi, |A|^{1/2}\phi\right),$$

is a Hilbert space; the associated norm is denoted by $\|\cdot\|_{Q(A)}$.

Example 4.2.1. *Perturbed periodic Schrödinger operators $A := -\Delta + V_{\mathrm{per}} + W$ are self-adjoint semibounded operators on $\mathcal{H} := L^2(\mathbb{R}^d)$, with domain $D(A) := H^2(\mathbb{R}^d)$ and form domain $Q(A) := H^1(\mathbb{R}^d)$.*

For any finite dimensional vector subspace X of $\mathcal{H}$ such that $X \subset Q(A)$, we introduce the following notation

- $i_X : X \hookrightarrow \mathcal{H}$ is the canonical embedding of X into $\mathcal{H}$;
- $i_X^* : \mathcal{H} \to X$ is the adjoint of i_X, that is the orthogonal projection from $\mathcal{H}$ onto X associated with the scalar product $m(\cdot,\cdot)$;
- $A|_X : X \to X$ is the self-adjoint operator on X defined by

$$\forall (\psi,\phi) \in X \times X, \quad m\left(A|_X \psi, \phi\right) = a(\psi,\phi);$$

- $\Pi_X^{\mathcal{H}} : \mathcal{H} \to \mathcal{H}$ and $\Pi_X^{Q(A)} : Q(A) \to Q(A)$ are the orthogonal projections onto X for $(\mathcal{H}, m(\cdot,\cdot))$ and $\big(Q(A), \langle\cdot,\cdot\rangle_{Q(A)}\big)$, respectively.

We set

$$\widehat{\sigma}_{\rm ess}(A) := \overline{\sigma(A)}^{\overline{\mathbb{R}}} \setminus \sigma_{\rm d}(A),$$

where $\sigma_{\rm d}(A)$ is the discrete spectrum of A, and where $\overline{\sigma(A)}^{\overline{\mathbb{R}}}$ is the closure of $\sigma(A)$, the spectrum of A, in $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$. A *spectral gap* of A is an interval (Σ^-, Σ^+) such that $\Sigma^-, \Sigma^+ \in \widehat{\sigma}_{\rm ess}(A) \cap \mathbb{R}$ and $(\Sigma^-, \Sigma^+) \cap \widehat{\sigma}_{\rm ess}(A) = \emptyset$ (which implies that $\mathrm{Tr}(\mathbb{1}_{(-\infty,\Sigma^-]}(A)) = \mathrm{Tr}(\mathbb{1}_{[\Sigma^+,\infty)}(A)) = \infty$). As usual, $\mathbb{1}_B$ denotes the characteristic function of the Borel set $B \subset \mathbb{R}$. The discrete eigenvalues of the operator A in a spectral gap (Σ^-, Σ^+), if any, are isolated and of finite multiplicities, but can accumulate at Σ^- and/or Σ^+ [163].

Let us finally recall the notions of limit superior and limit inferior of a sequence of sets of complex numbers (see for instance [56]).

Definition 4.2.1. *Let $(E_n)_{n\in\mathbb{N}}$ be a sequence of subsets of $\mathbb{C}$.*

- *The set $\overline{\lim\limits_{n\to\infty}} E_n$ (limit superior) is the set of all complex numbers $\lambda \in \mathbb{C}$ such that there exist a subsequence $(E_{n_k})_{k\in\mathbb{N}}$ of $(E_n)_{n\in\mathbb{N}}$ and a sequence $(\lambda_{n_k})_{k\in\mathbb{N}}$ of complex numbers such that for all $k \in \mathbb{N}$, $\lambda_{n_k} \in E_{n_k}$ and $\lim\limits_{k\to\infty} \lambda_{n_k} = \lambda$.*

- *The set $\underline{\lim\limits_{n\to\infty}} E_n$ (limit inferior) is the set of all complex numbers $\lambda \in \mathbb{C}$ such that there exists a sequence $(\lambda_n)_{n\in\mathbb{N}}$ of complex numbers such that for all $n \in \mathbb{N}$, $\lambda_n \in E_n$ and $\lim\limits_{n\to\infty} \lambda_n = \lambda$.*

- *If $\underline{\lim\limits_{n\to\infty}} E_n = \overline{\lim\limits_{n\to\infty}} E_n$, then $\lim\limits_{n\to\infty} E_n := \underline{\lim\limits_{n\to\infty}} E_n = \overline{\lim\limits_{n\to\infty}} E_n$.*

4.2.2 Consistent and non-consistent approximations

Definition 4.2.2. *An approximation $(\mathcal{T}_n)_{n\in\mathbb{N}}$ of a self-adjoint operator A is a sequence such that, for all $n \in \mathbb{N}$,*

$$\mathcal{T}_n := (X_n, a_n, m_n),$$

where

- *$(X_n)_{n\in\mathbb{N}}$ is a sequence of finite dimensional subspaces of $Q(A)$;*

- *$(a_n)_{n\in\mathbb{N}}$ is a sequence of symmetric bilinear forms on $Q(A)$;*

- *$(m_n)_{n\in\mathbb{N}}$ is a sequence of symmetric bilinear forms on $Q(A)$ such that the restriction of m_n to X_n forms a scalar product on X_n. We denote by $\|\cdot\|_{X_n}$ the associated norm: $\forall \phi_n \in X_n$, $\|\phi_n\|_{X_n} = m_n(\phi_n, \phi_n)^{1/2}$.*

The approximation $(\mathcal{T}_n)_{n\in\mathbb{N}}$ is called consistent if, for any (ψ, λ) solution of (4.1),

$$\forall \phi_n \in X_n, \quad a_n(\psi, \phi_n) = \lambda m_n(\psi, \phi_n),$$

and non-consistent otherwise.

The approximation $(\mathcal{T}_n)_{n\in\mathbb{N}}$ is referred to as a standard Galerkin method if, for all $n \in \mathbb{N}$, $a_n = a$ and $m_n = m$. Standard Galerkin methods are obviously consistent.

If $(\mathcal{T}_n)_{n\in\mathbb{N}}$ is an approximation of A, we denote by $\mathcal{A}_n$ and $\mathcal{M}_n$ the m-symmetric (i.e. symmetric w.r.t. the scalar product $m(\cdot,\cdot)$) linear operators on X_n defined by: $\forall \phi_n, \psi_n \in X_n$,

$$\begin{aligned} m\left(\mathcal{A}_n\phi_n, \psi_n\right) &= a_n(\phi_n, \psi_n), \\ m\left(\mathcal{M}_n\phi_n, \psi_n\right) &= m_n(\phi_n, \psi_n). \end{aligned}$$

Since m_n is a scalar product on X_n, the operator $\mathcal{M}_n$ is invertible and we can define the operator

$$A_n = \mathcal{M}_n^{-1/2}\mathcal{A}_n\mathcal{M}_n^{-1/2}$$

on X_n, which is m-symmetric as well. The generalized eigenvalue problem

$$\left\{ \begin{array}{l} \text{find } (\psi_n, \lambda_n) \in X_n \times \mathbb{R} \text{ such that } \|\psi_n\|_{X_n}^2 = 1 \text{ and} \\ \forall \phi_n \in X_n,\ a_n(\psi_n, \phi_n) = \lambda_n m_n(\psi_n, \phi_n), \end{array} \right. \tag{4.2}$$

is then equivalent, through the change of variable $\xi_n = \mathcal{M}_n^{1/2}\psi_n$, to the eigenvalue problem

$$\left\{ \begin{array}{l} \text{find } (\xi_n, \lambda_n) \in X_n \times \mathbb{R} \text{ such that } \|\xi_n\|_{\mathcal{H}}^2 = 1 \text{ and} \\ A_n\xi_n = \lambda_n\xi_n. \end{array} \right.$$

The main objective of this work is to provide sufficient conditions on such potentially non-consistent approximations $(\mathcal{T}_n)_{n\in\mathbb{N}}$ so that the discrete eigenvalues of A and the associated eigenvectors are well-approximated in a certain sense by eigenvalues and eigenvectors of the discretized problems (4.2). We wish to provide a framework which will enable us to deal with the supercell method for perturbed periodic linear Schrödinger operators described in Section 4.5.

4.3 An abstract convergence result

4.3.1 The general case

Let us consider an approximation $(\mathcal{T}_n)_{n\in\mathbb{N}}$ of A satisfying the following assumptions:

(A1) $\forall \psi \in Q(A),\ \left\| \left(1 - \Pi_{X_n}^{Q(A)}\right)\psi \right\|_{Q(A)} \underset{n\to\infty}{\longrightarrow} 0$;

(A2) there exists $0 < \gamma \leq \Gamma < \infty$ such that for all $n \in \mathbb{N}$ and all $\psi_n, \phi_n \in X_n$,

$$\begin{aligned} \gamma\|\psi_n\|_{\mathcal{H}}^2 \leq \quad & m_n(\psi_n, \psi_n) \quad \leq \Gamma\|\psi_n\|_{\mathcal{H}}^2, \\ & |a_n(\psi_n, \phi_n)| \quad \leq \Gamma\|\psi_n\|_{Q(A)}\|\phi_n\|_{Q(A)}; \end{aligned}$$

(A3) for any compact subset $K \subset \mathbb{C}$, if there exists a subsequence $(\mathcal{T}_{n_k})_{k\in\mathbb{N}}$ of $(\mathcal{T}_n)_{n\in\mathbb{N}}$ such that $\operatorname{dist}\left(K, \sigma(A_{n_k})\right) \geq \alpha_K$ for some $\alpha_K > 0$ independent of $k \in \mathbb{N}$, then there exists $c_K > 0$ such that for all $\mu \in K$ and all $k \in \mathbb{N}$,

$$\inf_{w_{n_k}\in X_{n_k}} \sup_{v_{n_k}\in X_{n_k}} \frac{|(a_{n_k} - \mu m_{n_k})(w_{n_k}, v_{n_k})|}{\|w_{n_k}\|_{Q(A)}\|v_{n_k}\|_{Q(A)}} \geq c_K;$$

(A4) there exist $\kappa \in \mathbb{R}_+$ and, for each $n \in \mathbb{N}$, two symmetric bilinear forms $\widetilde{a}_n$ and $\widetilde{m}_n$ on $Q(A)$, and four seminorms r_n^a, r_n^m, s_n^a, s_n^m on $Q(A)$ such that $\forall \phi_n, \psi_n \in X_n$,

$$\begin{aligned} \gamma\|\psi_n\|_{\mathcal{H}}^2 \leq \quad & \widetilde{m}_n(\psi_n, \psi_n) \quad \leq \Gamma\|\psi_n\|_{\mathcal{H}}^2, \\ & |\widetilde{a}_n(\psi_n, \phi_n)| \quad \leq \Gamma\|\psi_n\|_{Q(A)}\|\phi_n\|_{Q(A)}, \end{aligned}$$

and $\forall \phi, \psi \in Q(A)$,

$$
\begin{array}{ll}
|(a - \widetilde{a}_n)(\phi,\psi)| \leq r_n^a(\phi) r_n^a(\psi), & |(m - \widetilde{m}_n)(\phi,\psi)| \leq r_n^m(\phi) r_n^m(\psi), \\
r_n^a(\phi) \leq \kappa \|\phi\|_{Q(A)}, & r_n^m(\phi) \leq \kappa \|\phi\|_{\mathcal{H}}, \\
r_n^a(\phi) \underset{n\to\infty}{\longrightarrow} 0, & r_n^m(\phi) \underset{n\to\infty}{\longrightarrow} 0,
\end{array}
$$

and

$$
\sup_{w_n \in X_n} \frac{|(a_n - \widetilde{a}_n)(\Pi_{X_n}^{Q(A)}\phi, w_n)|}{\|w_n\|_{Q(A)}} \leq s_n^a(\phi), \qquad \sup_{w_n \in X_n} \frac{|(m_n - \widetilde{m}_n)(\Pi_{X_n}^{Q(A)}\phi, w_n)|}{\|w_n\|_{Q(A)}} \leq s_n^m(\phi),
$$

$$
\begin{array}{ll}
s_n^a(\phi) \leq \kappa \|\phi\|_{Q(A)}, & s_n^m(\phi) \leq \kappa \|\phi\|_{\mathcal{H}}, \\
s_n^a(\phi) \underset{n\to\infty}{\longrightarrow} 0, & s_n^m(\phi) \underset{n\to\infty}{\longrightarrow} 0.
\end{array}
$$

Before stating our main result, let us comment on these assumptions.

Conditions (A1) and (A2) are classical. The former means that any $\psi \in Q(A)$ can be approximated in $Q(A)$ by a sequence $(\psi_n)_{n\in\mathbb{N}}$ such that $\psi_n \in X_n$ for each $n \in \mathbb{N}$. The latter ensures that, *uniformly in* n, the norms $\|\cdot\|_{X_n}$ and $\|\cdot\|_{\mathcal{H}}$ are equivalent on X_n, and the bilinear forms a_n are continuous on X_n, the space X_n being endowed with the norm $\|\cdot\|_{Q(A)}$.

Assumption (A3) is important in our proof since it enables us to apply Strang's lemma (see Section 4.8) with a uniform discrete inf-sup condition.

For the supercell approximation, we will prove a stronger result:

(A3') for any compact subset $K \subset \mathbb{C}$, there exists $c_K > 0$ such that for all $n \in \mathbb{N}$ and all $\mu \in K$,

$$
\inf_{w_n \in X_n} \sup_{v_n \in X_n} \frac{|(a_n - \mu m_n)(w_n, v_n)|}{\|w_n\|_{Q(A)} \|v_n\|_{Q(A)}} \geq c_K \min(1, \operatorname{dist}(\mu, \sigma(A_n))).
$$

It is easily checked that (A3') implies (A3).

Let us finally comment on condition (A4) in the perspective of the analysis of the supercell method with numerical integration addressed in Section 4.5. In the latter setting, the introduction of the bilinear forms $\widetilde{a}_n$ and $\widetilde{m}_n$ aims at separating in the error bounds of Theorem 4.3.1 the contributions inherently due to the supercell method (truncation of the domain and artificial periodic boundary conditions) and those due to numerical integration. We postpone until Section 4.5 the precise definitions of a_n, m_n, $\widetilde{a}_n$ and $\widetilde{m}_n$ in this context.

Note that (A4) implies that the approximation $(\mathcal{T}_n)_{n\in\mathbb{N}}$ is weakly consistent in the sense that for all $\phi \in Q(A)$, the consistency errors $r_n^a(\phi)$, $r_n^m(\phi)$, $s_n^a(\phi)$ and $s_n^m(\phi)$ converge to 0 as n goes to infinity.

We are now in position to state our main result.

Theorem 4.3.1. *Let A be a self-adjoint operator on $\mathcal{H}$, $\lambda \in \sigma_{\mathrm{d}}(A)$ a discrete eigenvalue of A with multiplicity q, and $(\mathcal{T}_n)_{n\in\mathbb{N}}$ an approximation of A satisfying assumptions (A1)-(A4). Then,*

1. ***Convergence of the eigenvalues***

$$
\lambda \in \underline{\lim}\, \sigma(A_n). \tag{4.3}
$$

2. ***A priori error estimates in the absence of spectral pollution***

Assume that

(B1) $\exists \varepsilon > 0$ s.t. $(\lambda-\varepsilon,\lambda+\varepsilon)\cap\sigma(A)=\{\lambda\}$ and $\overline{\lim_{n\to\infty}}\,\sigma(A_n)\cap(\lambda-\varepsilon,\lambda+\varepsilon)=\{\lambda\}$.

Let $\mathcal{P}:=\mathbb{1}_{\{\lambda\}}(A)$ *be the orthogonal projection on* $\mathrm{Ker}(A-\lambda)$ *and*

$$\mathcal{P}_n := i_{X_n}\mathcal{M}_n^{-1/2}\mathbb{1}_{(\lambda-\varepsilon/2,\lambda+\varepsilon/2)}(A_n)\mathcal{M}_n^{1/2}i_{X_n}^*.$$

Then,

$$\mathrm{Rank}(\mathcal{P}_n)\geq q, \tag{4.4}$$

and there exists $C\in\mathbb{R}_+$ *such that, for* n *large enough,*

$$\|(\mathcal{P}-\mathcal{P}_n)\mathcal{P}\|_{\mathcal{L}(\mathcal{H},Q(A))}\leq C\left(\left\|\left(1-\Pi_{X_n}^{Q(A)}\right)\mathcal{P}\right\|_{\mathcal{L}(\mathcal{H},Q(A))}+\mathcal{R}_n^a+\mathcal{R}_n^m+\mathcal{S}_n^a+\mathcal{S}_n^m\right), \tag{4.5}$$

with

$$\begin{aligned}
\mathcal{R}_n^a &:= \sup_{\psi\in\mathrm{Ran}(\mathcal{P}),\,\|\psi\|_{\mathcal{H}}=1} r_n^a(\psi),\\
\mathcal{R}_n^m &:= \sup_{\psi\in\mathrm{Ran}(\mathcal{P}),\,\|\psi\|_{\mathcal{H}}=1} r_n^m(\psi),\\
\mathcal{S}_n^a &:= \sup_{\psi\in\mathrm{Ran}(\mathcal{P}),\,\|\psi\|_{\mathcal{H}}=1} s_n^a(\psi),\\
\mathcal{S}_n^m &:= \sup_{\psi\in\mathrm{Ran}(\mathcal{P}),\,\|\psi\|_{\mathcal{H}}=1} s_n^m(\psi).
\end{aligned}$$

If we assume in addition that

(B2) *for* n *large enough,* $\mathrm{Rank}(\mathcal{P}_n)=q$,

then there exists $C\in\mathbb{R}_+$ *such that, for* n *large enough,*

$$\|(\mathcal{P}-\mathcal{P}_n)\mathcal{P}_n\|_{\mathcal{L}(\mathcal{H},Q(A))}\leq C\left(\left\|\left(1-\Pi_{X_n}^{Q(A)}\right)\mathcal{P}\right\|_{\mathcal{L}(\mathcal{H},Q(A))}+\mathcal{R}_n^a+\mathcal{R}_n^m+\mathcal{S}_n^a+\mathcal{S}_n^m\right), \tag{4.6}$$

$$\max_{\lambda_n\in\sigma(A_n)\cap(\lambda-\varepsilon/2,\lambda+\varepsilon/2)}|\lambda_n-\lambda|\leq C\left(\left(\left\|\left(1-\Pi_{X_n}^{Q(A)}\right)\mathcal{P}\right\|_{\mathcal{L}(\mathcal{H},Q(A))}+\mathcal{R}_n^a+\mathcal{R}_n^m\right)^2+\mathcal{S}_n^a+\mathcal{S}_n^m\right). \tag{4.7}$$

It is easy to check that $P_n := \mathcal{M}_n^{-1/2}\mathbb{1}_{(\lambda-\varepsilon/2,\lambda+\varepsilon/2)}(A_n)\mathcal{M}_n^{1/2}$ is the m_n-orthogonal projection of X_n onto the space $Y_n\subset X_n$ spanned by the eigenvectors of (4.2) associated with the eigenvalues belonging to the interval $(\lambda-\varepsilon/2,\lambda+\varepsilon/2)$. The operator $\mathcal{P}_n = i_{X_n}P_n i_{X_n}^*\in\mathcal{L}(\mathcal{H})$ is therefore a (non-orthogonal) projection on the finite dimensional space $i_{X_n}Y_n\subset\mathcal{H}$.

Theorem 4.3.1 implies that, if $(\mathcal{T}_n)_{n\in\mathbb{N}}$ is an approximation of the operator A satisfying (A1)-(A4), for all discrete eigenvalue λ of A, there exists a sequence $(\lambda_n)_{n\in\mathbb{N}}$ of elements of $\sigma(A_n)$ converging to λ. Assumption (B1) states that there is no spurious eigenvalues in the vicinity of λ. Estimate (4.5) shows that under assumption (B1), for each eigenvector ψ of A associated with the discrete eigenvalue λ, there exists a sequence $(\psi_n)_{n\in\mathbb{N}}$ of elements of $\mathrm{Ran}(\mathcal{P}_n)$ which strongly converges towards ψ in $Q(A)$.

On the other hand, there may *a priori* exist a sequence $(\psi_n)_{n\in\mathbb{N}}$ of normalized elements of $\mathrm{Ran}(\mathcal{P}_n)$ weakly converging in $\mathcal{H}$ towards a vector that is not an eigenvector of A associated with λ. This is excluded when we make the additional assumption (B2). Assumption (B2) means that, for n large enough, the sum of the multiplicities of the eigenvalues of A_n close to λ is equal to the multiplicity q of λ. Under this assumption, if $(\psi_n)_{n\in\mathbb{N}}$ is a sequence of vectors of $\mathcal{H}$ such that for each n large enough, ψ_n is an $\mathcal{H}$-normalized eigenvector of A_n associated with an eigenvalue $\lambda_n \in (\lambda-\varepsilon, \lambda+\varepsilon)$, and if $(\psi_n)_{n\in\mathbb{N}}$ weakly converges in $\mathcal{H}$ towards some $\psi \in \mathcal{H}$, then estimate (4.6) implies that ψ is a $\mathcal{H}$-normalized eigenvector of A associated with the eigenvalue λ and that the convergence of $(\psi_n)_{n\in\mathbb{N}}$ to ψ holds strongly in $Q(A)$.

Lastly, estimate (4.7) shows that when $\widetilde{a}_n = a_n$ and $\widetilde{m}_n = m_n$ (which is the case in the supercell model when numerical integration errors are neglected), then $\mathcal{S}_n^a = \mathcal{S}_n^m = 0$, and the convergence rate of the eigenvalues is twice the convergence rate of the eigenvectors measured in the $Q(A)$ norm. Such a doubling of the convergence rate is expected in variational approximations of linear eigenvalue problems (see e.g. [56]).

4.3.2 Standard Galerkin method

Let us now consider the special case when for all $n \in \mathbb{N}$, $\mathcal{T}_n = (X_n, a, m)$ where $(X_n)_{n\in\mathbb{N}}$ is a sequence of finite dimensional subspaces of $Q(A)$ satisfying (A1). In this case, for all $n \in \mathbb{N}$, $A_n = A|_{X_n}$, $\mathcal{M}_n$ is the identity operator, and

$$\mathcal{P}_n = i_{X_n} \mathbb{1}_{(\lambda-\varepsilon/2,\lambda+\varepsilon/2)}(A|_{X_n}) i_{X_n}^*$$

is an orthogonal projector with respect to the scalar product m.

In this setting, (A2) and (A4) are obviously satisfied, and (A3) and (A3') respectively read

(C3) for any compact subset $K \subset \mathbb{C}$, if there exists a subsequence $(X_{n_k})_{k\in\mathbb{N}}$ of $(X_n)_{n\in\mathbb{N}}$ such that $\mathrm{dist}\,(K, \sigma(A|_{X_{n_k}})) \geq \alpha_K$ for some $\alpha_K > 0$ independent of $k \in \mathbb{N}$, then there exists $c_K > 0$ such that for all $\mu \in K$ and all $k \in \mathbb{N}$,

$$\inf_{w_{n_k}\in X_{n_k}} \sup_{v_{n_k}\in X_{n_k}} \frac{|(a-\mu m)(w_{n_k}, v_{n_k})|}{\|w_{n_k}\|_{Q(A)} \|v_{n_k}\|_{Q(A)}} \geq c_K;$$

and

(C3') for all compact subset $K \subset \mathbb{C}$, there exists $c_K > 0$ such that for all $n \in \mathbb{N}$ and all $\mu \in K$,

$$\inf_{w_n\in X_n} \sup_{v_n\in X_n} \frac{|(a-\mu m)(w_n, v_n)|}{\|w_n\|_{Q(A)} \|v_n\|_{Q(A)}} \geq c_K \min(1, \mathrm{dist}(\mu, \sigma(A|_{X_n}))).$$

It is proved in Appendix 4.9 that, when A is semibounded, (C3'), and thus (C3), automatically hold. On the other hand, when A is not semibounded, (C3) is not always satisfied. An explicit counterexample is given in Appendix 4.9.

The formulation of Theorem 4.3.1 simplifies in this case as follows:

Corollary 4.3.1. *Let A be a self-adjoint operator on $\mathcal{H}$, $\lambda \in \sigma_{\mathrm{d}}(A)$ a discrete eigenvalue of A with multiplicity q, and $(X_n)_{n\in\mathbb{N}}$ a sequence of finite dimensional subspaces of $Q(A)$ such that*

$$\forall \psi \in Q(A), \quad \left\| \left(1 - \Pi_{X_n}^{Q(A)}\right) \psi \right\|_{Q(A)} \underset{n\to\infty}{\longrightarrow} 0.$$

Let us assume that either A is semibounded or $(X_n)_{n\in\mathbb{N}}$ satisfies assumption (C3). Then,

1. ***Convergence of the eigenvalues***

$$\lambda \in \underline{\lim}\, \sigma(A|_{X_n}).$$

2. ***A priori error estimates in the absence of spectral pollution***

Assume that

$$\text{(D1)} \quad \exists \varepsilon > 0 \text{ s.t. } (\lambda-\varepsilon,\lambda+\varepsilon)\cap\sigma(A) = \{\lambda\} \text{ and } \overline{\lim_{n\to\infty}}\, \sigma(A_n)\cap(\lambda-\varepsilon,\lambda+\varepsilon) = \{\lambda\}.$$

Let $\mathcal{P} := \mathbb{1}_{\{\lambda\}}(A)$ *be the orthogonal projection on* $\mathrm{Ker}(A-\lambda)$ *and* $\mathcal{P}_n := i_{X_n}\mathbb{1}_{(\lambda-\varepsilon/2,\lambda+\varepsilon/2)}(A|_{X_n})i_{X_n}^*$. *Then,*

$$\mathrm{Rank}(\mathcal{P}_n) \geq q,$$

and there exists $C \in \mathbb{R}_+$ *such that, for* n *large enough,*

$$\|(\mathcal{P}-\mathcal{P}_n)\mathcal{P}\|_{\mathcal{L}(\mathcal{H},Q(A))} \leq C \left\| \left(1-\Pi_{X_n}^{Q(A)}\right)\mathcal{P}\right\|_{\mathcal{L}(\mathcal{H},Q(A))}. \tag{4.8}$$

If we assume in addition that

$$\text{(D2)} \quad \textit{for } n \textit{ large enough, } \mathrm{Rank}(\mathcal{P}_n) = q,$$

then there exists $C \in \mathbb{R}_+$ *such that, for* n *large enough,*

$$\|(\mathcal{P}-\mathcal{P}_n)\mathcal{P}_n\|_{\mathcal{L}(\mathcal{H},Q(A))} \leq C \left\| \left(1-\Pi_{X_n}^{Q(A)}\right)\mathcal{P}\right\|_{\mathcal{L}(\mathcal{H},Q(A))}, \tag{4.9}$$

$$\max_{\lambda_n\in\sigma(A_n)\cap(\lambda-\varepsilon/2,\lambda+\varepsilon/2)} |\lambda_n-\lambda| \leq C\left(\left\| \left(1-\Pi_{X_n}^{Q(A)}\right)\mathcal{P}\right\|_{\mathcal{L}(\mathcal{H},Q(A))}\right)^2. \tag{4.10}$$

The estimates (4.8), (4.9) and (4.10) are optimal. They are similar to the ones proved in [73, 149, 150], but our assumptions on the sequence of discretized operators $A|_{X_n}$ are different. In [73], these estimates are proved under the condition

$$\delta(A, A|_{X_n}) \underset{n\to\infty}{\longrightarrow} 0, \tag{4.11}$$

where

$$\delta(A, A|_{X_n}) := \sup_{\phi\in D(A),\ \|\phi\|_{\mathcal{H}}+\|A\phi\|_{\mathcal{H}}=1} \inf_{\phi_n\in X_n} \|\phi-\phi_n\|_{\mathcal{H}} + \|A\phi - A|_{X_n}\phi_n\|_{\mathcal{H}}.$$

In [149], the assumptions are that A is invertible and

$$\sup_{v_n\in X_n} \inf_{w_n\in X_n} \frac{\left\|A^{-1}v_n - w_n\right\|_{Q(A)}}{\|v_n\|_{Q(A)}} \underset{n\to\infty}{\longrightarrow} 0. \tag{4.12}$$

Each of the conditions (4.11) and (4.12) ensures that (D1) and (D2) hold for any discrete eigenvalue of A. In the case when A is semibounded, (C3) is automatically satisfied, so that our assumptions boil down to (A1), (D1) and (D2). These three conditions are weaker than those in [73, 149, 150], and more easy to check in some settings, as will be seen in Section 4.5 on the example of the supercell method. On the other hand, when A is not semibounded, the precise relationship between condition (C3) and (4.11) and (4.12) is still unclear to us.

4.4 Proof of Theorem 4.3.1

4.4.1 Proof of (4.3)

Let us argue by contradiction and assume that there exists a subsequence $(\mathcal{T}_{n_k})_{k\in\mathbb{N}}$ and $\eta > 0$ such that $(\lambda-\eta,\lambda+\eta)\cap\sigma(A) = \{\lambda\}$ and

$$\forall k \in \mathbb{N},\ \operatorname{dist}(\lambda, \sigma(A_{n_k})) \geq \eta. \tag{4.13}$$

Let $\psi \in D(A)$ be a $\mathcal{H}$-normalized eigenvector of A associated with the discrete eigenvalue λ and $\mu := \lambda + \eta/2$. As $(\mu-\frac{\eta}{2},\mu+\frac{\eta}{2})\cap\sigma(A) = \emptyset$, it holds

$$\alpha := \min_{\nu\in\sigma(A)} \frac{|\nu-\mu|}{1+|\nu|} > 0.$$

Let us consider the auxiliary problem

$$\left\{\begin{array}{l} \text{find } u \in Q(A) \text{ such that} \\ \forall v \in Q(A),\ (a-\mu m)(u,v) = (\lambda-\mu)m(\psi,v). \end{array}\right. \tag{4.14}$$

The bilinear form $a-\mu m$ is continuous on $Q(A)$ and satisfies $\|a-\mu m\|_{\mathcal{L}(Q(A)\times Q(A))} \leq 1+|\mu|$. The linear form $f : Q(A) \ni v \mapsto (\lambda-\mu)m(\psi,v)$ is also continuous. Furthermore, as $\mu \notin \sigma(A)$, if $v \in Q(A)$ is such that $(a-\mu m)(v,w) = 0$ for all $w \in Q(A)$, then necessarily $v = 0$. Lastly,

$$\inf_{w\in Q(A)} \sup_{v\in Q(A)} \frac{|(a-\mu m)(v,w)|}{\|v\|_{Q(A)}\|w\|_{Q(A)}} \geq \min_{\nu\in\sigma(A)} \frac{|\nu-\mu|}{1+|\nu|} = \alpha.$$

Thus, applying Banach-Nečas-Babuška's theorem (see Section 4.8), problem (4.14) is well-posed. Clearly, its unique solution is $u = \psi$.

Let us now introduce the following sequence of discretized problems for $k \in \mathbb{N}$:

$$\left\{\begin{array}{l} \text{find } u_{n_k} \in X_{n_k} \text{ such that} \\ \forall v_{n_k} \in X_{n_k},\ (a_{n_k}-\mu m_{n_k})(u_{n_k},v_{n_k}) = (\lambda-\mu)m_{n_k}\left(\Pi^{Q(A)}_{X_{n_k}}\psi, v_{n_k}\right). \end{array}\right. \tag{4.15}$$

From (4.13) and assumption (A3) (for $K = \{\mu\}$ and $\alpha_K = \eta/2$), we deduce the discrete inf-sup condition

$$\forall k \in \mathbb{N},\quad \inf_{w_{n_k}\in X_{n_k}} \sup_{v_{n_k}\in X_{n_k}} \frac{|(a_{n_k}-\mu m_{n_k})(v_{n_k},w_{n_k})|}{\|v_{n_k}\|_{Q(A)}\|w_{n_k}\|_{Q(A)}} \geq c > 0.$$

Thus, by Strang's lemma (see Section 4.8) and assumptions (A2), (A3) and (A4), for all $k \in \mathbb{N}$,

$$\begin{aligned}
\|\psi-u_{n_k}\|_{Q(A)} &\leq \frac{\eta}{2c} \sup_{v_{n_k}\in X_{n_k}} \frac{|m(\psi,v_{n_k}) - m_{n_k}(\Pi^{Q(A)}_{X_{n_k}}\psi, v_{n_k})|}{\|v_{n_k}\|_{Q(A)}} \\
&+ \inf_{w_{n_k}\in X_{n_k}} \left(\frac{c+1+|\mu|}{c}\|\psi-w_{n_k}\|_{Q(A)} + \frac{1}{c}\sup_{v_{n_k}\in X_{n_k}} \frac{|[(a_{n_k}-a)+\mu(m-m_{n_k})](w_{n_k},v_{n_k})|}{\|v_{n_k}\|_{Q(A)}} \right) \\
&\leq \frac{\eta}{2c}\left(\left\|\psi - \Pi^{Q(A)}_{X_{n_k}}\psi\right\|_{\mathcal{H}} + \kappa r^m_{n_k}\left(\Pi^{Q(A)}_{X_{n_k}}\psi\right) + s^m_{n_k}(\psi)\right) + \frac{c+1+|\mu|}{c}\left\|\psi - \Pi^{Q(A)}_{X_{n_k}}\psi\right\|_{Q(A)} \\
&\quad + \frac{1}{c}\left(\kappa r^a_{n_k}\left(\Pi^{Q(A)}_{X_{n_k}}\psi\right) + s^a_{n_k}(\psi) + |\mu|\kappa r^m_{n_k}\left(\Pi^{Q(A)}_{X_{n_k}}\psi\right) + |\mu| s^m_{n_k}(\psi)\right) \\
&\leq C\left(\left\|\psi - \Pi^{Q(A)}_{X_{n_k}}\psi\right\|_{Q(A)} + r^a_{n_k}(\psi) + r^m_{n_k}(\psi) + s^a_{n_k}(\psi) + s^m_{n_k}(\psi)\right),
\end{aligned}$$

where $C \in \mathbb{R}_+$ is a constant independent of k. The above inequality implies that the sequence $(u_{n_k})_{k\in\mathbb{N}}$ strongly converges to ψ in $Q(A)$, from which we infer that

$$\lim_{k\to\infty} \|\Pi_{X_{n_k}}^{Q(A)}\psi - u_{n_k}\|_{Q(A)} = 0. \tag{4.16}$$

On the other hand, (4.15) yields

$$\forall v_{n_k} \in X_{n_k},\ (a_{n_k} - \lambda m_{n_k})(u_{n_k}, v_{n_k}) = (\lambda-\mu) m_{n_k}\left(\Pi_{X_{n_k}}^{Q(A)}\psi - u_{n_k}, v_{n_k}\right).$$

The above equality also reads

$$(A_{n_k} - \lambda)(\mathcal{M}_{n_k}^{1/2} u_{n_k}) = (\lambda-\mu)\mathcal{M}_{n_k}^{1/2}\left(\Pi_{X_{n_k}}^{Q(A)}\psi - u_{n_k}\right).$$

It then follows from (A1), (A2) and (4.16) that

$$\lim_{k\to\infty} \|(A_{n_k} - \lambda)(\mathcal{M}_{n_k}^{1/2} u_{n_k})\|_{\mathcal{H}} = 0 \qquad \text{and} \qquad \liminf_{k\to\infty} \|\mathcal{M}_{n_k}^{1/2} u_{n_k}\|_{\mathcal{H}}^2 \geq \gamma > 0,$$

which proves that $\operatorname{dist}(\lambda, \sigma(A_{n_k})) \underset{k\to\infty}{\longrightarrow} 0$ and contradicts (4.13).

4.4.2 Proof of (4.4) and (4.5)

By assumption (B1), the approximation $(\mathcal{T}_n)_{n\in\mathbb{N}}$ is such that $\overline{\lim_{n\to\infty}}\, \sigma(A_n) \cap (\lambda-\varepsilon, \lambda+\varepsilon) = \{\lambda\}$. Hence, for n large enough,

$$\sigma(A_n) \cap ((\lambda - 2\varepsilon/3, \lambda - \varepsilon/3) \cup (\lambda + \varepsilon/3, \lambda + 2\varepsilon/3)) = \emptyset,$$

so that the circle $\mathcal{C}$ in the complex plane centered at λ and of radius $\varepsilon/2$ is such that $\operatorname{dist}(\mathcal{C}, \sigma(A_n)) \geq \varepsilon/6$. This implies in particular that, for n large enough,

$$\mathcal{P}_n = \frac{1}{2i\pi}\oint_{\mathcal{C}} i_{X_n}\mathcal{M}_n^{-1/2}(z - A_n)^{-1}\mathcal{M}_n^{1/2} i_{X_n}^*\, dz.$$

Consequently, for all $\psi \in \operatorname{Ran}(\mathcal{P})$, it holds

$$(\mathcal{P} - \mathcal{P}_n)\psi = \frac{1}{2i\pi}\oint_{\mathcal{C}} \left((z-A)^{-1}\psi - i_{X_n}\mathcal{M}_n^{-1/2}(z-A_n)^{-1}\mathcal{M}_n^{1/2} i_{X_n}^*\psi\right) dz.$$

In the following, C will denote a constant independent of $n \in \mathbb{N}^*$ and $z \in \mathcal{C}$, which may change along the calculations.

For $z \in \mathcal{C}$, we introduce the auxiliary problem

$$\begin{cases} \text{find } u^z \in Q(A) \text{ such that} \\ \forall v \in Q(A),\ (zm - a)(u^z, v) = m(\psi, v), \end{cases}$$

whose unique solution is $u^z = (z-A)^{-1}\psi = \frac{\psi}{z-\lambda}$, since $\psi \in \operatorname{Ran}(\mathcal{P})$. We also introduce the discretized problem

$$\begin{cases} \text{find } u_n^z \in X_n \text{ such that} \\ \forall v_n \in X_n,\ (zm_n - a_n)(u_n^z, v_n) = m_n\left(\Pi_{X_n}^{\mathcal{H}}\psi, v_n\right), \end{cases}$$

whose unique solution is $u_n^z = i_{X_n}\mathcal{M}_n^{-1/2}(z-A_n)^{-1}\mathcal{M}_n^{1/2} i_{X_n}^*\psi$. From assumption (A3), since $\mathcal{C}$ is a compact subset of $\mathbb{C}$, there exists $c > 0$ such that for all $z \in \mathcal{C}$ and $n \in \mathbb{N}$,

$$\inf_{w_n\in X_n}\sup_{v_n\in X_n} \frac{|(a_n - z m_n)(v_n, w_n)|}{\|v_n\|_{Q(A)}\|w_n\|_{Q(A)}} \geq c.$$

Reasoning as in Section 4.4.1, we infer from Strang's lemma, assumptions (A2)-(A4) and the fact that r_n^a, r_n^m, s_n^a and s_n^m are semi-norms, that for all $z \in \mathcal{C}$,

$$\begin{aligned}
\|u^z - u_n^z\|_{Q(A)} &\leq \frac{1}{c} \sup_{v_n \in X_n} \frac{|m(\psi, v_n) - m_n(\Pi_{X_n}^{\mathcal{H}}\psi, v_n)|}{\|v_n\|_{Q(A)}} \\
&\quad + \inf_{w_n \in X_n} \left(\frac{c+1+|z|}{c} \|u^z - w_n\|_{Q(A)} + \frac{1}{c} \sup_{v_n \in X_n} \frac{|[(a_n - a) + z(m_n - m)](w_n, v_n)|}{\|v_n\|_{Q(A)}} \right) \\
&\leq \frac{\kappa}{c} r_n^m(\psi) + \frac{\Gamma}{c} \left\|\psi - \Pi_{X_n}^{\mathcal{H}}\psi\right\|_{\mathcal{H}} + \frac{1}{c} s_n^m(\psi) + \frac{\Gamma}{c} \left\|\Pi_{X_n}^{Q(A)}\psi - \Pi_{X_n}^{\mathcal{H}}\psi\right\|_{\mathcal{H}} \\
&\quad + \frac{c+1+|z|}{c} \left\|u^z - \Pi_{X_n}^{Q(A)} u^z\right\|_{Q(A)} \\
&\quad + \frac{1}{c} \left(\kappa r_n^a \left(\Pi_{X_n}^{Q(A)} u^z\right) + s_n^a(u^z) + |z| \kappa r_n^m \left(\Pi_{X_n}^{Q(A)} u^z\right) + |z| s_n^m(u^z) \right) \\
&\leq \frac{\kappa}{c} r_n^m(\psi) + \frac{3\Gamma}{c} \left\|\psi - \Pi_{X_n}^{Q(A)}\psi\right\|_{Q(A)} + \frac{c + (1+|z|)(1+\kappa^2)}{c} \left\|u^z - \Pi_{X_n}^{Q(A)} u^z\right\|_{Q(A)} \\
&\quad + \frac{1}{c} \left(\kappa r_n^a(u^z) + s_n^a(u^z) + |z| \kappa r_n^m(u^z) + |z| s_n^m(u^z) \right) \\
&\leq C \left(\left\| \left(1 - \Pi_{X_n}^{Q(A)}\right) \psi \right\|_{Q(A)} + r_n^a(\psi) + r_n^m(\psi) + s_n^a(\psi) + s_n^m(\psi) \right),
\end{aligned}$$

since $u^z = \frac{\psi}{z - \lambda}$. Thus, for all $z \in \mathcal{C}$,

$$\begin{aligned}
&\left\| (z - A)^{-1}\psi - i_{X_n} \mathcal{M}_n^{-1/2} (z - A_n)^{-1} \mathcal{M}_n^{1/2} i_{X_n}^* \psi \right\|_{Q(A)} \\
&\qquad \leq C \left(\left\| \left(1 - \Pi_{X_n}^{Q(A)}\right) \psi \right\|_{Q(A)} + r_n^a(\psi) + r_n^m(\psi) + s_n^a(\psi) + s_n^m(\psi) \right).
\end{aligned}$$

Since $\mathcal{C}$ is of finite length, we obtain that, for n large enough, for all $\psi \in \mathrm{Ran}(\mathcal{P})$,

$$\|(\mathcal{P} - \mathcal{P}_n)\psi\|_{Q(A)} \leq C \left(\left\| \left(1 - \Pi_{X_n}^{Q(A)}\right) \psi \right\|_{Q(A)} + r_n^a(\psi) + r_n^m(\psi) + s_n^a(\psi) + s_n^m(\psi) \right),$$

which readily leads to (4.5).

Let us finally consider a $\mathcal{H}$-orthonormal basis $(\zeta_1, \cdots, \zeta_q)$ of $\mathrm{Ran}(\mathcal{P}) = \mathrm{Ker}(\lambda - A)$. Since for all $1 \leq i \leq q$, $\mathcal{P}_n \zeta_i \underset{n \to \infty}{\longrightarrow} \mathcal{P}\zeta_i = \zeta_i$ strongly in $\mathcal{H}$, the family $(\mathcal{P}_n\zeta_1, \cdots, \mathcal{P}_n\zeta_q)$ is free for n large enough, so that $\mathrm{Rank}(\mathcal{P}_n) \geq q$.

4.4.3 Proof of (4.6) and (4.7)

We just have shown that for all $1 \leq i \leq q$, $\mathcal{P}_n \zeta_i \underset{n \to \infty}{\longrightarrow} \mathcal{P}\zeta_i = \zeta_i$ strongly in $\mathcal{H}$. Under the additional assumption that, for n large enough, $\mathrm{Rank}(\mathcal{P}_n) = q$, this implies that there exists $n_0 \in \mathbb{N}$, such that, for $n \geq n_0$, $(\mathcal{P}_n\zeta_1, \cdots, \mathcal{P}_n\zeta_q)$ forms a basis of $\mathrm{Ran}(\mathcal{P}_n)$, with

$$\min_{1 \leq i \leq q} \|\mathcal{P}_n \zeta_i\|_{\mathcal{H}}^2 \geq \frac{3}{4} \quad \text{and} \quad \max_{1 \leq i,j \leq q,\, i \neq j} |m(\mathcal{P}_n\zeta_i, \zeta_j)| \leq \frac{1}{4q}.$$

Thus, any $\xi_n \in \mathrm{Ran}(\mathcal{P}_n)$ can be decomposed as

$$\xi_n = \sum_{i=1}^{q} \alpha^i(\xi_n) \mathcal{P}_n \zeta_i,$$

the coefficients $(\alpha^1(\xi_n),\cdots,\alpha^q(\xi_n))$ of ξ_n in the basis $(\mathcal{P}_n\zeta_1,\cdots,\mathcal{P}_n\zeta_q)$ being such that

$$\max_{1\le i\le q}|\alpha^i(\xi_n)|\le 2\|\xi_n\|_{\mathcal{H}}.$$

We have

$$\begin{aligned}\mathcal{P}\xi_n-\xi_n &= \sum_{i=1}^q \alpha^i(\xi_n)\left(\sum_{j=1}^q m\left(\mathcal{P}_n\zeta_i,\zeta_j\right)\zeta_j-\mathcal{P}_n\zeta_i\right)\\ &= \sum_{i=1}^q \alpha^i(\xi_n)\left(\sum_{j\neq i} m\left(\mathcal{P}_n\zeta_i-\zeta_i,\zeta_j\right)\zeta_j-\left(\mathcal{P}_n\zeta_i-\zeta_i\right)+m\left(\mathcal{P}_n\zeta_i-\zeta_i,\zeta_i\right)\zeta_i\right),\end{aligned}$$

and we deduce from (4.5) that for all $1\le i\le q$,

$$\|\zeta_i-\mathcal{P}_n\zeta_i\|_{Q(A)}\le C\left(\left\|\left(1-\Pi_{X_n}^{Q(A)}\right)\mathcal{P}\right\|_{\mathcal{L}(\mathcal{H},Q(A))}+\mathcal{R}_n^a+\mathcal{R}_n^m+\mathcal{S}_n^a+\mathcal{S}_n^m\right).$$

Hence,

$$\forall\xi_n\in\mathrm{Ran}(\mathcal{P}_n),\quad \|\mathcal{P}\xi_n-\xi_n\|_{Q(A)}\le C\left(\left\|\left(1-\Pi_{X_n}^{Q(A)}\right)\mathcal{P}\right\|_{\mathcal{L}(\mathcal{H},Q(A))}+\mathcal{R}_n^a+\mathcal{R}_n^m+\mathcal{S}_n^a+\mathcal{S}_n^m\right)\|\xi_n\|_{\mathcal{H}},$$

where the constant C is independent of n. Besides, it also follows from (A2) and the definition of $\mathcal{P}_n$ that

$$\forall n\in\mathbb{N},\quad \|\mathcal{P}_n\|_{\mathcal{L}(\mathcal{H})}\le\sqrt{\frac{\Gamma}{\gamma}}.$$

Therefore,

$$\begin{aligned}\|(\mathcal{P}-\mathcal{P}_n)\mathcal{P}_n\|_{\mathcal{L}(\mathcal{H},Q(A))} &\le \sup_{\xi_n\in\mathrm{Ran}(\mathcal{P}_n)\setminus\{0\}}\frac{\|\mathcal{P}\xi_n-\xi_n\|_{Q(A)}}{\|\xi_n\|_{\mathcal{H}}}\|\mathcal{P}_n\|_{\mathcal{L}(\mathcal{H})}\\ &\le C\left(\left\|\left(1-\Pi_{X_n}^{Q(A)}\right)\mathcal{P}\right\|_{\mathcal{L}(\mathcal{H},Q(A))}+\mathcal{R}_n^a+\mathcal{R}_n^m+\mathcal{S}_n^a+\mathcal{S}_n^m\right),\end{aligned}$$

and (4.6) is proved.

For each n large enough, let $(\psi_n,\lambda_n)\in X_n\times\mathbb{R}$ be a solution to the generalized eigenvalue problem (4.2) such that $\lambda_n\in(\lambda-\varepsilon/2,\lambda+\varepsilon/2)$, $\phi_n=\frac{\psi_n}{\|\psi_n\|_{\mathcal{H}}}$, and $\chi_n=\frac{\mathcal{P}\psi_n}{\|\mathcal{P}\psi_n\|_{\mathcal{H}}}=\frac{\mathcal{P}\phi_n}{\|\mathcal{P}\phi_n\|_{\mathcal{H}}}$. It follows from (4.6) that

$$\|\mathcal{P}\phi_n-\phi_n\|_{\mathcal{H}}\le\|\mathcal{P}\phi_n-\phi_n\|_{Q(A)}\le C\left(\left\|\left(1-\Pi_{X_n}^{Q(A)}\right)\mathcal{P}\right\|_{\mathcal{L}(\mathcal{H},Q(A))}+\mathcal{R}_n^a+\mathcal{R}_n^m+\mathcal{S}_n^a+\mathcal{S}_n^m\right)\underset{n\to\infty}{\longrightarrow}0,$$

from which we infer that $\|\mathcal{P}\phi_n\|_{\mathcal{H}}\to 1$, $(\phi_n)_{n\in\mathbb{N}}$ is bounded in $Q(A)$, $\|\phi_n-\chi_n\|_{Q(A)}\to 0$, and

$$\begin{aligned}\|\chi_n-\phi_n\|_{Q(A)} &\le \left\|\frac{\mathcal{P}\phi_n}{\|\mathcal{P}\phi_n\|_{\mathcal{H}}}-\mathcal{P}\phi_n\right\|_{Q(A)}+\|\mathcal{P}\phi_n-\phi_n\|_{Q(A)}\\ &\le \|\mathcal{P}\phi_n\|_{\mathcal{H}}^{-1}\|\phi_n-\mathcal{P}\phi_n\|_{\mathcal{H}}\|\mathcal{P}\phi_n\|_{Q(A)}+\|\mathcal{P}\phi_n-\phi_n\|_{Q(A)}\\ &\le C\left(\left\|\left(1-\Pi_{X_n}^{Q(A)}\right)\mathcal{P}\right\|_{\mathcal{L}(\mathcal{H},Q(A))}+\mathcal{R}_n^a+\mathcal{R}_n^m+\mathcal{S}_n^a+\mathcal{S}_n^m\right).\end{aligned}$$

Besides, it holds

$$\begin{aligned}|\lambda_n-\lambda| &= |a_n(\psi_n,\psi_n)-a(\chi_n,\chi_n)|\\ &\le \left|\frac{a_n(\phi_n,\phi_n)}{m_n(\phi_n,\phi_n)}-a(\phi_n,\phi_n)\right|+|a(\phi_n,\phi_n)-a(\chi_n,\chi_n)|.\end{aligned}$$

On the one hand, we have

$$\begin{aligned}|a(\phi_n,\phi_n)-a(\chi_n,\chi_n)| &= |a(\phi_n-\chi_n,\phi_n-\chi_n)+2a(\chi_n,\phi_n-\chi_n)|\\ &= |a(\phi_n-\chi_n,\phi_n-\chi_n)+2\lambda m(\chi_n,\phi_n-\chi_n)|\\ &= |a(\phi_n-\chi_n,\phi_n-\chi_n)-\lambda\|\chi_n-\phi_n\|^2_{\mathcal{H}}|\\ &\le C\|\phi_n-\chi_n\|^2_{Q(A)}.\end{aligned}$$

On the other hand,

$$\begin{aligned}|(a-a_n)(\phi_n,\phi_n)| &\le |(a-\widetilde{a}_n)(\phi_n,\phi_n)|+|(\widetilde{a}_n-a_n)(\phi_n,\phi_n)|\\ &\le r_n^a(\phi_n)^2+\left|(\widetilde{a}_n-a_n)\left(\phi_n-\Pi_{X_n}^{Q(A)}\chi_n,\phi_n-\Pi_{X_n}^{Q(A)}\chi_n\right)\right|\\ &\quad +2\left|(\widetilde{a}_n-a_n)\left(\Pi_{X_n}^{Q(A)}\chi_n,\phi_n-\Pi_{X_n}^{Q(A)}\chi_n\right)\right|\\ &\quad +\left|(\widetilde{a}_n-a_n)\left(\Pi_{X_n}^{Q(A)}\chi_n,\Pi_{X_n}^{Q(A)}\chi_n\right)\right|\\ &\le \left(r_n^a(\chi_n)+\kappa\|\phi_n-\chi_n\|_{Q(A)}\right)^2+(\Gamma+\kappa^2)\left\|\phi_n-\Pi_{X_n}^{Q(A)}\chi_n\right\|^2_{Q(A)}\\ &\quad +\left(2\left\|\phi_n-\Pi_{X_n}^{Q(A)}\chi_n\right\|_{Q(A)}+\left\|\Pi_{X_n}^{Q(A)}\chi_n\right\|_{Q(A)}\right)s_n^a(\chi_n)\\ &\le C\left[\left(r_n^a(\chi_n)+\|\phi_n-\chi_n\|_{Q(A)}+\left\|\left(1-\Pi_{X_n}^{Q(A)}\right)\chi_n\right\|_{Q(A)}\right)^2+s_n^a(\chi_n)\right]\\ &\le C\left[\left(\left\|\left(1-\Pi_{X_n}^{Q(A)}\right)\mathcal{P}\right\|_{\mathcal{L}(\mathcal{H},Q(A))}+\mathcal{R}_n^a+\mathcal{R}_n^m+\mathcal{S}_n^a+\mathcal{S}_n^m\right)^2+\mathcal{S}_n^a\right],\end{aligned}$$

and a similar calculation leads to

$$\begin{aligned}|m_n(\phi_n,\phi_n)-1| &= |m_n(\phi_n,\phi_n)-m(\phi_n,\phi_n)|,\\ &\le C\left[\left(\left\|\left(1-\Pi_{X_n}^{Q(A)}\right)\mathcal{P}\right\|_{\mathcal{L}(\mathcal{H},Q(A))}+\mathcal{R}_n^a+\mathcal{R}_n^m+\mathcal{S}_n^a+\mathcal{S}_n^m\right)^2+\mathcal{S}_n^m\right].\end{aligned}$$

Consequently,

$$\begin{aligned}\left|\frac{a_n(\phi_n,\phi_n)}{m_n(\phi_n,\phi_n)}-a(\phi_n,\phi_n)\right| &\le \frac{|(a-a_n)(\phi_n,\phi_n)|}{m_n(\phi_n,\phi_n)}+|a(\phi_n,\phi_n)|\left|\frac{m_n(\phi_n,\phi_n)-1}{m_n(\phi_n,\phi_n)}\right|\\ &\le \gamma^{-1}\left(|(a-a_n)(\phi_n,\phi_n)|+|a(\phi_n,\phi_n)|\,|m_n(\phi_n,\phi_n)-1|\right)\\ &\le C\left[\left(\left\|\left(1-\Pi_{X_n}^{Q(A)}\right)\mathcal{P}\right\|_{\mathcal{L}(\mathcal{H},Q(A))}+\mathcal{R}_n^a+\mathcal{R}_n^m+\mathcal{S}_n^a+\mathcal{S}_n^m\right)^2+\mathcal{S}_n^a+\mathcal{S}_n^m\right].\end{aligned}$$

Collecting the above results, we obtain

$$|\lambda-\lambda_n|\le C\left[\left(\left\|\left(1-\Pi_{X_n}^{Q(A)}\right)\mathcal{P}\right\|_{\mathcal{L}(\mathcal{H},Q(A))}+\mathcal{R}_n^a+\mathcal{R}_n^m\right)^2+\mathcal{S}_n^a+\mathcal{S}_n^m\right],$$

which proves estimate (4.7).

4.5 Application to the supercell method

The aim of this section is to show that the theoretical framework presented in Section 4.3 can be applied to the numerical analysis of the supercell method for perturbed periodic Schrödinger operators.

Note that the supercell method was previously studied from a mathematical viewpoint by Soussi [172], for the special case of a two-dimensional periodic Schrödinger operator in the presence of a compactly supported perturbation W of the form $W(x) = w\mathbb{1}_\Omega(x)$, where w is a real constant and Ω a bounded domain of $\mathbb{R}^2$.

4.5.1 The supercell method with exact integration

Let $\mathcal{R}$ be a periodic lattice of $\mathbb{R}^d$, $\mathcal{R}^*$ its reciprocal lattice and Γ a unit cell of $\mathcal{R}$ such that 0 is in the interior of Γ. Typically, in the case of the cubic lattice $\mathcal{R} = \mathbb{Z}^d$, $\mathcal{R}^* = 2\pi\mathbb{Z}^d$ and $\Gamma = (-1/2, 1/2]^d$ is an admissible unit cell.

Let us introduce the perturbed periodic Schrödinger operator

$$A := -\Delta + V_{\rm per} + W,$$

where Δ is the Laplace operator, $V_{\rm per}$ a real-valued $\mathcal{R}$-periodic function of $L^p_{\rm loc}(\mathbb{R}^d)$, with $p = 2$ if $d \leq 3$, $p > 2$ if $d = 4$ and $p = d/2$ for $d \geq 5$, and $W \in L^\infty(\mathbb{R}^d)$ a real-valued function such that $W(x) \underset{|x|\to\infty}{\longrightarrow} 0$.

The operator A is self-adjoint and bounded from below on $\mathcal{H} := L^2(\mathbb{R}^d)$, endowed with its natural inner product

$$\forall \phi, \psi \in \mathcal{H}, \quad m(\phi, \psi) := \int_{\mathbb{R}^d} \phi\psi,$$

with domain $D(A) = H^2(\mathbb{R}^d)$ and form domain $Q(A) = H^1(\mathbb{R}^d)$. The associated bilinear form $a(\cdot,\cdot)$ is defined by

$$\forall \phi, \psi \in Q(A), \quad a(\phi, \psi) := \int_{\mathbb{R}^d} \nabla\phi \cdot \nabla\psi + \int_{\mathbb{R}^d} (V_{\rm per} + W)\phi\psi.$$

We denote by $A^0 := -\Delta + V_{\rm per}$ the corresponding periodic Schrödinger operator on $L^2(\mathbb{R}^d)$.

The supercell method is the current state-of-the-art technique in solid state physics to compute the spectrum of the operator A. For $L \in \mathbb{N}^*$, we denote by $\Gamma_L := L\Gamma$ the supercell of size L and set

$$\begin{aligned}
L^2_{\rm per}(\Gamma_L) &:= \left\{ u_L \in L^2_{\rm loc}(\mathbb{R}^d) \mid u_L \ L\mathcal{R}\text{-periodic} \right\}, \\
H^1_{\rm per}(\Gamma_L) &:= \left\{ u_L \in L^2_{\rm per}(\Gamma_L) \mid \nabla u_L \in \left(L^2_{\rm per}(\Gamma_L)\right)^d \right\}, \\
C^0_{\rm per}(\Gamma_L) &:= \left\{ u_L \in C^0(\mathbb{R}^d) \mid u_L \ L\mathcal{R}\text{-periodic} \right\}, \\
L^\infty_{\rm per}(\Gamma_L) &:= \left\{ u_L \in L^\infty(\mathbb{R}^d) \mid u_L \ L\mathcal{R}\text{-periodic} \right\}.
\end{aligned}$$

For $u_L \in L^2_{\rm per}(\Gamma_L)$ and $k \in L^{-1}\mathcal{R}^*$, we denote by

$$\widehat{u}_L(k) := \frac{1}{|\Gamma_L|^{1/2}} \int_{\Gamma_L} u_L(x) e^{ik\cdot x}\, dx$$

the Fourier coefficient of u_L corresponding to the k mode. For $r \in \mathbb{R}$, the Sobolev space $H^r_{\rm per}(\Gamma_L)$ can be defined as

$$H^r_{\rm per}(\Gamma_L) := \left\{ u_L \in L^2_{\rm per}(\Gamma_L) \mid \sum_{k \in L^{-1}\mathcal{R}^*} \left(1+|k|^2\right)^r |\widehat{u}_L(k)|^2 < \infty \right\}.$$

The supercell method relies on the resolution of the following (non-consistent and non-conforming) eigenvalue problem:

$$\left\{ \begin{array}{l} \text{find } (u_{L,N}, \lambda_{L,N}) \in Y_{L,N} \times \mathbb{R} \text{ such that} \\ \forall v_{L,N} \in Y_{L,N},\ \widehat{a}_L(u_{L,N}, v_{L,N}) = \lambda_{L,N} \widehat{m}_L\left(u_{L,N}, v_{L,N}\right), \end{array} \right.$$

where

$$\begin{aligned} \forall u_L, v_L \in L^2_{\rm per}(\Gamma_L), \qquad & \widehat{m}_L\left(u_L, v_L\right) := \int_{\Gamma_L} u_L v_L, \\ \forall u_L, v_L \in H^1_{\rm per}(\Gamma_L), \qquad & \widehat{a}_L(u_L, v_L) := \int_{\Gamma_L} \nabla u_L \cdot \nabla v_L + \int_{\Gamma_L} (V_{\rm per} + W) u_L v_L, \end{aligned}$$

and $Y_{L,N}$ is a finite dimensional subspace of $H^1_{\rm per}(\Gamma_L)$.

We set $H_{L,N} = H_L|_{Y_{L,N}}$, where H_L denotes the unique self-adjoint operator on $L^2_{\rm per}(\Gamma_L)$ associated with the quadratic form $\widehat{a}_L$. We have $D(H_L) = H^2_{\rm per}(\Gamma_L)$ and

$$\forall u_L \in H^2_{\rm per}(\Gamma_L), \quad H_L u_L = -\Delta u_L + (V_{\rm per} + W_L) u_L,$$

where $W_L \in L^\infty_{\rm per}(\Gamma_L)$ denotes the $L\mathcal{R}$-periodic extension of $W|_{\Gamma_L}$.

For the sake of clarity, our analysis will be restricted to the case of the cubic lattice $\mathcal{R} = \mathbb{Z}^d$ and the planewave discretization method, for which

$$Y_{L,N} := \left\{ \sum_{k \in L^{-1}\mathcal{R}^* \mid |k| \le 2\pi N L^{-1}} c_k e_{L,k} \mid \forall k,\ c_{-k} = c_k^* \right\},$$

where $e_{L,k}(x) := |\Gamma_L|^{-1/2} e^{ik\cdot x}$. We denote by $\Pi_{L,N}$ the orthogonal projection of $L^2_{\rm per}(\Gamma_L)$ on $Y_{N,L}$ for the $L^2_{\rm per}(\Gamma_L)$ inner product (actually $\Pi_{L,N}$ is also the orthogonal projection of $H^s_{\rm per}(\Gamma_L)$ on $Y_{N,L}$ for the $H^s_{\rm per}(\Gamma_L)$ inner product, for any $s \in \mathbb{R}$).

The discretization spaces $Y_{L,N}$ possess the following properties:

$$\forall u_{L,N} \in Y_{L,N},\ \Pi_{L,N}\left(-\Delta u_{L,N}\right) = -\Delta u_{L,N},$$

and for all real numbers r and s such that $0 \le r \le s$, there exists a constant $C \in \mathbb{R}_+$ such that for all $L \in \mathbb{N}^*$ and all $u_L \in H^s_{\rm per}(\Gamma_L)$,

$$\|u_L - \Pi_{L,N} u_L\|_{H^r_{\rm per}(\Gamma_L)} \le C \left(\frac{L}{N}\right)^{s-r} \|u_L\|_{H^s_{\rm per}(\Gamma_L)}. \tag{4.17}$$

As in [40], we will assume that $V_{\rm per}$ belongs to the functional space $\mathcal{Z}_{\rm per}(\Gamma)$ (denoted by $\mathcal{M}_{\rm per}(\Gamma)$ in [40]), defined by

$$\mathcal{Z}_{\rm per}(\Gamma) := \left\{ V \in L^2_{\rm per}(\Gamma) \mid \|V\|_{\mathcal{Z}_{\rm per}(\Gamma)} := \sup_{L \in \mathbb{N}^*} \sup_{w_L \in H^1_{\rm per}(\Gamma_L)\setminus\{0\}} \frac{\|V w_L\|_{L^2_{\rm per}(\Gamma_L)}}{\|w_L\|_{H^1_{\rm per}(\Gamma_L)}} < +\infty \right\}.$$

The space $\mathcal{Z}_{\rm per}(\Gamma)$ is a normed space and the space of the $\mathcal{R}$-periodic functions of class C^∞ is dense in $\mathcal{Z}_{\rm per}(\Gamma)$.

Our main result concerning the supercell method in the absence of numerical integration error is the following:

Theorem 4.5.1. *Assume that* $V_{\rm per} \in \mathcal{Z}_{\rm per}(\Gamma)$ *and that* $W \in L^\infty(\mathbb{R}^d)$ *with* $W(x) \underset{|x|\to\infty}{\longrightarrow} 0$. *Let* $(N_L)_{L\in\mathbb{N}^*}$ *be a sequence of integers such that* $\dfrac{N_L}{L} \underset{L\to\infty}{\longrightarrow} +\infty$. *Then,*

1. ***Absence of pollution***

$$\lim_{L\to\infty} \sigma(H_{L,N_L}) = \sigma(A). \tag{4.18}$$

2. ***A priori error estimates***

Assume that, in addition, $V_{\rm per} \in H^{r-2}_{\rm per}(\Gamma)$ *and* $W \in H^{r-2}(\mathbb{R}^d)$, *for some* $r \geq 2$. *Let* λ *be a discrete eigenvalue of* A *and* $\varepsilon > 0$ *be such that* $\sigma(A) \cap (\lambda-\varepsilon, \lambda+\varepsilon) = \{\lambda\}$. *Let* $\mathcal{P} := \mathbb{1}_{\{\lambda\}}(A)$ *be the* $L^2(\mathbb{R}^d)$*-orthogonal projection onto the eigenspace of* A *associated with* λ *and* $\mathfrak{P}_L := \mathbb{1}_{(\lambda-\varepsilon/2,\lambda+\varepsilon/2)}(H_{L,N_L})$ *the* $L^2_{\rm per}(\Gamma_L)$*-orthogonal spectral projection of* H_{L,N_L} *associated with the eigenvalues belonging to the interval* $(\lambda-\varepsilon/2, \lambda+\varepsilon/2)$. *Consider finally a sequence of cut-off functions* $(\chi_L)_{L\in\mathbb{N}^*}$ *such that*

$$0 \leq \chi_L \leq 1 \text{ on } \mathbb{R}^d,\ \chi_L = 1 \text{ on } \Gamma_L,\ \mathrm{Supp}(\chi_L) \subset (L+\sqrt{L})\Gamma,\ \|\nabla\chi_L\|_{L^\infty} \leq c, \tag{4.19}$$

for some constant $c \in \mathbb{R}_+$ *independent of* $L \in \mathbb{N}^*$.

Then, $\mathrm{Ran}(\mathcal{P}) \subset H^r(\mathbb{R}^d)$, *and there exists* $C, \delta > 0$ *such that for* L *large enough,*

$$\mathrm{Tr}(\mathcal{P}) = \mathrm{Tr}(\mathfrak{P}_L), \tag{4.20}$$

$$\sup_{\psi\in\mathrm{Ran}(\mathcal{P}),\ \|\psi\|_{L^2(\mathbb{R}^d)}=1} \inf_{u_L\in\mathrm{Ran}(\mathfrak{P}_L)} \|\psi - \chi_L u_L\|_{H^1(\mathbb{R}^d)} \leq C\left(e^{-\delta L} + \left(\frac{L}{N_L}\right)^{r-1}\right), \tag{4.21}$$

$$\sup_{u_L\in\mathrm{Ran}(\mathfrak{P}_L),\ \|u_L\|_{L^2_{\rm per}(\Gamma_L)}=1} \inf_{\psi\in\mathrm{Ran}(\mathcal{P})} \|\psi - \chi_L u_L\|_{H^1(\mathbb{R}^d)} \leq C\left(e^{-\delta L} + \left(\frac{L}{N_L}\right)^{r-1}\right), \tag{4.22}$$

$$\max_{\lambda_L\in\sigma(H_{L,N_L})\cap(\lambda-\varepsilon/2,\lambda+\varepsilon/2)} |\lambda_L - \lambda| \leq C\left(e^{-\delta L} + \left(\frac{L}{N_L}\right)^{r-1}\right)^2. \tag{4.23}$$

4.5.2 The supercell method with numerical integration

In general, the computation of the integral $\int_{\Gamma_L}(V_{\rm per}+W)u_L v_L$ with $u_L, v_L \in Y_{L,N_L}$ cannot be carried out explicitly, and a numerical integration procedure is needed. We assume in this section that $V_{\rm per}$ and W are continuous functions.

For $M \in \mathbb{N}^*$ and $u_L \in C^0_{\rm per}(\Gamma_L)$, we denote by $\widehat{u}_L^{FFT,M}$ the discrete Fourier transform of u_L on the cartesian grid $\mathcal{G}_{L,M} := \frac{L}{M}\mathbb{Z}^d$. Recall that if

$$u_L = \sum_{k\in L^{-1}\mathcal{R}^*} \widehat{u}_L(k) e_{L,k},$$

the discrete Fourier transform of u_L is the $ML^{-1}\mathcal{R}^*$-periodic sequence $\widehat{u}_L^{FFT,M} = \left(\widehat{u}_L^{FFT,M}(k)\right)_{k\in L^{-1}\mathcal{R}^*}$ where

$$\widehat{u}_L^{FFT,M}(k) = \frac{1}{M^d} \sum_{x\in\mathcal{G}_{L,M}\cap\Gamma_L} u_L(x) e^{-ik\cdot x} = |\Gamma_L|^{-1/2} \sum_{K\in L^{-1}\mathcal{R}^*} \widehat{u}_L(k+MK).$$

We now introduce the subspaces

$$W_{L,M}^{1D} := \left| \begin{array}{ll} \mathrm{Span}\left\{ e^{ily} \mid l \in 2\pi L^{-1}\mathbb{Z},\ |l| \leq \dfrac{2\pi}{L}\left(\dfrac{M-1}{2}\right)\right\} & (M \text{ odd}), \\ \mathrm{Span}\left\{ e^{ily} \mid l \in 2\pi L^{-1}\mathbb{Z},\ |l| \leq \dfrac{2\pi}{L}\left(\dfrac{M-1}{2}\right)\right\} \oplus \mathbb{C}\left(e^{i\pi My/L} + e^{-i\pi My/L}\right) & (M \text{ even}), \end{array} \right.$$

and denote by $W_{L,M}$ the d-tensor product space $W_{L,M} := W_{L,M}^{1D} \otimes \cdots \otimes W_{L,M}^{1D}$. In particular, when M is odd,

$$W_{L,M} = \mathrm{Span}\left\{ e_{L,k},\ k \in L^{-1}\mathcal{R}^*,\ |k|_\infty \leq 2\pi L^{-1}\left(\frac{M-1}{2}\right)\right\}.$$

It is then possible to define the interpolation projector $\mathcal{I}_{L,M}$ from $C^0_{\mathrm{per}}(\Gamma_L)$ onto $W_{L,M}$ by $[\mathcal{I}_{L,M}(u_L)](x) = u_L(x)$ for all $x \in \mathcal{G}_{L,M}$. In particular, when M is odd, we have the simple relation

$$\mathcal{I}_{L,M}(u_L) = |\Gamma_L|^{1/2} \sum_{k \in L^{-1}\mathcal{R}^* \mid |k|_\infty \leq 2\pi L^{-1}\left(\frac{M-1}{2}\right)} \widehat{u}_L^{FFT,M}(k) e_{L,k}.$$

It is easy to check that if the function u_L is real-valued, then so is the function $\mathcal{I}_{L,M}(u_L)$.

Besides, when $M \geq 4N+1$, it holds that for all $u_L, v_L \in Y_{L,N}$,

$$\int_{\Gamma_L} \mathcal{I}_{L,M}(V_L u_L v_L) = \int_{\Gamma_L} \mathcal{I}_{L,M}(V_L)\, u_L v_L,$$

for any $V_L \in L^2_{\mathrm{per}}(\Gamma_L)$.

The supercell method with numerical integration then consists in considering the following eigenvalue problem for a given $M \geq 4N+1$,

$$\left\{ \begin{array}{l} \text{find } (u_{L,N}, \lambda_{L,N}) \in Y_{L,N} \times \mathbb{R} \text{ such that} \\ \forall v_{L,N} \in Y_{L,N},\ \widehat{a}_{L,M}(u_{L,N}, v_{L,N}) = \lambda_{L,N} \widehat{m}_L(u_{L,N}, v_{L,N}), \end{array} \right.$$

where

$$\forall u_L, v_L \in H^1_{\mathrm{per}}(\Gamma_L), \qquad \widehat{a}_{L,M}(u_L, v_L) := \int_{\Gamma_L} \nabla u_L \cdot \nabla v_L + \int_{\Gamma_L} \mathcal{I}_{L,M}(V_{\mathrm{per}} + \widetilde{W}_L) u_L v_L,$$

and where $\widetilde{W}_L$ is the $L\mathcal{R}$-periodic extension of $\xi_L W|_{\Gamma_L}$, ξ_L being a $C^{[r-1]}(\mathbb{R}^d)$ cut-off function such that $0 \leq \xi_L \leq 1$, $\xi_L = 1$ on Γ_{L-1}, $\mathrm{Supp}(\xi_L) \subset (L-1/2)\,\Gamma$, and the sequences $\left(\|\partial^\alpha \xi_L\|_{L^\infty(\mathbb{R}^d)}\right)_{L \in \mathbb{N}^*}$ are uniformly bounded in L, for all $|\alpha| \leq [r-1]$ (here and above, $[r-1]$ denotes the integer part of $r-1$).

As in the preceding section, we denote by $H_{L,N,M} = H_{L,M}|_{Y_{L,N}}$, where $H_{L,M}$ is the unique self-adjoint operator on $L^2_{\mathrm{per}}(\Gamma_L)$ with domain $D(H_{L,M}) = H^2_{\mathrm{per}}(\Gamma_L)$ associated with the quadratic form $\widehat{a}_{L,M}$.

Theorem 4.5.2. *Let $(N_L)_{L\in\mathbb{N}^*}$ and $(G_L)_{L\in\mathbb{N}^*}$ be sequences of integers such that $\frac{N_L}{L} \underset{L\to\infty}{\longrightarrow} +\infty$ and $G_L \underset{L\to\infty}{\longrightarrow} +\infty$, and $M_L := LG_L$. We assume that $V_{\mathrm{per}} \in C^0_{\mathrm{per}}(\Gamma) \cap H^{r-2}_{\mathrm{per}}(\Gamma)$ and $W \in C^0(\mathbb{R}^d) \cap H^{r-2}(\mathbb{R}^d)$ for some $r > 2$. Then,*

1. ***Absence of pollution***

$$\lim_{L\to\infty} \sigma(H_{L,N_L,M_L}) = \sigma(A). \tag{4.24}$$

2. ***A priori error estimates***

Let λ be a discrete eigenvalue of A and $\varepsilon > 0$ be such that $\sigma(A) \cap (\lambda-\varepsilon, \lambda+\varepsilon) = \{\lambda\}$. Let $\mathcal{P} := \mathbb{1}_{\{\lambda\}}(A)$ be the $L^2(\mathbb{R}^d)$-orthogonal spectral projection onto the eigenspace of A associated with λ, and $\mathfrak{P}_L := \mathbb{1}_{(\lambda-\varepsilon/2,\lambda+\varepsilon/2)}(H_{L,N_L,M_L})$ the $L^2_{\rm per}(\Gamma_L)$-orthogonal spectral projection of H_{L,N_L,M_L} associated with the eigenvalues belonging to the interval $(\lambda-\varepsilon/2, \lambda+\varepsilon/2)$. We finally consider a sequence $(\chi_L)_{L\in\mathbb{N}^*}$ of cut-off functions such that

$$0 \le \chi_L \le 1 \text{ on } \mathbb{R}^d,\ \chi_L = 1 \text{ on } \Gamma_L,\ \mathrm{Supp}(\chi_L) \subset (L+\sqrt{L})\Gamma,\ \|\nabla\chi_L\|_{L^\infty} \le c,$$

for some constant $c \in \mathbb{R}_+$ independent of $L \in \mathbb{N}^*$.

Then, $\mathrm{Ran}(\mathcal{P}) \subset H^r(\mathbb{R}^d)$*, and there exists $C, \delta > 0$ such that for L large enough,*

$$\mathrm{Tr}(\mathcal{P}) = \mathrm{Tr}(\mathfrak{P}_L), \tag{4.25}$$

$$\sup_{\psi\in\mathrm{Ran}(\mathcal{P}),\ \|\psi\|_{L^2(\mathbb{R}^d)}=1} \inf_{u_L\in\mathrm{Ran}(\mathfrak{P}_L)} \|\psi - \chi_L u_L\|_{H^1(\mathbb{R}^d)} \le C\left(\epsilon_1(L) + \epsilon_2(L)\right), \tag{4.26}$$

$$\sup_{u_L\in\mathrm{Ran}(\mathfrak{P}_L),\ \|u_L\|_{L^2_{\rm per}(\Gamma_L)}=1} \inf_{\psi\in\mathrm{Ran}(\mathcal{P})} \|\psi - \chi_L u_L\|_{H^1(\mathbb{R}^d)} \le C\left(\epsilon_1(L) + \epsilon_2(L)\right), \tag{4.27}$$

$$\max_{\lambda_L\in\sigma(H_{L,N_L})\cap(\lambda-\varepsilon/2,\lambda+\varepsilon/2)} |\lambda_L - \lambda| \le C\left(\epsilon_1(L)^2 + \epsilon_2(L)\right), \tag{4.28}$$

where

$$\epsilon_1(L) := e^{-\delta L} + \left(\frac{L}{N_L}\right)^{r-1} \quad \text{and} \quad \epsilon_2(L) := \left(\frac{L}{M_L}\right)^{r-2} + \|W\|_{L^\infty(\mathbb{R}^d\setminus\Gamma_{L-1})}\left(e^{-\delta L} + \left(\frac{L}{N_L}\right)^{r}\right).$$

4.5.3 Formulation in terms of non-consistent approximations

The supercell method can be rewritten as a non-consistent approximation of the operator A (in the sense introduced in Section 4.2.2), based on the approximation spaces $(X_L)_{L\in\mathbb{N}^*}$ and the symmetric bilinear forms $(a_L)_{L\in\mathbb{N}^*}$, $(\widetilde{a}_L)_{L\in\mathbb{N}^*}$, and $(m_L)_{L\in\mathbb{N}^*}$ defined for all $L \in \mathbb{N}^*$ by

$$X_L := \{\chi_L u_L,\ u_L \in Y_{L,N_L}\} \subset H^1(\mathbb{R}^d),$$

and

$$\begin{aligned}
\forall \phi,\psi \in H^1(\mathbb{R}^d), \qquad a_L(\phi,\psi) &:= \int_{\Gamma_L} \nabla\phi\cdot\nabla\psi + \int_{\Gamma_L} \mathcal{I}_{L,M_L}(V_{\rm per} + \widetilde{W}_L)\phi\psi,\\
\widetilde{a}_L(\phi,\psi) &:= \int_{\Gamma_L} \nabla\phi\cdot\nabla\psi + \int_{\Gamma_L} (V_{\rm per} + W)\phi\psi,\\
m_L(\phi,\psi) &:= \int_{\Gamma_L} \phi\psi,
\end{aligned}$$

where we recall that $(\chi_L)_{L\in\mathbb{N}^*}$ is a sequence of cut-off functions satisfying (4.19). It is easily checked that for all $L \in \mathbb{N}^*$, $m_L(\cdot,\cdot)$ defines a scalar product on X_L.

Let us introduce, for each $L \in \mathbb{N}^*$, the unitary operator

$$\begin{array}{rccc}
j_L\ :\ \left(Y_{L,N_L}, \langle\cdot,\cdot\rangle_{L^2_{\rm per}(\Gamma_L)}\right) & \to & (X_L, m_L(\cdot,\cdot)), \\
u_L & \mapsto & \chi_L u_L.
\end{array}$$

Its adjoint (and inverse) j_L^* is given by: $\forall \phi_L \in X_L$, $j_L^*(\phi_L) = u_L$ where u_L is the $L\mathcal{R}$-periodic extension of $\phi_L|_{\Gamma_L}$. The supercell problems

$$\left\{ \begin{array}{l} \text{find } (\lambda_L, u_L) \in \mathbb{R} \times Y_{L,N_L} \text{ such that } \|u_L\|_{L^2_{\text{per}}(\Gamma_L)} = 1, \\ \forall v_L \in Y_{L,N_L}, \quad \widehat{a}_L(u_L, v_L) = \lambda_L \widehat{m}_L(u_L, v_L), \end{array} \right.$$

and

$$\left\{ \begin{array}{l} \text{find } (\lambda_L, u_L) \in \mathbb{R} \times Y_{L,N_L} \text{ such that } \|u_L\|_{L^2_{\text{per}}(\Gamma_L)} = 1, \\ \forall v_L \in Y_{L,N_L}, \quad \widehat{a}_{L,M_L}(u_L, v_L) = \lambda_L \widehat{m}_L(u_L, v_L), \end{array} \right.$$

are then respectively equivalent, through the change of variable $\psi_L = j_L u_L$, to the generalized eigenproblems

$$\left\{ \begin{array}{l} \text{find } (\lambda_L, \psi_L) \in \mathbb{R} \times X_L \text{ such that } m_L(\psi_L, \psi_L) = 1 \text{ and} \\ \forall \phi_L \in X_L, \quad \widetilde{a}_L(\psi_L, \phi_L) = \lambda_L m_L(\psi_L, \phi_L), \end{array} \right.$$

and

$$\left\{ \begin{array}{l} \text{find } (\lambda_L, \psi_L) \in \mathbb{R} \times X_L \text{ such that } m_L(\psi_L, \psi_L) = 1 \text{ and} \\ \forall \phi_L \in X_L, \quad a_L(\psi_L, \phi_L) = \lambda_L m_L(\psi_L, \phi_L). \end{array} \right.$$

Thus, considering the supercell method with exact and numerical integrations is equivalent to considering the non-consistent but conforming approximations $(\mathcal{T}_L)_{L\in\mathbb{N}^*}$ and $(\widetilde{\mathcal{T}}_L)_{L\in\mathbb{N}^*}$ respectively defined by

$$\widetilde{\mathcal{T}}_L = (X_L, \widetilde{a}_L, m_L) \quad \text{and} \quad \mathcal{T}_L = (X_L, a_L, m_L).$$

Taking the same notation as in Section 4.3, it holds that $\widetilde{A}_L = j_L H_{L,N_L} j_L^*$ and $A_L = j_L H_{L,N_L,M_L} j_L^*$ so that $\sigma(\widetilde{A}_L) = \sigma(H_{L,N_L})$, $\sigma(A_L) = \sigma(H_{L,N_L,M_L})$ and, in both cases, $\mathcal{P}_L = i_{X_L} j_L \mathfrak{P}_L j_L^* i_{X_L}^*$. The following section is devoted to the proof of Theorems 4.5.1 and 4.5.2, which are in fact corollaries of Theorem 4.3.1. We will first check that all the assumptions of Theorem 4.3.1 are satisfied for the approximations $(\widetilde{\mathcal{T}}_L)_{L\in\mathbb{N}^*}$ and $(\mathcal{T}_L)_{L\in\mathbb{N}^*}$, and then derive more explicit expressions of the right hand sides of (4.5), (4.6) and (4.7) in terms of L, N_L and M_L.

We prove in Section 4.6.1 that the supercell method with exact integration satisfies assumptions (A1)-(A4). In Section 4.6.2, we prove (4.18) and (4.20), which imply that this method also satisfies assumptions (B1) and (B2) for any discrete eigenvalue λ of the operator A. Estimating the terms involved in estimates (4.5), (4.6) and (4.7) will then lead to estimates (4.21), (4.22) and (4.23) and conclude the proof of Theorem 4.5.1. Section 4.6.4 is devoted to the proof of Theorem 4.5.2, in which numerical integration errors are taken into account.

4.6 Proof of Theorem 4.5.1 and Theorem 4.5.2

In the sequel, C will denote an arbitrary constant independent on $L \in \mathbb{N}^*$ which may vary along the calculations.

4.6.1 Proof of (A1)-(A4) for $\widetilde{\mathcal{T}}_L = (X_L, \widetilde{a}_L, m_L)$

Proof of (A1): Let us prove that

$$\forall \phi \in H^1(\mathbb{R}^d), \ \inf_{\phi_L \in X_L} \|\phi - \phi_L\|_{H^1(\mathbb{R}^d)} \underset{L\to\infty}{\longrightarrow} 0.$$

Let $\phi \in H^1(\mathbb{R}^d)$ and $\varepsilon > 0$. Since $C_c^\infty(\mathbb{R}^d)$ is dense in $H^1(\mathbb{R}^d)$, there exists $\eta \in C_c^\infty(\mathbb{R}^d)$ such that $\|\phi - \eta\|_{H^1(\mathbb{R}^d)} \leq \varepsilon$. Let $L_0 \in \mathbb{N}^*$ be such that $\mathrm{Supp}(\eta) \subset (L_0 - \sqrt{L_0})\Gamma$. For all $L \geq L_0$, if η_L denotes the $L\mathcal{R}$-periodic extension of $\eta|_{\Gamma_L}$, we infer from (4.17) that

$$\|\eta_L - \Pi_{L,N_L}\eta_L\|_{H^1_{\mathrm{per}}(\Gamma_L)} \leq C\frac{L}{N_L}\|\eta_L\|_{H^2_{\mathrm{per}}(\Gamma_L)} = C\frac{L}{N_L}\|\eta\|_{H^2(\mathbb{R}^d)} \underset{L\to\infty}{\longrightarrow} 0,$$

with $C \in \mathbb{R}_+$ independent of L. Let us then consider the sequence $(\phi_L)_{L\in\mathbb{N}^*}$ defined as $\phi_L := \chi_L\Pi_{L,N_L}\eta_L \in X_L$ for all $L \in \mathbb{N}^*$, for which

$$\begin{aligned}\|\phi - \phi_L\|_{H^1(\mathbb{R}^d)} &\leq \|\phi - \eta\|_{H^1(\mathbb{R}^d)} + \|\eta - \chi_L\Pi_{L,N_L}\eta_L\|_{H^1(\mathbb{R}^d)},\\ &\leq \varepsilon + \|\eta_L - \Pi_{L,N_L}\eta_L\|_{H^1_{\mathrm{per}}(\Gamma_L)} + \|\chi_L\Pi_{L,N_L}\eta_L\|_{H^1((L+\sqrt{L})\Gamma\setminus\Gamma_L)}.\end{aligned}$$

Furthermore, since $0 \leq \chi_L \leq 1$, and $\eta_L = 0$ on $(L+\sqrt{L})\Gamma \setminus \Gamma_L$, it holds

$$\begin{aligned}\|\chi_L\Pi_{L,N_L}\eta_L\|^2_{H^1((L+\sqrt{L})\Gamma\setminus\Gamma_L)} &\leq \|\chi_L\Pi_{L,N_L}\eta_L\|^2_{L^2((L+\sqrt{L})\Gamma\setminus\Gamma_L)} + 2\|\nabla\chi_L\Pi_{L,N_L}\eta_L\|^2_{L^2((L+\sqrt{L})\Gamma\setminus\Gamma_L)}\\ &\quad +2\|\chi_L\nabla(\Pi_{L,N_L}\eta_L)\|^2_{L^2((L+\sqrt{L})\Gamma\setminus\Gamma_L)}\\ &\leq \|\Pi_{L,N_L}\eta_L - \eta_L\|^2_{L^2((L+\sqrt{L})\Gamma\setminus\Gamma_L)} + \|\nabla(\Pi_{L,N_L}\eta_L) - \nabla\eta_L\|^2_{L^2((L+\sqrt{L})\Gamma\setminus\Gamma_L)}\\ &\quad +\|\nabla\chi_L\|_{L^\infty(\mathbb{R}^d)}\|\Pi_{L,N_L}\eta_L - \eta_L\|^2_{L^2((L+\sqrt{L})\Gamma\setminus\Gamma_L)}\\ &\leq 3^d(4 + \|\nabla\chi_L\|_{L^\infty(\mathbb{R}^d)})\|\Pi_{L,N_L}\eta_L - \eta_L\|^2_{H^1_{\mathrm{per}}(\Gamma_L)} \underset{L\to\infty}{\longrightarrow} 0.\end{aligned}$$

Hence the result.

Proof of (A2): Let $\phi_L, \psi_L \in X_L$, and $u_L, v_L \in Y_{L,N_L} \subset H^1_{\mathrm{per}}(\Gamma_L)$ such that $\phi_L = \chi_L u_L$ and $\psi_L = \chi_L v_L$. It holds

$$\int_{\mathbb{R}^d}|\phi_L|^2 = \int_{\Gamma_{3L}}|\phi_L|^2 = \int_{\Gamma_{3L}}\chi_L^2|u_L|^2 \leq 3^d\int_{\Gamma_L}|u_L|^2 = 3^d\int_{\Gamma_L}|\phi_L|^2 \leq 3^d\int_{\mathbb{R}^d}|\phi_L|^2.$$

Therefore,

$$\frac{1}{3^d}\|\phi_L\|^2_{L^2(\mathbb{R}^d)} \leq m_L(\phi_L,\phi_L) \leq \|\phi_L\|_{L^2(\mathbb{R}^d)}. \tag{4.29}$$

Besides,

$$\begin{aligned}|\widetilde{a}_L(\phi_L,\psi_L)| &= \left|\int_{\Gamma_L}\nabla\phi_L\cdot\nabla\psi_L + \int_{\Gamma_L}(V_{\mathrm{per}} + W)\phi_L\psi_L\right|\\ &\leq (1 + \|W\|_{L^\infty(\mathbb{R}^d)})\|\phi_L\|_{H^1(\mathbb{R}^d)}\|\psi_L\|_{H^1(\mathbb{R}^d)} + \|V_{\mathrm{per}}u_L\|_{L^2_{\mathrm{per}}(\Gamma_L)}\|v_L\|_{L^2_{\mathrm{per}}(\Gamma_L)}\\ &\leq (1 + \|W\|_{L^\infty(\mathbb{R}^d)} + \|V_{\mathrm{per}}\|_{\mathcal{Z}_{\mathrm{per}}(\Gamma)})\|\phi_L\|_{H^1(\mathbb{R}^d)}\|\psi_L\|_{H^1(\mathbb{R}^d)}.\end{aligned}$$

Thus, assumption (A2) is satisfied.

Proof of (A3): For all $\alpha > 0$ arbitrarily small, there exists a constant C_α such that for all $\phi \in H^1(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d}|V_{\mathrm{per}}||\phi|^2 \leq \alpha\int_{\mathbb{R}^d}|\nabla\phi|^2 + C_\alpha\int_{\mathbb{R}^d}|\phi|^2. \tag{4.30}$$

Besides, for all $\phi_L \in X_L$, if $\phi_L = \chi_L u_L$ with $u_L \in Y_{L,N_L}$, it holds that

$$\begin{aligned}\int_{\mathbb{R}^d}|\nabla\phi_L|^2 &\leq 2\int_{(L+\sqrt{L})\Gamma}|\nabla\chi_L u_L|^2 + |\chi_L\nabla u_L|^2\\ &\leq 2\times 3^d\left(\|\nabla\chi_L\|_{L^\infty(\mathbb{R}^d)}\int_{\Gamma_L}|u_L|^2 + \int_{\Gamma_L}|\nabla u_L|^2\right),\end{aligned}$$

which, together with (4.29), yields that, for L large enough

$$\int_{\mathbb{R}^d} |\phi_L|^2 + \int_{\mathbb{R}^d} |\nabla \phi_L|^2 \leq 3^{d+1} \left(\int_{\Gamma_L} |\phi_L|^2 + \int_{\Gamma_L} |\nabla \phi_L|^2 \right). \tag{4.31}$$

Using (4.30) and (4.31), we obtain that for all $\alpha > 0$ arbitrarily small, there exists $D_\alpha \in \mathbb{R}_+$ such that for all $L \in \mathbb{N}^*$ and all $\phi_L \in X_L$,

$$\begin{aligned} \int_{\Gamma_L} (V_{\rm per} + W)|\phi_L|^2 &\leq \int_{\mathbb{R}^d} (|V_{\rm per}| + |W|)\,|\phi_L|^2 \\ &\leq \alpha \int_{\Gamma_L} |\nabla \phi_L|^2 + D_\alpha \int_{\Gamma_L} |\phi_L|^2. \end{aligned}$$

This last inequality implies that there exists $\beta > 0$ independent on $L \in \mathbb{N}^*$ such that for all $\phi_L \in X_L$,

$$\|\phi_L\|^2_{H^1(\mathbb{R}^d)} \leq 3^{d+1} \|\phi_L\|^2_{H^1(\Gamma_L)} \leq \beta \left(|\widetilde{a}_L(\phi_L, \phi_L)| + m_L(\phi_L, \phi_L) \right).$$

Thus, for all $\mu \in \mathbb{C}$, it holds that

$$\begin{aligned} &\inf_{\psi_L \in X_L} \sup_{\phi_L \in X_L} \frac{|(\widetilde{a}_L - \mu m_L)(\phi_L, \psi_L)|}{\|\phi_L\|_{H^1(\mathbb{R}^d)} \|\psi_L\|_{H^1(\mathbb{R}^d)}} \\ &\geq \frac{1}{\beta} \inf_{\psi_L \in X_L} \sup_{\phi_L \in X_L} \frac{|(\widetilde{a}_L - \mu m_L)(\phi_L, \psi_L)|}{\left(|\widetilde{a}_L(\phi_L, \phi_L)| + m_L(\phi_L, \phi_L) \right)^{1/2} \left(|\widetilde{a}_L(\psi_L, \psi_L)| + m_L(\psi_L, \psi_L) \right)^{1/2}}. \end{aligned}$$

Let $(\zeta_L^i)_{1 \leq i \leq \dim(X_L)}$ be an m_L-orthonormal basis of X_L, such that for all $1 \leq i \leq \dim(X_L)$,

$$H_{L,N_L} j_L^* \zeta_L^{(i)} = \nu_L^i j_L^* \zeta_L^i,\ 1 \leq i \leq \dim(X_L),$$

where $\{\nu_L^i,\ 1 \leq i \leq \dim(X_L)\} = \sigma(H_{L,N_L})$. Then, any $\phi_L \in X_L$ can be expanded in the basis $(\zeta_L^i)_{1 \leq i \leq \dim(X_L)}$:

$$\phi_L = \sum_{i=1}^{\dim(X_L)} c_i \zeta_L^i,\ c_i \in \mathbb{R},\ 1 \leq i \leq \dim(X_L),$$

and it holds that $|\widetilde{a}_L(\phi_L, \phi_L)| + m_L(\phi_L, \phi_L) \leq \sum_{i=1}^{\dim(X_L)} |c_i|^2 (1 + |\nu_L^i|)$. Considering

$$\psi_L := \sum_{i=1}^{\dim(X_L)} \mathrm{sgn}(\nu_L^i - \mu) c_i \zeta_L^i,$$

we obtain that

$$\inf_{\psi_L \in X_L} \sup_{\phi_L \in X_L} \frac{|(\widetilde{a}_L - \mu m_L)(\phi_L, \psi_L)|}{\|\phi_L\|_{H^1(\mathbb{R}^d)} \|\psi_L\|_{H^1(\mathbb{R}^d)}} \geq \frac{1}{\beta} \inf_{\nu_L \in \sigma(H_{L,N_L})} \frac{|\nu_L - \mu|}{1 + |\nu_L|}. \tag{4.32}$$

Since (4.32) holds for any $\mu \in \mathbb{C}$, this implies that for any compact subset $K \subset \mathbb{C}$, there exists a constant $c_K > 0$ such that for all $L \in \mathbb{N}^*$ and all $\mu \in K$,

$$\inf_{\psi_L \in X_L} \sup_{\phi_L \in X_L} \frac{|(\widetilde{a}_L - \mu m_L)(\phi_L, \psi_L)|}{\|\phi_L\|_{H^1(\mathbb{R}^d)} \|\psi_L\|_{H^1(\mathbb{R}^d)}} \geq c_K \min\left(1, \mathrm{dist}(\mu, \sigma(H_{L,N_L}))\right).$$

Thus, condition (A3'), and condition (A3), hold for the approximation $(\widetilde{\mathcal{T}}_L)_{L\in\mathbb{N}^*}$.

Proof of (A4): For all $\phi \in H^1(\mathbb{R}^d)$, we denote by

$$r_L^m(\phi) := \left(\int_{\mathbb{R}^d\setminus\Gamma_{L-1}} |\phi|^2\right)^{1/2} \leq \|\phi\|_{L^2(\mathbb{R}^d)},$$

and

$$r_L^a(\phi) := \left(\int_{\mathbb{R}^d\setminus\Gamma_{L-1}} |\phi|^2 + |\nabla\phi|^2\right)^{1/2} \leq \|\phi\|_{H^1(\mathbb{R}^d)}.$$

Then, r_L^m and r_L^a are seminorms on $H^1(\mathbb{R}^d)$ such that for all $\phi \in H^1(\mathbb{R}^d)$, $r_L^m(\phi) \underset{L\to\infty}{\longrightarrow} 0$ and $r_L^a(\phi) \underset{L\to\infty}{\longrightarrow} 0$. For all $\phi, \psi \in H^1(\mathbb{R}^d)$, it holds

$$|(m - m_L)(\phi,\psi)| = \left|\int_{\mathbb{R}^d\setminus\Gamma_L} \phi\psi\right| \leq r_L^m(\phi) r_L^m(\psi).$$

Let $(\omega_L)_{L\in\mathbb{N}^*}$ be a sequence of C^∞ cut-off functions such that for all $L \in \mathbb{N}^*$, $0 \leq \omega_L \leq 1$, $\omega_L = 1$ on $\mathbb{R}^d \setminus \Gamma_L$, $\omega_L = 0$ on Γ_{L-1} and the sequence $(\|\nabla\omega_L\|_{L^\infty(\mathbb{R}^d)})_{L\in\mathbb{N}^*}$ is uniformly bounded in $L \in \mathbb{N}^*$. Then, for all $\phi, \psi \in H^1(\mathbb{R}^d)$,

$$\begin{aligned}
|(a - \widetilde{a}_L)(\phi,\psi)| &= \left|\int_{\mathbb{R}^d\setminus\Gamma_L} \nabla\phi\cdot\nabla\psi + \int_{\mathbb{R}^d\setminus\Gamma_L} (V_{\rm per} + W)\phi\psi\right| \\
&\leq (1 + \|W\|_{L^\infty}) r_L^a(\phi) r_L^a(\psi) + \int_{\mathbb{R}^d} |V_{\rm per}\omega_L\phi\omega_L\psi| \\
&\leq (1 + \|W\|_{L^\infty}) r_L^a(\phi) r_L^a(\psi) + \left(\int_{\mathbb{R}^d} |V_{\rm per}||\omega_L\phi|^2\right)^{1/2} \left(\int_{\mathbb{R}^d} |V_{\rm per}||\omega_L\psi|^2\right)^{1/2}.
\end{aligned}$$

Using (4.30),

$$\int_{\mathbb{R}^d} |V_{\rm per}||\omega_L\phi|^2 \leq \frac{1}{2}\int_{\mathbb{R}^d} |\nabla(\omega_L\phi)|^2 + C\int_{\mathbb{R}^d} |\omega_L\phi|^2 \leq C r_L^a(\phi)^2.$$

Thus, there exists $\kappa \in \mathbb{R}_+$ independent on $L \in \mathbb{N}^*$ such that

$$|(a - \widetilde{a}_L)(\phi,\psi)| \leq \kappa r_L^a(\phi) r_L^a(\psi).$$

4.6.2 Absence of pollution

Proposition 4.6.1. *It holds*

$$\sigma(A) = \lim_{L\to\infty} \sigma(H_{L,N_L}). \tag{4.33}$$

Besides, for any discrete eigenvalue λ of the operator A and for all $\varepsilon > 0$ such that $(\lambda - \varepsilon, \lambda + \varepsilon) \cap \sigma(A) = \{\lambda\}$, we have, for L large enough,

$$\mathrm{Tr}(\mathfrak{P}_L) = \mathrm{Tr}(\mathcal{P}), \tag{4.34}$$

where $\mathcal{P} := \mathbb{1}_{\{\lambda\}}(A)$ and $\mathfrak{P}_L := \mathbb{1}_{(\lambda-\varepsilon/2,\lambda+\varepsilon/2)}(H_{L,N_L})$.

Let us notice that (4.33) implies that (B1) is satisfied for any discrete eigenvalue of A, and that (4.34) is nothing but a reformulation of (B2). We refer to [40, Theorem 3.1] for a proof of (4.33).

Proof of (4.34). If follows from (4.33) that (B1) is satisfied and therefore that for n large enough, $\mathrm{Tr}(\mathfrak{P}_{L_n}) \geq \mathrm{Tr}(\mathcal{P})$. Let us assume that there exists an increasing sequence $(L_k)_{k\in\mathbb{N}^*}$ of integers such that

$$\mathrm{Tr}(\mathfrak{P}_{L_k}) > q := \mathrm{Tr}(\mathcal{P}).$$

For all $k \in \mathbb{N}$, let $(\zeta_{L_k}^{(i)})_{1\leq i\leq q+1}$ be an $L^2_{\mathrm{per}}(\Gamma_{L_k})$-orthonormal family of vectors of $Y_{L_k,N_{L_k}}$ such that for all $1 \leq i \leq q+1$,

$$H_{L_k,N_{L_k}}\zeta_{L_k}^{(i)} = \lambda_{L_k}^{(i)}\zeta_{L_k}^{(i)} \quad \text{with } \lambda_{L_k}^{(i)} \in (\lambda - \varepsilon/2, \lambda + \varepsilon/2).$$

Then, for all $k \in \mathbb{N}$, $(\chi_{L_k}\zeta_{L_k}^{(i)})_{1\leq i\leq q+1}$ forms a free family of X_{L_k} and there exists $g_k \in \mathrm{Span}(\zeta_{L_k}^{(i)})_{1\leq i\leq q+1}$ such that $\|g_k\|_{L^2_{\mathrm{per}}(\Gamma_{L_k})} = 1$ and

$$\widetilde{g}_k := \chi_{L_k} g_k \in \mathrm{Ker}(\mathcal{P}).$$

Reasoning as above, it can be easily checked that $\left(\|g_k\|_{H^1_{\mathrm{per}}(\Gamma_{L_k})}\right)_{k\in\mathbb{N}^*}$ is bounded, which implies that $\left(\|\widetilde{g}_k\|_{H^1(\mathbb{R}^d)}\right)_{k\in\mathbb{N}^*}$ is bounded as well. Thus, up to the extraction of a subsequence, there exists $g \in H^1(\mathbb{R}^d) \cap \mathrm{Ker}(\mathcal{P})$ such that $\widetilde{g}_k \underset{k\to\infty}{\rightharpoonup} g$ in $H^1(\mathbb{R}^d)$ and $\widetilde{g}_k \underset{k\to\infty}{\longrightarrow} g$ in $L^2_{\mathrm{loc}}(\mathbb{R}^d)$. Since $\widetilde{g}_k = \chi_{L_k} g_k$, this also implies that

$$g_k \underset{k\to\infty}{\longrightarrow} g \quad \text{strongly in } L^2_{\mathrm{loc}}(\mathbb{R}^d),$$

which readily leads to

$$\left(H_{L_k,N_{L_k}} - \lambda\right) g_k \underset{k\to\infty}{\longrightarrow} -\Delta g + (V_{\mathrm{per}} + W - \lambda)g \quad \text{in } \mathcal{D}'(\mathbb{R}^d).$$

Besides, since $g_k \in \mathrm{Ran}(\mathfrak{P}_{L_k})$ and $\lim_{k\to\infty} \sigma(H_{L_k,N_{L_k}}) = \sigma(H)$, we have,

$$\left\|\left(H_{L_k,N_{L_k}} - \lambda\right) g_k\right\|_{L^2_{\mathrm{per}}(\Gamma_{L_k})} \underset{k\to\infty}{\longrightarrow} 0,$$

which, in turn, implies that

$$\left(H_{L_k,N_{L_k}} - \lambda\right) g_k \underset{k\to\infty}{\longrightarrow} 0 \quad \text{in } \mathcal{D}'(\mathbb{R}^d).$$

Therefore,

$$-\Delta g + (V_{\mathrm{per}} + W - \lambda)g = 0.$$

Consequently, $g \in \mathrm{Ker}(\mathcal{P}) \cap \mathrm{Ran}(\mathcal{P}) = \{0\}$. Using similar arguments as those used in the proof of [40, Theorem 3.1], we infer from the fact that $(g_k)_{k\in\mathbb{N}}$ strongly converges to 0 in $L^2_{\mathrm{loc}}(\mathbb{R}^d)$ that $\left(\frac{\widetilde{g}_k}{\|\widetilde{g}_k\|_{L^2(\mathbb{R}^d)}}\right)_{k\in\mathbb{N}}$ is a Weyl sequence for $A^0 = -\Delta + V_{\mathrm{per}}$ associated with λ, which contradicts the fact that $\lambda \notin \sigma(A^0)$. □

4.6.3 Proof of Theorem 4.5.1

We have proved that the supercell method with planewave discretization and exact integration satisfies assumptions (A1)-(A4), and that for each discrete eigenvalue located in a spectral gap of A, assumptions (B1) and (B2) are satisfied. Thus, Theorem 4.3.1 can be applied and there exists $C \in \mathbb{R}_+$ such that for L large enough,

$$\begin{aligned}
\mathrm{Tr}(\mathcal{P}_L) &= \mathrm{Tr}(\mathcal{P}) = \mathrm{Tr}(\mathfrak{P}_L), \\
\|(\mathcal{P}-\mathcal{P}_L)\mathcal{P}\|_{\mathcal{L}(L^2(\mathbb{R}^d),H^1(\mathbb{R}^d))} &\leq C\left(\left\|\left(1-\Pi_{X_L}^{H^1(\mathbb{R}^d)}\right)\mathcal{P}\right\|_{\mathcal{L}(L^2(\mathbb{R}^d),H^1(\mathbb{R}^d))} + \mathcal{R}_L^a + \mathcal{R}_L^m\right), \\
\|(\mathcal{P}-\mathcal{P}_L)\mathcal{P}_L\|_{\mathcal{L}(L^2(\mathbb{R}^d),H^1(\mathbb{R}^d))} &\leq C\left(\left\|\left(1-\Pi_{X_L}^{H^1(\mathbb{R}^d)}\right)\mathcal{P}\right\|_{\mathcal{L}(L^2(\mathbb{R}^d),H^1(\mathbb{R}^d))} + \mathcal{R}_L^a + \mathcal{R}_L^m\right), \\
\max_{\lambda_L\in\sigma(H_{L,N_L})\cap(\lambda-\varepsilon/2,\lambda+\varepsilon/2)} |\lambda_L-\lambda| &\leq C\left(\left\|\left(1-\Pi_{X_L}^{H^1(\mathbb{R}^d)}\right)\mathcal{P}\right\|_{\mathcal{L}(L^2(\mathbb{R}^d),H^1(\mathbb{R}^d))} + \mathcal{R}_L^a + \mathcal{R}_L^m\right)^2,
\end{aligned}$$

where $\mathcal{P}_L := i_{X_L} j_L \mathfrak{P}_L j_L^* i_{X_L}^*$ and

$$\begin{aligned}
\mathcal{R}_L^m &:= \sup_{\psi\in\mathrm{Ran}(\mathcal{P}),\ \|\psi\|_{L^2(\mathbb{R}^d)}=1} r_L^m(\psi), \\
\mathcal{R}_L^a &:= \sup_{\psi\in\mathrm{Ran}(\mathcal{P}),\ \|\psi\|_{L^2(\mathbb{R}^d)}=1} r_L^a(\psi).
\end{aligned}$$

Since we have

$$\sup_{\psi\in\mathrm{Ran}(\mathcal{P}),\ \|\psi\|_{L^2(\mathbb{R}^d)}=1} \inf_{u_L\in\mathrm{Ran}(\mathfrak{P}_L)} \|\psi-\chi_L u_L\|_{H^1(\mathbb{R}^d)} \leq \|(\mathcal{P}-\mathcal{P}_L)\mathcal{P}\|_{\mathcal{L}(L^2(\mathbb{R}^d),H^1(\mathbb{R}^d))},$$

and

$$\sup_{u_L\in\mathrm{Ran}(\mathfrak{P}_L),\ \|u_L\|_{L^2_{\mathrm{per}}(\Gamma_L)}=1} \inf_{\psi\in\mathrm{Ran}(\mathcal{P})} \|\psi-\chi_L u_L\|_{H^1(\mathbb{R}^d)} \leq \|(\mathcal{P}-\mathcal{P}_L)\mathcal{P}_L\|_{\mathcal{L}(L^2(\mathbb{R}^d),H^1(\mathbb{R}^d))},$$

it just remains to prove that there exists $\delta > 0$ independent on L such that

$$\left\|\left(1-\Pi_{X_L}^{H^1(\mathbb{R}^d)}\right)\mathcal{P}\right\|_{\mathcal{L}(L^2(\mathbb{R}^d),H^1(\mathbb{R}^d))} + \mathcal{R}_L^a + \mathcal{R}_L^m \leq C\left(e^{-\delta L} + \left(\frac{N_L}{L}\right)^{r-1}\right).$$

This estimate is based on exponential decay results for the bound states of Schrödinger operators [168]. A real-valued function V on $\mathbb{R}^d$ is said to lie in the class K_d if and only if

$$\begin{aligned}
&\text{if } d \geq 3, \qquad \lim_{\alpha\downarrow 0} \sup_{x\in\mathbb{R}^d} \int_{|x-y|\leq\alpha} \frac{|V(y)|}{|x-y|^{d-2}}\, dy = 0; \\
&\text{if } d = 2, \qquad \lim_{\alpha\downarrow 0} \sup_{x\in\mathbb{R}^d} \int_{|x-y|\leq\alpha} |V(y)| \ln\left(|x-y|^{-1}\right)\, dy = 0; \\
&\text{if } d = 1, \qquad \sup_{x\in\mathbb{R}^d} \int_{|x-y|\leq 1} |V(y)|\, dy < \infty.
\end{aligned}$$

Under our assumptions on V_{per} and W, $V = V_{\mathrm{per}} + W \in K_d$. It then follows from Theorem C.3.4 and Corollary C.2.3 in [168] that there exists $C, \delta > 0$ such that for all $L^2(\mathbb{R}^d)$-normalized $\psi \in \mathrm{Ran}(\mathcal{P})$,

$$\forall x \in \mathbb{R}^d, \quad |\psi(x)| \leq Ce^{-3\delta|x|} \quad \text{and} \quad e^{3\delta|\cdot|}\nabla\psi \in \left(L^2(\mathbb{R}^d)\right)^d. \tag{4.35}$$

For all $L \geq 6$, let $\eta_L \in C_c^\infty(\mathbb{R}^d)$ such that $0 \leq \eta_L \leq 1$, $\eta_L = 1$ on $\Gamma_{L/2-2}$, $\mathrm{Supp}(\eta_L) \subset \Gamma_{L/2-1}$ and all its derivative up to the $[r+1]^{\mathrm{st}}$ order are bounded in $L^\infty(\mathbb{R}^d)$, uniformly in $L \in \mathbb{N}^*$. Let $\psi \in \mathrm{Ran}(\mathcal{P})$ such that $\|\psi\|_{L^2(\mathbb{R}^d)} = 1$, $\zeta_L = \eta_L \psi$, and $\widetilde{\zeta}_L$ the $L\mathcal{R}$-periodic extension of ζ_L. Then, $\chi_L \Pi_{L,N_L} \widetilde{\zeta}_L \in X_L$, and it holds

$$\begin{aligned}
\|\psi - \chi_L \Pi_{L,N_L} \widetilde{\zeta}_L\|_{H^1(\mathbb{R}^d)} &\leq \|\psi - \eta_L\psi\|_{H^1(\mathbb{R}^d)} + \|\zeta_L - \chi_L \Pi_{L,N_L} \widetilde{\zeta}_L\|_{H^1(\mathbb{R}^d)} \\
&= \|\psi - \eta_L\psi\|_{H^1(\mathbb{R}^d)} + \|\chi_L \left(\zeta_L - \Pi_{L,N_L} \widetilde{\zeta}_L\right)\|_{H^1(\mathbb{R}^d)} \\
&\leq Ce^{-\delta L} + C\|\widetilde{\zeta}_L - \Pi_{L,N_L} \widetilde{\zeta}_L\|_{H^1_{\mathrm{per}}(\Gamma_L)} \\
&\leq Ce^{-\delta L} + C\left(\frac{N_L}{L}\right)^{r-1} \|\widetilde{\zeta}_L\|_{H^r_{\mathrm{per}}(\Gamma_L)} \\
&\leq C\left(e^{-\delta L} + \left(\frac{N_L}{L}\right)^{r-1} \|\zeta_L\|_{H^r(\mathbb{R}^d)}\right) \\
&\leq C\left(e^{-\delta L} + \left(\frac{N_L}{L}\right)^{r-1} \|\psi\|_{H^r(\mathbb{R}^d)}\right) \\
&\leq C\left(e^{-\delta L} + \left(\frac{N_L}{L}\right)^{r-1}\right).
\end{aligned}$$

This yields the estimate

$$\left\| \left(1 - \Pi_{X_L}^{H^1(\mathbb{R}^d)}\right) \mathcal{P} \right\|_{\mathcal{L}(L^2(\mathbb{R}^d), H^1(\mathbb{R}^d))} \leq C\left(e^{-\delta L} + \left(\frac{N_L}{L}\right)^{r-1}\right).$$

The remaining estimate

$$\mathcal{R}_L^a + \mathcal{R}_L^m \leq Ce^{-\delta L},$$

is a straightforward consequence of (4.35).

4.6.4 Proof of Theorem 4.5.2

Let us first remark that since $\dfrac{M_L}{L} = G_L \in \mathbb{N}^*$, $\mathcal{I}_{L,M_L}(V_{\mathrm{per}}) = \mathcal{I}_{1,G_L}(V_{\mathrm{per}})$ is a $\mathcal{R}$-periodic function. Let $\phi_L, \psi_L \in X_L$ be such that $\phi_L = \chi_L u_L$ and $\psi_L = \chi_L v_L$ with $u_L, v_L \in Y_{L,N_L}$. Then, we have

$$\left| \int_{\Gamma_L} \left(V_{\mathrm{per}} - \mathcal{I}_{L,M_L}(V_{\mathrm{per}})\right) \phi_L \psi_L \right| \leq \left| \int_{\Gamma_L} |V_{\mathrm{per}} - \mathcal{I}_{L,M_L}(V_{\mathrm{per}})|\, u_L^2 \right|^{1/2} \left| \int_{\Gamma_L} |V_{\mathrm{per}} - \mathcal{I}_{L,M_L}(V_{\mathrm{per}})|\, v_L^2 \right|^{1/2}.$$

As

$$\begin{aligned}
\int_{\Gamma_L} |V_{\mathrm{per}} - \mathcal{I}_{L,M_L}(V_{\mathrm{per}})|\, u_L^2 &= \sum_{R \in \mathcal{R} \cap \Gamma_L} \int_\Gamma |V_{\mathrm{per}} - \mathcal{I}_{1,G_L}(V_{\mathrm{per}})|\, u_L(\cdot + R)^2 \\
&\leq \|V_{\mathrm{per}} - \mathcal{I}_{1,G_L}(V_{\mathrm{per}})\|_{L^2_{\mathrm{per}}(\Gamma)} \sum_{R \in \mathcal{R} \cap \Gamma_L} \|u_L(\cdot + R)\|^2_{L^4(\Gamma)} \\
&\leq C\, \|V_{\mathrm{per}} - \mathcal{I}_{1,G_L}(V_{\mathrm{per}})\|_{L^2_{\mathrm{per}}(\Gamma)} \sum_{R \in \mathcal{R} \cap \Gamma_L} \|u_L(\cdot + R)\|^2_{H^1(\Gamma)} \\
&= C\, \|V_{\mathrm{per}} - \mathcal{I}_{1,G_L}(V_{\mathrm{per}})\|_{L^2_{\mathrm{per}}(\Gamma)} \|u_L\|^2_{H^1_{\mathrm{per}}(\Gamma_L)} \\
&\leq C\, \|V_{\mathrm{per}} - \mathcal{I}_{1,G_L}(V_{\mathrm{per}})\|_{L^2_{\mathrm{per}}(\Gamma)} \|\phi_L\|^2_{H^1(\mathbb{R}^d)},
\end{aligned}$$

we obtain

$$\left|\int_{\Gamma_L}(V_{\rm per}-\mathcal{I}_{L,M_L}(V_{\rm per}))\phi_L\psi_L\right|\leq C\|V_{\rm per}-\mathcal{I}_{1,G_L}(V_{\rm per})\|_{L^2_{\rm per}(\Gamma)}\|\phi_L\|_{H^1(\mathbb{R}^d)}\|\psi_L\|_{H^1(\mathbb{R}^d)}, \tag{4.36}$$

for a constant C independent of L, with

$$\|V_{\rm per}-\mathcal{I}_{1,G_L}(V_{\rm per})\|_{L^2_{\rm per}(\Gamma)}\leq CG_L^{-(r-2)}\|V_{\rm per}\|_{H^{r-2}_{\rm per}(\Gamma)}=C\left(\frac{L}{M_L}\right)^{r-2}\|V_{\rm per}\|_{H^{r-2}_{\rm per}(\Gamma)}\underset{L\to\infty}{\longrightarrow}0. \tag{4.37}$$

Besides, since $W\in C^0(\mathbb{R}^d)\cap H^{r-2}(\mathbb{R}^d)$,

$$\begin{aligned}\|\widetilde{W}_L-\mathcal{I}_{L,M_L}(\widetilde{W}_L)\|_{L^2(\Gamma_L)} &\leq C\left(\frac{L}{M_L}\right)^{r-2}\|\widetilde{W}_L\|_{H^{r-2}_{\rm per}(\Gamma_L)}\\ &\leq C\left(\frac{L}{M_L}\right)^{r-2}\|W\|_{H^{r-2}(\mathbb{R}^d)},\end{aligned} \tag{4.38}$$

and $\|W-\widetilde{W}_L\|_{L^\infty(\mathbb{R}^d)}\leq\|W\|_{L^\infty(\mathbb{R}^d\setminus\Gamma_{L-1})}\underset{L\to\infty}{\longrightarrow}0$. Thus,

$$\sup_{\phi_L\in X_L}\sup_{\psi_L\in X_L}\frac{|(\widetilde{a}_L-a_L)(\phi_L,\psi_L)|}{\|\phi_L\|_{H^1(\mathbb{R}^d)}\|\psi_L\|_{H^1(\mathbb{R}^d)}}\underset{L\to\infty}{\longrightarrow}0.$$

Together with the results proved in Section 4.6.1, this implies (A2), (A3), (B1) and (B2) are satisfied for $\mathcal{T}_L=(X_L,a_L,m_L)$. Assumption (A4) is also satisfied for $\mathcal{T}_L=(X_L,a_L,m_L)$, with $\widetilde{a}_L(\cdot,\cdot)$ playing the role of $\widetilde{a}_n(\cdot,\cdot)$ and $\widetilde{m}_n(\cdot,\cdot)=m_n(\cdot,\cdot)=m_L(\cdot,\cdot)$. To obtain the estimates (4.26), (4.27) and (4.28), it remains to prove that

$$\mathcal{S}_L^a\leq C\left[\left(\frac{L}{M_L}\right)^{r-2}+\|W\|_{L^\infty(\mathbb{R}^d\setminus\Gamma_{L-1})}\left(e^{-\delta L}+\left(\frac{L}{N_L}\right)^r\right)\right],$$

where

$$\mathcal{S}_L^a:=\sup_{\psi\in{\rm Ran}(\mathcal{P}),\,\|\psi\|_{L^2(\mathbb{R}^d)}=1}\sup_{\phi_L\in X_L}\frac{\left|(a_L-\widetilde{a}_L)\left(\Pi_{X_L}^{H^1(\mathbb{R}^d)}\psi,\phi_L\right)\right|}{\|\phi_L\|_{H^1(\mathbb{R}^d)}}.$$

Using (4.36) and (4.37), we already have for all $\psi\in{\rm Ran}(\mathcal{P})$ such that $\|\psi\|_{L^2(\mathbb{R}^d)}=1$,

$$\left|\int_{\Gamma_L}(V_{\rm per}-\mathcal{I}_{L,M_L}(V_{\rm per}))\left(\Pi_{X_L}^{H^1(\mathbb{R}^d)}\psi\right)\phi_L\right|\leq C\left(\frac{L}{M_L}\right)^{r-2}\|\phi_L\|_{H^1(\mathbb{R}^d)}. \tag{4.39}$$

Besides, using (4.38), it holds that

$$\left|\int_{\Gamma_L}(\widetilde{W}_L-\mathcal{I}_{L,M_L}(\widetilde{W}_L))\left(\Pi_{X_L}^{H^1(\mathbb{R}^d)}\psi\right)\phi_L\right|\leq C\left(\frac{L}{M_L}\right)^{r-2}\|\phi_L\|_{H^1(\mathbb{R}^d)}. \tag{4.40}$$

It also follows from (4.35) that

$$\begin{aligned}\left|\int_{\Gamma_L}(\widetilde{W}_L-W)\left(\Pi_{X_L}^{H^1(\mathbb{R}^d)}\psi\right)\phi_L\right| &\leq \left|\int_{\Gamma_L\setminus\Gamma_{L-1}}(\widetilde{W}_L-W)\psi\phi_L\right|+\left|\int_{\Gamma_L\setminus\Gamma_{L-1}}(\widetilde{W}_L-W)\left(\psi-\Pi_{X_L}^{H^1(\mathbb{R}^d)}\psi\right)\phi_L\right|\\ &\leq C\|W\|_{L^\infty(\mathbb{R}^d\setminus\Gamma_{L-1})}\left(e^{-\delta L}+\left\|\psi-\Pi_{X_L}^{H^1(\mathbb{R}^d)}\psi\right\|_{L^2(\mathbb{R}^d)}\right)\|\phi_L\|_{H^1(\mathbb{R}^d)}.\end{aligned}$$

Reasoning as in the proof of (A1) in Section 4.6.1, and using (4.35), we can prove that

$$\left\|\psi - \Pi_{X_L}^{H^1(\mathbb{R}^d)}\psi\right\|_{L^2(\mathbb{R}^d)} \leq C\left(e^{-\delta L} + \left(\frac{L}{N_L}\right)^r\right).$$

Thus,

$$\left|\int_{\Gamma_L}(\widetilde{W}_L - W)\left(\Pi_{X_L}^{H^1(\mathbb{R}^d)}\psi\right)\phi_L\right| \leq C\left[\left(\frac{L}{M_L}\right)^{r-2} + \|W\|_{L^\infty(\mathbb{R}^d\setminus\Gamma_{L-1})}\left(e^{-\delta L} + \left(\frac{L}{N_L}\right)^r\right)\right]\|\phi_L\|_{H^1(\mathbb{R}^d)}. \tag{4.41}$$

Finally, using (4.39), (4.40) and (4.41), we obtain

$$\mathcal{S}_L^a \leq C\left[\left(\frac{L}{M_L}\right)^{r-2} + \|W\|_{L^\infty(\mathbb{R}^d\setminus\Gamma_{L-1})}\left(e^{-\delta L} + \left(\frac{L}{N_L}\right)^r\right)\right],$$

which ends the proof of Theorem 4.5.2.

4.7 Numerical results

In this section, we present some numerical results obtained with the software Scilab, illustrating the *a priori* estimates given in Theorem 4.5.1 and 4.5.2. These results have been obtained with $d = 1$, $V_{\rm per}(x) = |\sin x|$, $W(x) = -2\exp(-|x|)$ and $\Gamma = (-\pi, \pi]$. The particular form of these potentials enables us to compute the mass and stiffness matrices analytically (and therefore with no numerical integration error). The operator $A = -\Delta + V_{\rm per} + W$ then possesses a discrete simple eigenvalue $\lambda \approx 1.69$ located in the spectral gap $[\alpha, \beta]$ of the operator $A^0 = -\Delta + V_{\rm per}$ where $\alpha \approx 1.43$ and $\beta \approx 1.84$. The reference values for λ and the associated eigenvector (considered in our numerical study as the limits $L, N_L \to \infty$) are obtained with $L_{\rm ref} = 40$ and $N_{\rm ref} = 1400$.

Fig. 4.1 shows $\sigma(H_{L,N_{\rm ref}}) \cap [1, 2]$ for $L = 6, 8, 10, 12, 14, 16, 18$ and $N_{\rm ref} = 1400$. We can see that there is no spectral pollution, as predicted by [40] and Proposition 4.6.1.

The next series of numerical tests confirms the exponential convergence of the supercell method with respect to the size of the supercell. We have compared the eigenvalue closest to λ and the associated eigenvector obtained for different values of L ($L = 6, 8, 10, 12, 14, 16, 18$) to the reference eigenvalue and eigenvector obtained with $L = 40$, all these calculations being done with $N_{\rm ref} = 1400$. Fig. 4.2 shows the relative errors on the eigenvalue, and the square of the L^2 and H^1 norms of the error on the eigenvector. More precisely, for all $L \in \mathbb{N}^*$, we consider the eigenvector u_L of $H_{L,N_{\rm ref}}$ associated with the eigenvalue λ_L of $H_{L,N_{\rm ref}}$ closest to 1.69, and set $\phi_L = \chi_L u_L$, where χ_L is the unique C^2 function defined by $\chi_L = 1$ on $[-\pi L, \pi L]$, $\chi_L = 0$ on $\mathbb{R}\setminus[-\pi(L+\sqrt{L}), \pi(L+\sqrt{L})]$, and χ_L is a sixth degree polynomial on $[-\pi(L+\sqrt{L}), -\pi L]$ and on $[\pi L, \pi(L+\sqrt{L})]$. Fig. 4.2 shows the decay rate of $\log_{10}\left(\frac{|\lambda_L - \lambda_{L_{\rm ref}}|}{\lambda_{L_{\rm ref}}}\right)$, $\log_{10}\left(\frac{\|\phi_L - \phi_{L_{\rm ref}}\|^2_{L^2(\mathbb{R})}}{\|\phi_{L_{\rm ref}}\|^2_{L^2(\mathbb{R})}}\right)$ and $\log_{10}\left(\frac{\|\phi_L - \phi_{L_{\rm ref}}\|^2_{H^1(\mathbb{R})}}{\|\phi_{L_{\rm ref}}\|^2_{H^1(\mathbb{R})}}\right)$. These numerical results show the exponential decay of the error as a function of L, as well as the doubling of the convergence rate of the eigenvalue with respect to the convergence rate of the eigenvector.

The last series of numerical tests aims at testing the effect of numerical integration. For all $L \in \mathbb{N}^*$, we denote by λ_{L,N_L,M_L} the eigenvalue of H_{L,N_L,M_L} which is closest to λ, by u_{L,N_L,M_L} an associated normalized eigenvector, and by $\phi_{L,N_L,M_L} = \chi_L u_{L,N_L,M_L}$ (we choose the sign of u_{L,N_L,M_L} in such a way that $\|\phi_{L,N_L,M_L} - \phi_L\|_{L^2(\mathbb{R}^d)} \simeq 0$). In Fig. 4.3,

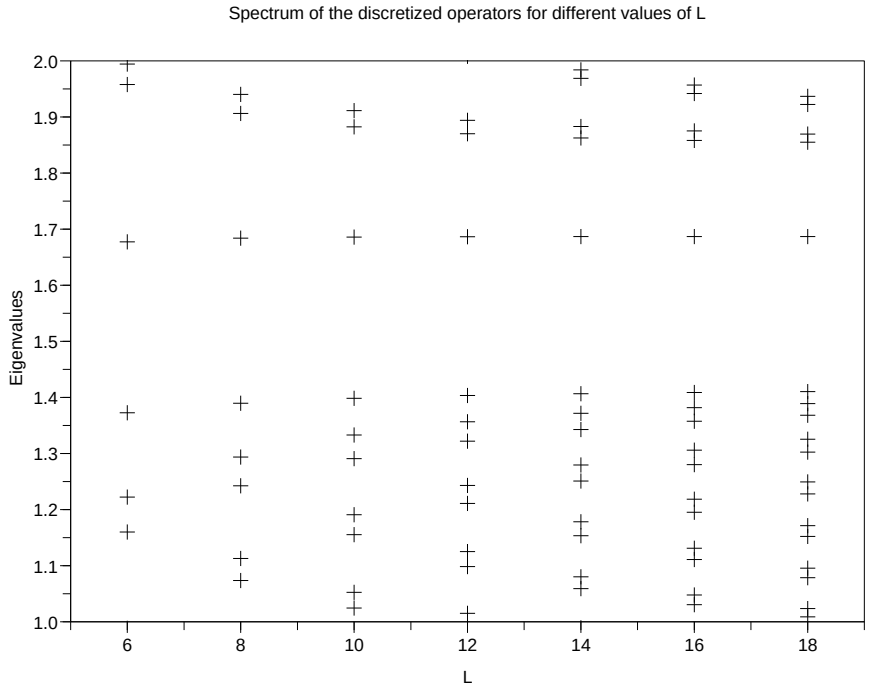

Figure 4.1: Spectrum of $H_{L,N_{\text{ref}}}$ in the range $[1,2]$ for different values of L, with $N_{\text{ref}} = 1400$.

Fig. 4.4 and Fig. 4.5 below are drawn the errors $|\lambda_{L,N_L,M_L} - \lambda_L|$, $\|\phi_{L,N_L,M_L} - \phi_L\|_{L^2(\mathbb{R}^d)}$ and $\|\phi_L - \phi_{L,N_L,M_L}\|_{H^1(\mathbb{R}^d)}$ for the following values:

- $L = 6, 8, 10, 12, 14, 16, 18$,
- $N_L = NL$ where $N = 2, 4, 6, 8, 10, 12, 14$,
- $M_L = ML$ where $M = 56, 112, 224, 448$,

as well as the results obtained with exact integration ($M = \infty$).

4.8 Appendix: Banach-Nečas-Babuška's Theorem and Strang's lemma

In this appendix, we recall the Banach-Nečas-Babuška theorem and the Strang lemma (see e.g. [29, 83]).

Theorem 4.8.1. (Banach-Nečas-Babuška) *Let W be a Banach space and V a reflexive Banach space. Let $a \in \mathcal{L}(W \times V; \mathbb{R})$ and $f \in V'$. Then the problem*

$$\begin{cases} \text{find } u \in W \text{ such that} \\ \forall v \in V, \quad a(u,v) = f(v), \end{cases} \tag{4.42}$$

is well-posed if and only if

- $\exists \alpha > 0, s.t. \inf\limits_{w \in W} \sup\limits_{v \in V} \dfrac{|a(w,v)|}{\|w\|_W \|v\|_V} \geq \alpha;$

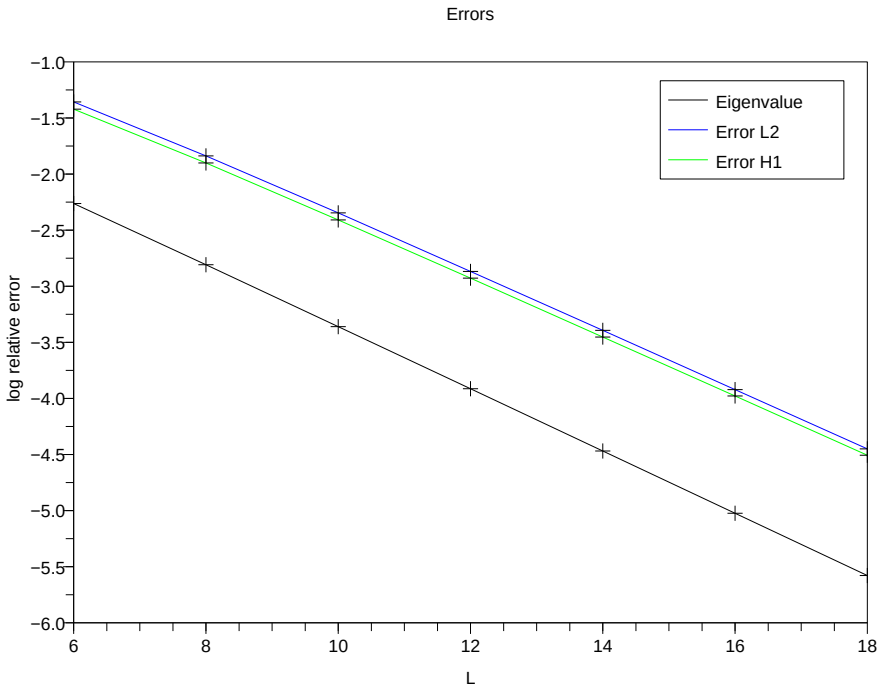

Figure 4.2: Decay rates of $\log_{10}\left(\frac{|\lambda_L - \lambda_{L_{\text{ref}}}|}{\lambda_{L_{\text{ref}}}}\right)$ (Eigenvalue), $\log_{10}\left(\frac{\|\phi_L - \phi_{L_{\text{ref}}}\|^2_{L^2(\mathbb{R})}}{\|\phi_{L_{\text{ref}}}\|^2_{L^2(\mathbb{R})}}\right)$ (Error L2) and $\log_{10}\left(\frac{\|\phi_L - \phi_{L_{\text{ref}}}\|^2_{H^1(\mathbb{R})}}{\|\phi_{L_{\text{ref}}}\|^2_{H^1(\mathbb{R})}}\right)$ (Error H1) for different values of L.

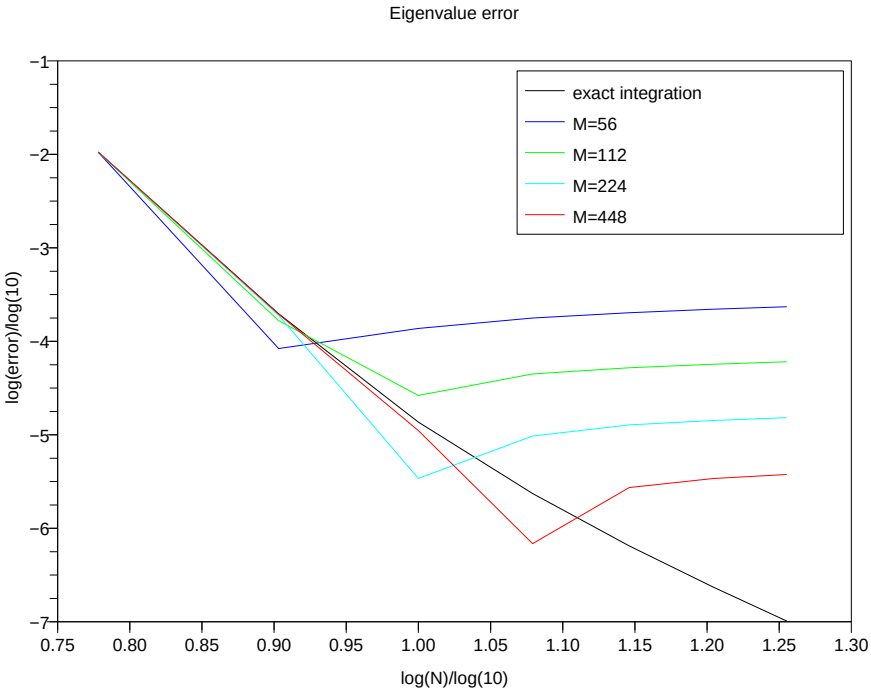

Figure 4.3: Error on the eigenvalue $\log_{10}\left(|\lambda_{L,N_L,M_L} - \lambda_L|\right)$ as a function of $\log_{10}(N)$.

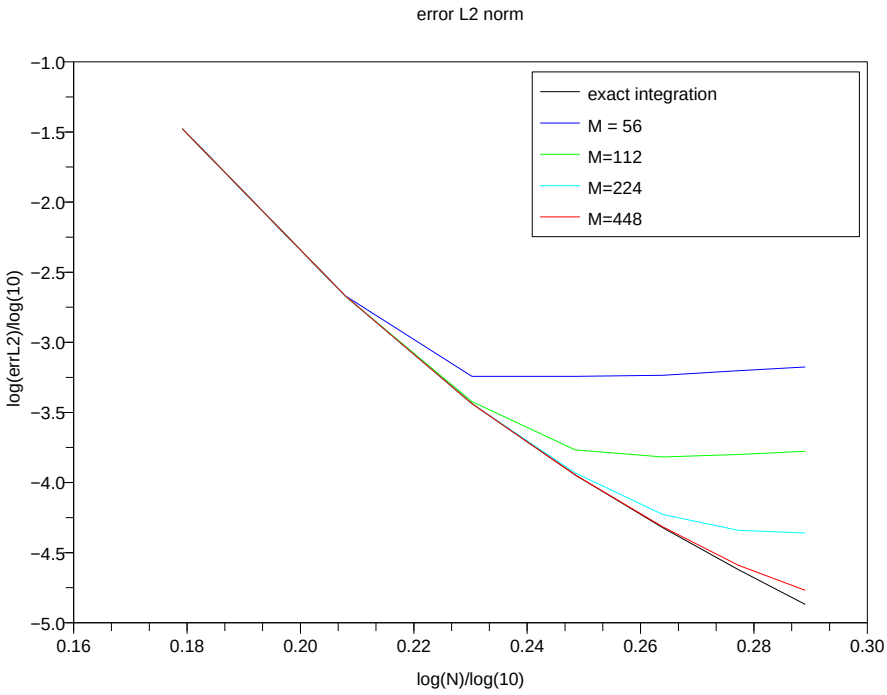

Figure 4.4: Error on the eigenvector $\log_{10}\left(\|\phi_{L,N_L,M_L} - \phi_L\|_{L^2(\mathbb{R}^d)}\right)$ as a function of $\log_{10}(N)$.

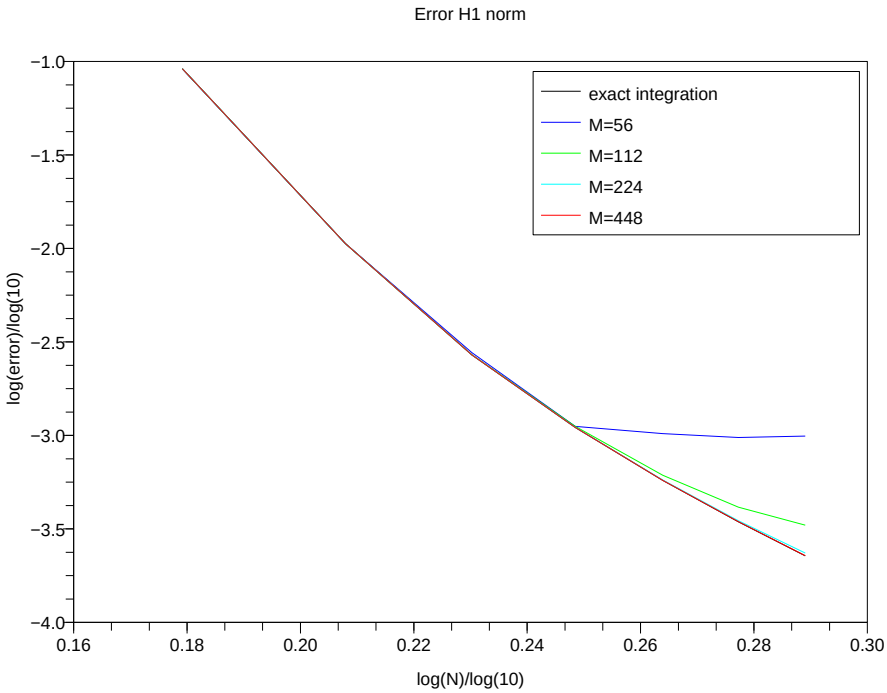

Figure 4.5: Error on the eigenvector $\log_{10}\left(\|\phi_{L,N_L,M_L} - \phi_L\|_{H^1(\mathbb{R}^d)}\right)$ as a function of $\log_{10}(N)$.

- $\forall v \in V,\ (\forall w \in W, \quad a(w,v) = 0) \Rightarrow (v = 0)$.

Moreover, the following a priori estimate holds:

$$\forall f \in V', \quad \|u\|_W \leq \frac{1}{\alpha}\|f\|_{V'}. \tag{4.43}$$

Lemma 4.8.1. (Strang) *Let us consider the following approximate problem*

$$\begin{cases} \text{find } u_n \in W_n \text{ such that} \\ \forall v_n \in V_n, \quad a_n(u_n, v_n) = f_n(v_n), \end{cases} \tag{4.44}$$

and let us assume that

- $W_n \subset W$ *and* $V_n \subset V$*;*
- $\exists \alpha_n > 0$, *s.t.* $\inf\limits_{w_n \in W_n} \sup\limits_{v_n \in V_n} \dfrac{|a_n(w_n, v_n)|}{\|w_n\|_W \|v_n\|_V} \geq \alpha_n$, *and* $\dim(W_n) = \dim(V_n)$*;*
- *the bilinear form* a_n *is bounded on* $W_n \times V_n$.

Then, the following error estimate holds:

$$\begin{aligned} \|u - u_n\|_W \quad \leq \quad & \frac{1}{\alpha_n}\|f - f_n\|_{\mathcal{L}(V_n)} \\ & + \inf_{w_n \in W_n} \left[\left(1 + \frac{\|a\|_{\mathcal{L}(W,V_n)}}{\alpha_n} \right) \|u - w_n\|_W + \frac{1}{\alpha_n} \sup_{v_n \in V_n} \frac{|a(w_n, v_n) - a_n(w_n, v_n)|}{\|v_n\|_V} \right]. \end{aligned}$$

4.9 Appendix: Standard Galerkin methods

Let us first prove that if A is a semibounded operator and if $(X_n)_{n\in\mathbb{N}}$ is a sequence of finite-dimensional subspaces of $Q(A)$ satisfying (A1), then condition (C3') is satisfied.

Proposition 4.9.1. *Let* A *be a semibounded self-adjoint operator on* $\mathcal{H}$ *and* $(X_n)_{n\in\mathbb{N}}$ *a sequence of finite-dimensional subspaces of* $Q(A)$ *satisfying*

$$\forall \phi \in Q(A), \ \inf_{\phi_n \in X_n} \|\phi - \phi_n\|_{Q(A)} \underset{n\to\infty}{\longrightarrow} 0. \tag{4.45}$$

Then, for all compact subsets $K \subset \mathbb{C}$, *there exists* $c_K > 0$ *such that for all* $n \in \mathbb{N}$ *and all* $\mu \in K$,

$$\inf_{w_n \in X_n} \sup_{v_n \in X_n} \frac{|(a - \mu m)(w_n, v_n)|}{\|w_n\|_{Q(A)} \|v_n\|_{Q(A)}} \geq c_K \min(1, \operatorname{dist}(\mu, \sigma(A|_{X_n}))).$$

Proof. Let us consider for instance an operator A bounded from below by a constant $c \in \mathbb{R}$. Then,

$$\begin{cases} Q(A) & \to & \mathbb{R}_+ \\ \phi & \mapsto & \|\phi\|_* \ := \sqrt{(1 + |c|)\|\phi\|_{\mathcal{H}}^2 + a(\phi, \phi)}, \end{cases}$$

defines a norm on $Q(A)$ equivalent to the norm $\|\cdot\|_{Q(A)}$. Thus, there exists a constant $\kappa > 0$ such that for all sequences $(X_n)_{n\in\mathbb{N}}$ of finite-dimensional subspaces of $Q(A)$ satisfying (4.45), all $\mu \in \mathbb{C}$ and all $n \in \mathbb{N}$,

$$\begin{aligned}\inf_{w_n\in V_n}\sup_{v_n\in V_n}\frac{|(a-\mu)(w_n,v_n)|}{\|w_n\|_{Q(A)}\|v_n\|_{Q(A)}} &\geq \inf_{w_n\in V_n}\sup_{v_n\in V_n}\kappa\frac{|(a-\mu)(w_n,v_n)|}{\|w_n\|_*\|v_n\|_*},\\ &\geq \inf_{\nu_n\in\sigma(A|_{V_n})}\kappa\frac{|\nu_n-\mu|}{1+|c|+|\nu_n|}.\end{aligned}$$

By studying the function $f_\mu : \mathbb{R} \ni x \mapsto \kappa\frac{|x-\mu|}{1+|c|+|x|}$, one can easily prove that for all $n \in \mathbb{N}$,

$$\inf_{\nu_n\in\sigma(A|_{V_n})}\kappa\frac{|\nu_n-\mu|}{1+|c|+|\nu_n|} \geq c_\mu \min\left(1, \operatorname{dist}\left(\mu, \sigma(A|_{V_n})\right)\right),$$

with $c_\mu := \frac{\kappa}{2+2|c|+2|\mu|} > 0$. Hence the result. □

However, in the case when A is not semibounded, there exists some sequence $(X_n)_{n\in\mathbb{N}}$ of finite-dimensional subspaces of $Q(A)$ satisfying (A1) such that (C3) is not satisfied. An explicit counter-example is given below.

Example 4.9.1. *Let $\mathcal{H}$ be a separable Hilbert space, $(e_k)_{k\in\mathbb{Z}}$ an orthonormal basis of $\mathcal{H}$ and A the unbounded self-adjoint operator on $\mathcal{H}$ defined on the domain*

$$D(A) := \left\{ v = \sum_{k\in\mathbb{Z}} v_k e_k \,\middle|\, \sum_{k\in\mathbb{Z}} |k|^2|v_k|^2 < \infty \right\}$$

by

$$\forall v \in D(A), \quad Av := \sum_{k\in\mathbb{Z}} k\langle v, e_k\rangle_{\mathcal{H}} e_k.$$

The spectrum of A is purely discrete and each eigenvalue is simple. More precisely, $\sigma(A) = \sigma_d(A) = \mathbb{Z}$ and for all $k \in \mathbb{Z}$, $Ae_k = ke_k$. Note that A can be identified with the momentum operator of a quantum particle in a one-dimensional torus (take $\mathcal{H} = L^2_{\mathrm{per}}((0, 2\pi), \mathbb{C})$, $e_k(x) = (2\pi)^{-1/2}e^{ik\cdot x}$, $D(A) = H^1_{\mathrm{per}}((0, 2\pi), \mathbb{C})$, $A = -i\frac{d}{dx}$).

Let us consider the sequence of finite dimensional spaces $(X_n)_{n\in\mathbb{N}}$ defined as

$$X_n := \mathbb{C}e_{0,n} \oplus \mathbb{C}\widetilde{e}_{0,n} \oplus \operatorname{Span}\left\{e_k,\ 1 \leq |k| \leq n-1\right\},$$

where

$$e_{0,n} := \cos(1/n)e_0 + \frac{\sin(1/n)}{\sqrt{2}}e_n + \frac{\sin(1/n)}{\sqrt{2}}e_{-n}$$

and

$$\widetilde{e}_{0,n} := \frac{1}{\sqrt{2}}e_n - \frac{1}{\sqrt{2}}e_{-n}.$$

The family $(e_{0,n}, \widetilde{e}_{0,n}, (e_k,\ 1 \leq |k| \leq n-1))$ forms an $\mathcal{H}$-orthonormal basis of the discretization space X_n and the matrix of $A|_{X_n}$ in this basis reads

$$\begin{pmatrix} 0 & n\sin(1/n) & 0^T_{\mathbb{R}^{2(n-1)}} \\ n\sin(1/n) & 0 & 0^T_{\mathbb{R}^{2(n-1)}} \\ 0_{\mathbb{R}^{2(n-1)}} & 0_{\mathbb{R}^{2(n-1)}} & \operatorname{diag}(-(n-1), \cdots, -1, 1, \cdots, n-1) \end{pmatrix}.$$

Consequently,

$$\sigma(A|_{X_n}) = (-n\sin(1/n), n\sin(1/n), -(n-1), \cdots, -1, 1, \cdots, n-1),$$

so that

$$\lim_{n\to\infty} \sigma(A|_{X_n}) = \mathbb{Z}^* \subsetneq \mathbb{Z} = \sigma(A).$$

Yet, the vector e_0 *satisfies the condition*

$$\begin{aligned} \left\| \left(1 - \Pi_{X_n}^{Q(A)}\right) e_0 \right\|_{Q(A)} &\leq \|e_0 - e_{0,n}\|_{Q(A)}, \\ &= \sqrt{(1-\cos(1/n))^2 + \sin(1/n)^2 + n\sin(1/n)^2}, \\ &\underset{n\to+\infty}{\sim} n^{-1/2} \underset{n\to\infty}{\longrightarrow} 0, \end{aligned}$$

and it straightforwardly follows from the above result that

$$\forall \phi \in Q(A), \quad \inf_{\phi_n \in X_n} \|\phi - \phi_n\|_{Q(A)} \underset{n\to\infty}{\longrightarrow} 0.$$

As $0 \in \sigma_d(A)$ *whereas* $0 \notin \overline{\lim}_{n\to\infty} \sigma(A|_{X_n})$*, Corollary 4.3.1 implies that condition (C3) does not hold.*

Part II

Uncertainty quantification and greedy algorithms

Chapter 5

Uncertainty quantification and high-dimensional problems

5.1 Introduction

The work presented in this chapter is motivated by a collaboration with the Michelin company on uncertainty quantification (UQ).

Let us first introduce the notation which will be used throughout this chapter. The modeling of uncertainties is achieved through the definition of a suitable probability space $(\Omega, \mathcal{B}, \mathbb{P})$ where Ω denotes the space of elementary events, $\mathcal{B}$ a σ-algebra defined on Ω and $\mathbb{P}$ a probability measure. Elements (or events) of Ω will be denoted by $\omega \in \Omega$. Let $p, d \in \mathbb{N}^*$. Random variables will be denoted with capital letters by $T_1, \cdots, T_p$ and take their values in $\mathcal{T}_1, \cdots, \mathcal{T}_p$ where $\mathcal{T}_l$ is an open subset of $\mathbb{R}$ for all $1 \leq l \leq p$. By default, the random vector $T := (T_1, \cdots, T_p)$ takes its values in $\mathcal{T} := \mathcal{T}_1 \times \cdots \times \mathcal{T}_p \subset \mathbb{R}^p$. When it is precised at the beginning of a section, $\mathcal{T}$ may also refer to a general open subset of $\mathbb{R}^p$. We will denote with small letters $t, t_1, \cdots, t_p$ elements of $\mathcal{T}, \mathcal{T}_1, \cdots, \mathcal{T}_p$ respectively. Deterministic variables in the system (such as space, time, etc.) will be denoted by $x \in \mathcal{X}$ where $\mathcal{X} \subset \mathbb{R}^d$ is an open subset and $d \in \mathbb{N}^*$. The notation V, V_x, V_X, V_l with $1 \leq l \leq p$ will refer to Hilbert spaces of functions which will respectively depend on $(x, t) \in \mathcal{X} \times \mathcal{T}$, $x \in \mathcal{X}$, $t \in \mathcal{T}$, $t_l \in \mathcal{T}_l$. They will respectively be endowed with the scalar products $\langle \cdot, \cdot \rangle_V$, $\langle \cdot, \cdot \rangle_{V_x}$, $\langle \cdot, \cdot \rangle_{V_t}$, $\langle \cdot, \cdot \rangle_{V_l}$ and their associated norms will be denoted by $\| \cdot \|_V$, $\| \cdot \|_{V_x}$, $\| \cdot \|_{V_t}$, $\| \cdot \|_{V_l}$.

Let us now turn to the description of the problem posed by Michelin. The fabrication process of a tyre is a sequence of intricate steps. Even controlled, some uncertainties remain on the characteristics of the materials, pretension states or the exact location of the interfaces after molding and vulcanization. As a result, some dispersion is inevitably seen on various final performances such as wear for instance. Such a deviation is potentially accentuated by the diversity of the microscopic road profiles, possible road hazards, or various driving styles.

These sources of uncertainties can be modeled by a large number of random variables $T = (T_1, \cdots, T_p)$, which influence the mechanical behavior of the tyre.

Michelin is interested in studying the dependency of the behavior of a tyre on the random variables, and more precisely of output quantities reflecting tyre performance such as the mean and variance of the pressure on the part of the tyre in

contact with the soil.

In a Lagrangian approach of solid mechanics, the actual configuration $\mathcal{X}' \subset \mathbb{R}^3$ of the solid is considered with respect to a reference configuration $\mathcal{X} \subset \mathbb{R}^3$. Let (e_1, e_2, e_3) denote the canonical basis of $\mathbb{R}^3$. Let us introduce u the function which, to each particle located at the position $x = (x_1, x_2, x_3) \in \mathcal{X}$ in the reference configuration, associates its position $x' = (x'_1, x'_2, x'_3) = u(x) \in \mathcal{X}'$ in the actual configuration. Let us assume that the soil is modeled for the sake of simplicity by a plane located at the altitude $x_3 = 0$ (thus without uncertainty).

The problem solved by Michelin in order to model the stationary behavior of a tyre for a given value of the random vector T can be written under the following form

$$u(T) = \underset{v \in \mathcal{K}}{\operatorname{argmin}} \int_{\mathcal{X}} W(\nabla v, T) - \int_{\mathcal{X}} f(T)v, \tag{5.1}$$

where $W(\nabla v, T)$ represents the hyper-elastic energy of the system, $f(T)$ the external volume forces that acts on the tyre, and

$$\mathcal{K} := \{v \in V,\ v \cdot e_3|_\Gamma \geq 0\},$$

where V is the Hilbert space of admissible displacements and Γ the part of the frontier of the tyre which is likely to be in contact with the soil.

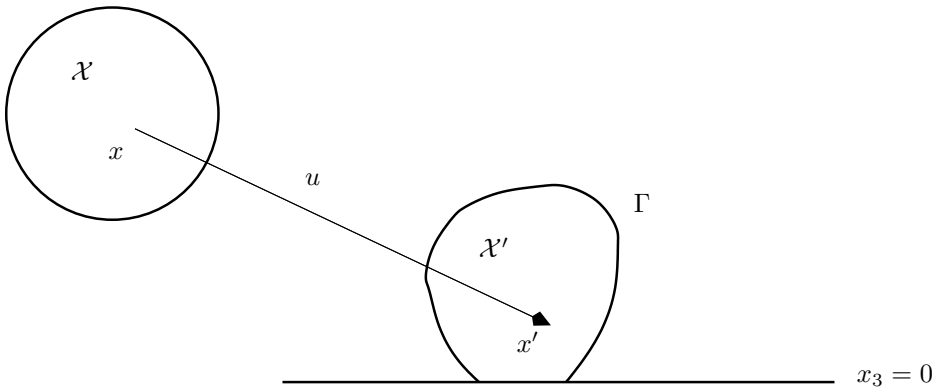

Figure 5.1: The tyre obstacle problem

This leads to a very high-dimensional problem, namely a parametrized obstacle (thus nonlinear) problem.

There exists a wide variety of approaches for the resolution of partial differential equations with random coefficients. A review of the most classical methods is presented in Section 5.2.

Unfortunately, when the number of random variables is very large, most of these approaches are limited. Statistical methods, such as Monte-Carlo algorithms, can always be used to compute expectations of output quantities of interest. Their convergence is in general very slow though. Getting a good reduced order model is one way to circumvent this slow convergence.

The main focus of this thesis work targets the development of numerical methods in order to compute efficient reduced-order model for

high-dimensional problems. In the case when the number of random variables is very large, standard algorithms cannot be carried out in practice because of the so-called *curse of dimensionality* [14].

A way to understand the curse of dimensionality is the following. Assume for the sake of simplicity that $\mathcal{X} = \mathcal{T}_1 = \cdots = \mathcal{T}_p = [0,1]$. Let $u : [0,1]^{p+1} \to \mathbb{R}$ be a C^m function with $m \in \mathbb{N}^*$. We would like to reconstruct the function u from an ensemble of N values $\{u(y_i)\}_{1 \leq i \leq N}$ where $y_1, \cdots, y_N \in [0,1]^{p+1}$. In this case, it is well-known that if $(y_i)_{1 \leq i \leq N}$ are the nodes of a uniform grid of $[0,1]^{p+1}$ with mesh size $h > 0$, and if a polynomial reconstruction scheme is used, then

$$\|u - R(u)\|_{L^\infty([0,1]^{p+1})} \leq Ch^m,$$

where $C > 0$ is a constant independent on h, and $R(u)$ denotes the reconstructed function. Since the number of sample points N scales like $h^{-(p+1)}$, the approximation error reads

$$\|u - R(u)\|_{L^\infty([0,1]^{p+1})} \leq CN^{-m/(p+1)}.$$

Thus, the higher the dimension, the slower the decay rate of the reconstruction error with respect to the number of sample points N.

Actually, it is proved in [74] that it is impossible to design reconstruction schemes which would achieve better results. This can be explained in terms of nonlinear width. Let L be a normed space with associated norm $\|\cdot\|_L$ and $K \subset L$. Let us consider continuous maps $E : K \to \mathbb{R}^N$ (encoding) and $R : \mathbb{R}^N \to L$ (reconstruction). The distortion of the pair (E, R) over K is defined as

$$\sup_{u \in K} \|u - R(E(u))\|_L,$$

i.e., it is the largest error made for all functions $u \in K$ by the encoding-reconstruction scheme. The nonlinear N-width of K is defined as the infimum of the distortion of all pairs of continuous maps (E, R):

$$d_N(K) := \inf_{\left.\begin{array}{c} E : K \to \mathbb{R}^N \\ R : \mathbb{R}^N \to L \end{array}\right\} \text{continuous}} \sup_{u \in K} \|u - R(E(u))\|_L.$$

Then it is known that in the case when $L = L^\infty([0,1]^{p+1})$ and

$$K = \left\{u \in C^m([0,1]^{p+1}) \mid \forall \alpha \in \mathbb{N}^{p+1},\ |\alpha| \leq m,\ \|\partial^\alpha u\|_{L^\infty([0,1]^{p+1})} \leq 1\right\}$$

is the unit ball of $C^m([0,1]^{p+1})$, then there exists $c, C > 0$ independent on p such that for all $N \in \mathbb{N}^*$,

$$cN^{-m/(p+1)} \leq d_N(K) \leq CN^{-m/(p+1)}.$$

In other words, if one wants to approximate a function $u \in C^m([0,1]^{p+1})$ so that the relative error is lower than a given threshold $\varepsilon > 0$, the number N of samples will necessarily scale exponentially with respect to the dimension $p + 1$.

In our UQ context, the Galerkin approximation is a typical example of a method whose complexity scales badly with the number of random parameters. Indeed, basis

functions are typically constructed as the tensorization of univariate function bases $\{\phi_i(x)\}_{1\leq i\leq N_x}$, $\{\psi^{(1)}_{j_1}(t_1)\}_{1\leq j_1\leq N_1}, \cdots, \{\psi^{(p)}_{j_p}(t_p)\}_{1\leq j_p\leq N_p}$ of the spaces $V_x, V_1, \cdots, V_p$. The approximation of a function $u \in V = V_x \otimes V_1 \otimes \cdots \otimes V_p$ is given under the following form

$$u(x, t_1, \cdots, t_p) \approx \sum_{1\leq i\leq N_x, 1\leq j_1\leq N_1, \cdots, 1\leq j_p\leq N_p} \lambda_{i,j_1,\cdots,j_p} \phi_i(x)\psi^{(1)}_{j_1}(t_1)\cdots\psi^{(p)}_{j_p}(t_p),$$

where $\{\lambda_{i,j_1,\cdots,j_p},\ 1 \leq i \leq N_x,\ 1 \leq j_1 \leq N_1, \cdots, 1 \leq j_p \leq N_p\}$ are real numbers to be determined. The size of the discretized problem is then equal to $N_x N_1 \cdots N_p$. In the case when $N_1 = \cdots = N_p = N$, the size is equal to $N_x N^p$ and scales exponentially with the number of random parameters.

That is the reason why we chose in this thesis work to study a class of algorithms, called *Progressive Generalized Decomposition* (PGD), which were introduced in different contexts by Ladevèze [123], Chinesta [5], and particularly adapted by Anthony Nouy [155] to the context of UQ. These methods are closely related to the greedy algorithms [174] used in nonlinear approximation. An introduction to general greedy algorithms and their application to high-dimensional problems are presented in Section 5.3. We will detail the contributions of this thesis work in Section 5.4. Lastly, in Appendix 5.5, we give a short and non-exhaustive review of other numerical methods used in the context of high-dimensional problems.

5.2 Classical uncertainty quantification methods

We present here a short review of the main classical methods used in the field of UQ. We refer to [154] for a very good and more exhaustive introduction to this subject.

5.2.1 Partial differential equations with stochastic coefficients

For many physical problems, the response of the system under consideration can be modeled via a Partial Differential Equation with Stochastic Coefficients (PDESC). The response u of such a model is a random field $(u(x;\omega))_{x\in\mathcal{X}}$ satisfying almost surely a set of equations

$$\mathcal{A}(u(x;\omega); c(x;\omega)) = b(x;\omega) \quad \text{in } \mathcal{D}'(\mathcal{X}), \tag{5.2}$$

where $\mathcal{A}$ denotes a differential operator, $(c(x;\omega))_{x\in\mathcal{X}}$ and $(b(x;\omega))_{x\in\mathcal{X}}$ are random fields denoting respectively the stochastic coefficients the PDESC (5.2) depends on, and the right-hand side associated with the source terms.

Dealing with a general PDESC of the form (5.2) is an intricate task. The fact that the response of the system depends on the two random fields $(c(x,\omega))_{x\in\mathcal{X}}$ and $(b(x,\omega))_{x\in\mathcal{X}}$ is an inherent difficulty of the system. Indeed, the characterization of a stochastic field requires the determination of a large number of random variables (possibly infinite and even uncountable). Let z denote either c or b. Actually, a random field z can be completely characterized by its finite dimensional probability laws, which are the joint probability laws of all finite sets of random variables $\{z(x_1;\omega), \cdots, z(x_m;\omega)\}$, with $m \in \mathbb{N}^*$ and $x_1, \cdots, x_m \in \mathcal{X}$ [3].

However, in the case when z is a square-integrable random field, i.e. in the case when $z \in Z_x \otimes L^2(\Omega, \mathbb{P})$ where Z_x is a Hilbert space of functions defined on $\mathcal{X}$, it is possible to characterize z with a countable number of random variables. Besides, a priori error estimates are available in order to evaluate the quality of the approximation when only a finite number of variables is considered. In this case, we can hopefully approximate accurately a general PDESC of the form (5.2) by a set of equations of the form

$$\mathcal{A}(u(x;\omega);x,T(\omega)) = b(x;T(\omega)) \quad \text{in } \mathcal{D}'(\mathcal{X}), \tag{5.3}$$

satisfied almost surely and where $T(\omega) = (T_1(\omega), \cdots, T_p(\omega))$ is a random vector. The wide majority of methods for uncertainty propagation actually concerns the approximation of PDESC of the form (5.3).

A large variety of discretization techniques for square-integrable random fields z can be found in the literature. We present here the two approaches that are most widely used in standard UQ algorithms.

Karhunen-Loève expansion

The *Hilbert Karhunen-Loève* (HKL) decomposition [115, 143, 78, 132] is an extension of the so-called *Karhunen-Loève decomposition* first introduced in the case when $Z_x = L^2(\mathcal{X})$. The space $Z_x \otimes L^2(\Omega, \mathbb{P})$ is endowed with the natural inner product

$$\forall z_1, z_2 \in Z_x \otimes L^2(\Omega, \mathbb{P}), \quad \langle z_1, z_2 \rangle_{Z_x \otimes L^2(\Omega,\mathbb{P})} = \mathbb{E}\left[\langle z_1(.,\omega), z_2(.,\omega)\rangle_{Z_x}\right],$$

where $\langle \cdot, \cdot \rangle_{Z_x}$ is the inner product of Z_x.

For almost all $x \in \mathcal{X}$, let $\mu_z(x) = \mathbb{E}[z(x,\omega)]$ be the mean-value of z at a point $x \in \mathcal{X}$. We define the *covariance function* of the random process z as a function $C_z \in Z_x \otimes Z_x$ where

$$\text{for almost all } x, y \in \mathcal{X}, \quad C_z(x,y) = \mathbb{E}\left[(z(x,\omega) - \mu_z(x))(z(y,\omega) - \mu_z(y))\right].$$

Let us also introduce the linear operator T_z from Z_x to Z_x defined for all $\widetilde{z} \in Z_x$ and almost all $x \in \mathcal{X}$ by

$$T_z(\widetilde{z})(x) = \langle C_z, \widetilde{z}\rangle_{Z_x}(x) := \mathbb{E}\left[(z(x,\omega) - \mu_z(x))\langle z(\cdot,\omega) - \mu_z(\cdot), \widetilde{z}\rangle_{Z_x}\right].$$

Under the assumption that the covariance function C_z is regular enough (for instance, has analytic smoothness in x and y), T_z defines a continuous self-adjoint positive semi-definite and compact operator on Z_x and can then be decomposed using spectral theory [163]. There exists $(\sigma_i)_{i\in\mathbb{N}^*}$ a non-increasing sequence of positive real numbers converging to 0 and $(e_i)_{i\in\mathbb{N}^*}$ an orthonormal family of Z_x such that

$$\forall i \in \mathbb{N}^*, \quad T_z e_i = \sigma_i e_i.$$

Then, by defining for all $i \in \mathbb{N}^*$,

$$T_i(\omega) := \frac{1}{\sqrt{\sigma_i}}\langle z(\cdot,\omega) - \mu_z, e_i\rangle_{Z_x},$$

the family $(T_i)_{i\in\mathbb{N}^*}$ forms a family of square-integrable centered uncorrelated random variables with unit variance. Such a spectral decomposition actually holds under weaker assumptions than analyticity of the covariance function, see for example [58].

The HKL decomposition then consists in decomposing the random process z as

$$z(x,\omega) = \mu_z(x) + \sum_{i=1}^{\infty} \sqrt{\sigma_i} e_i(x) T_i(\omega). \tag{5.4}$$

This expansion is convergent in $Z_x \otimes L^2(\Omega,\mathbb{P})$ in the sense that

$$\left\| z - \mu_z - \sum_{i=1}^{n} \sqrt{\sigma_i} e_i \otimes T_i \right\|^2_{Z_x\otimes L^2(\Omega,\mathbb{P})} = \|z-\mu_z\|^2_{Z_x\otimes L^2(\Omega,\mathbb{P})} - \sum_{i=1}^{n} \sigma_i \underset{n\to\infty}{\longrightarrow} 0.$$

One can then obtain a discretized version of the random process by truncating the decomposition

$$z(x,\omega) \approx z_n(x,\omega) = \mu_z(x) + \sum_{i=1}^{n} \sqrt{\sigma_i} e_i(x) T_i(\omega). \tag{5.5}$$

The truncated expansion (5.5) is the optimal decomposition of the random process z with respect to the natural norm in $Z_x \otimes L^2(\Omega,\mathbb{P})$ over the set of approximations of z under the form $\mu_z(x) + \sum_{i=1}^{n} w_i(x) S_i(\omega)$, with $w_i \in Z_x$ and $S_i \in L^2(\Omega,\mathbb{P})$. In other words,

$$\|z - z_n\|^2_{Z_x\otimes L^2(\Omega,\mathbb{P})} = \min_{(w_i)_{1\le i\le n}\in Z_x^n,\ (S_i)_{1\le i\le n}\in L^2(\Omega,\mathbb{P})^n} \left\| z - \mu_z - \sum_{i=1}^{n} w_i \otimes S_i \right\|^2_{Z_x\otimes L^2(\Omega,\mathbb{P})}.$$

Polynomial Chaos expansion

However, the covariance function C_z and hence the families $(e_i)_{i\in\mathbb{N}^*}$, $(T_i)_{i\in\mathbb{N}^*}$ and $(\sigma_i)_{i\in\mathbb{N}^*}$ involved in the HKL expansion of the process z may be difficult to compute in practice. We present here another very common discretization approach, the *polynomial chaos* (PC) expansion. As suggested by Wiener [182], any random variable in $L^2(\Omega,\mathbb{P})$ can be represented as a series of polynomials of independent standard Gaussian random variables.

Let us denote by $\mathbf{G} = (G_i)_{i\in\mathbb{N}^*}$ a countable set of independent standard Gaussian random variables. For all $\beta \in \mathbb{N}$, let us denote by $h_\beta : \mathbb{R} \to \mathbb{R}$ the one-dimensional Hermite polynomial of degree β. The family $(h_\beta)_{\beta\in\mathbb{N}}$ then forms an orthonormal basis of $L^2(\mathbb{R},\phi)$ where ϕ denotes the gaussian measure $\phi(dg) = \frac{1}{\sqrt{2\pi}} e^{-g^2/2}\, dg$. In other words,

$$\mathbb{E}\left[h_\beta(G_1) h_{\beta'}(G_1)\right] = \delta_{\beta\beta'},\ \forall \beta,\beta' \in \mathbb{N}.$$

Let us now introduce the set of finite-length multi-indices

$$\mathcal{I} = \left\{ \alpha = (\beta_i)_{i\in\mathbb{N}^*} \in \mathbb{N}^{\mathbb{N}^*},\ |\alpha| = \sum_{i\in\mathbb{N}^*} \beta_i < +\infty \right\}.$$

For all $\alpha = (\beta_i)_{i\in\mathbb{N}^*} \in \mathcal{I}$, multidimensional Hermite polynomials can be written as

$$H_\alpha(\mathbf{g}) = \Pi_{i=1}^{\infty} h_{\beta_i}(g_i),$$

where $\mathbf{g} = (g_i)_{i\in\mathbb{N}^*} \in \mathbb{R}^{\mathbb{N}^*}$. The family $(H_\alpha(\mathbf{G}))_{\alpha\in\mathcal{I}}$ then forms an orthonormal family in the sense that

$$\mathbb{E}\left[H_\alpha(\mathbf{G})H_{\alpha'}(\mathbf{G})\right] = \delta_{\alpha\alpha'},\ \forall \alpha, \alpha' \in \mathcal{I}.$$

The homogeneous chaos of degree p is the space

$$\mathcal{M}_p := \mathrm{Span}\{H_\alpha(\mathbf{g}); \alpha \in \mathcal{I}, |\alpha| = p\},$$

and the polynomial chaos of degree p is defined by

$$\bigoplus_{k=0}^{p} \mathcal{M}_k.$$

The following orthogonal decomposition holds

$$L^2(\Omega, \mathbb{P}) = \overline{\bigoplus_{k=1}^{\infty} \mathcal{M}_k}.$$

In other terms, the set of polynomials $(H_\alpha(\mathbf{G}))_{\alpha\in\mathcal{I}}$ forms an orthonormal basis of $L^2(\Omega, \mathbb{P})$. A stochastic process $z \in Z_x \otimes L^2(\Omega, \mathbb{P})$ can therefore be decomposed as follows (compare with (5.4))

$$z(x, \omega) = \sum_{\alpha\in\mathcal{I}} z_\alpha(x) H_\alpha(\mathbf{G}(\omega)),$$

where

$$z_\alpha(x) := \mathbb{E}\left[z(x, \omega)H_\alpha(\mathbf{G}(\omega))\right].$$

The Cameron-Martin theorem [34] states that this decomposition converges in $Z_x \otimes L^2(\Omega, \mathbb{P})$. An approximation of the random process can then be obtained by truncating the polynomial chaos to a finite degree and a finite number of Gaussian random variables.

The PC decomposition can be generalized in the following way: if the family of independent identically distributed random variables $\mathbf{G}$ is no more assumed to have a standard Gaussian distribution, can we still find polynomial families $(H_\alpha)_{\alpha\in\mathbb{N}}$ that allow similar expansions of a square-integrable process, i.e. such that $(H_\alpha(\mathbf{G}))_{\alpha\in\mathbb{N}}$ forms a Hilbertian basis of $L^2(\Omega, \mathbb{P})$? The answer is positive. For example, when the variables $\mathbf{G}$ are assumed to be independent variables uniformly distributed on the interval $[0, 1]$, it can be proved that the corresponding orthonormal family of polynomials are the Legendre polynomials.

More generally, Xiu and Karnadakis [183] have proved that for a large class of probability distributions, the corresponding polynomial families, which are called *generalized Polynomial Chaos* (gPC), can be determined using an Askey scheme. We refer to [178] for more correspondences between polynomial families and probability laws.

Uncertainty quantification problem

Thus, HKL or PC expansions enable us to represent the coefficients $(c(x,\omega))_{x\in\mathcal{X}}$ and $(b(x,\omega))_{x\in\mathcal{X}}$ appearing in PDESC (5.2) as random fields of the form $(c(x,T(\omega)))_{x\in\mathcal{X}}$ and $(b(x,T(\omega)))_{x\in\mathcal{X}}$ where $T(\omega) = (T_1(\omega),\cdots,T_p(\omega))$ provided that the random fields are regular enough (typically in $Z_x \otimes L^2(\Omega,\mathbb{P})$ with Z_x a Hilbert space of functions depending on $x \in \mathcal{X}$).

Thus, the PDESC (5.2) can be rewritten as a set of equations of the form

$$\mathcal{A}(u(x;\omega);x,T(\omega)) = b(x,T(\omega)) \quad \text{in } \mathcal{D}'(\mathcal{X}),$$

which are satisfied almost surely and where the randomness is modeled through the finite-size vector $T(\omega)$.

In general, the process $(u(x,\omega))_{x\in\mathcal{X}}$ is then measurable with respect to the σ-algebra spanned by the random vector $T(\omega)$. Then, there exists a measurable function $\widetilde{u}$ on $\mathcal{X}\times\mathcal{T}$ [27] such that, almost surely,

$$u(x,\omega) = \widetilde{u}(x,T(\omega)), \text{ for almost all } x \in \mathcal{X}.$$

In the sequel, the function $\widetilde{u} : \mathcal{X}\times\mathcal{T} \to \mathbb{R}$ will also be denoted by u for the sake of simplicity.

Thus, solving equations of the form (5.3) amounts to solving almost surely the set of equations

$$\mathcal{A}(u(x;X(\omega));x,T(\omega)) = b(x,T(\omega)) \quad \text{in } \mathcal{D}'(\mathcal{X}). \tag{5.6}$$

UQ methods deal with the practical resolution of equations of the form (5.6).

There are three main fields of application of UQ methods, namely, the computation of

- statistical moments, such as mean or variance, of certain output quantities of the mode;
- the probability of rare events, linked to reliability issues;
- a response surface to the model.

Prototypical examples of output quantities, whose mean values are of interest, are the following:

$$Y = \int_{\mathcal{X}} u(x,T)\,dx \quad \text{or} \quad Y = u(a,T) \text{ for some } a \in \mathcal{X}.$$

More generally, such an output can be expressed as a function of the random vector T: $Y = f(T)$, with $f : \mathcal{T} \to \mathbb{R}$. Computing the expectation $\mathbb{E}[Y]$ of the output Y is usually done via *Monte-Carlo* algorithms, which will be detailed in Section 5.2.2. The speed of convergence of these methods heavily relies on the use of good approximations of the solution u to design efficient variance reduction techniques.

Computing rare event probabilities usually involves the so-called *reliability* methods. These algorithms aim at obtaining good estimates of probabilities of the form

$\mathbb{P}[Y > s]$ where s is a given threshold value so that $\mathbb{P}[Y > s]$ is extremely small. Naive Monte-Carlo methods are not adequate in this particular case since they would require a huge number of samples. The principle of a reliability method consists in efficiently "exploring" the so-called failure domain $G = \{t \in \mathcal{T}, y = f(t) > s\}$. Information on the geometry of this complicate set is usually obtained via efficient reduced-order models. These methods will be presented in Section 5.2.3.

Lastly, the construction of accurate response surface models consists in directly computing functional representations of the function $u : \mathcal{X} \times \mathcal{T} \mapsto u(x,t)$, generally under the form

$$u(x,t) \approx \sum_{i=1} u_i(x)\Phi_i(t).$$

If a good reduced-order model for the function u is known, classical approaches can then be carried out in order to compute means, variances, rare event probabilities or even estimation of the full probability distribution of an output quantity Y. Section 5.2.4 will be devoted to the most classical methods used in UQ to derive such approximations.

Before presenting the above methods, let us introduce here a distinction which is crucial in the field of UQ: the notion of *intrusive* and *non-intrusive* methods. In any of the algorithms presented below, an approximation of the solution u is always computed under the form

$$u(x,t) \approx \sum_{i=1} u_i(x)\Phi_i(t),$$

where the functions $(u_i)_{1\leq i\leq n}$ only depend on the deterministic variable $x \in \mathcal{X}$ and the functions $(\Phi_i)_{1\leq i\leq n}$ only depend on the stochastic variable $t \in \mathcal{T}$. The problems which determine the functions $(u_i)_{1\leq i\leq n}$ (respectively the functions $(\Phi_i)_{1\leq i\leq n}$) are called the *deterministic* problems (respectively the *stochastic* problems).

Non-intrusive algorithms are methods where the deterministic problems to solve to determine the functions $(u_i)_{1\leq i\leq n}$ only require the resolution of equations of the form

$$\mathcal{A}(u(x,t_i);x,t_i) = b(x,t_i) \text{ in } \mathcal{D}'(\mathcal{X}), \tag{5.7}$$

for all $1 \leq i \leq n$ for some particular values $t_i \in \mathcal{T}$ of the random vector T. In numerous practical cases, a black-box deterministic code solving problems of type (5.7) is available but cannot be modified. The terminology "non-intrusive" refers to the fact that such methods can be applied using the black-box code.

Intrusive methods, on the contrary, require the resolution of deterministic problems which are not of the form (5.7) and thus imply the modification of the available deterministic codes.

5.2.2 Monte-Carlo methods

Monte-Carlo (MC) methods [33] are among the most widely used methods in order to compute the mean or the variance of a given random variable. In our context, the MC method consists in drawing n independent samples of the random variable T:

$t_1, \cdots, t_n \in \mathcal{T}$. For each $k \in \{1, \cdots, n\}$, the quantity $y_k = f(t_k)$ is evaluated. The mean of Y, i.e. $\mathbb{E}[Y] = \mathbb{E}[f(T)]$, is then approximated by the empirical estimator

$$\widehat{f}_n^{MC} := \frac{y_1 + \cdots + y_n}{n}.$$

The variance of this estimator is $\mathrm{Var}(\widehat{f}_n^{MC}) = \frac{1}{n}\mathrm{Var}(Y)$. Although the algorithm is always implementable, even for high-dimensional problems, the rate of convergence is very slow, of the order of $\mathcal{O}\left(\frac{\sqrt{\mathrm{Var}(Y)}}{\sqrt{n}}\right)$.

There exists a very large number of improvements of this MC method. Better sampling errors may be achieved through the choice of more efficient sampling strategies, for example the *Latin Hypercube Sampling* [148] or Quasi Monte-Carlo methods [33].

Other techniques aim at decreasing the variance of the empirical estimator. We present here a few of them.

One of them is the so-called *control variate* technique. This method consists in using a reduced-order model $f_r(T)$ whose evaluations are much cheaper from a computational point of view than for the true model $f(T)$. Thus, the mean $I_r := \mathbb{E}\left[f_r(T)\right]$ is hopefully much easier to compute. The empirical estimator used for $\mathbb{E}\left[f(T)\right]$ is then

$$\widehat{f}_n^{CV} := I_r + \frac{1}{n}\sum_{k=1}^{n}(f(t_k) - f_r(t_k)),$$

whose variance is given by: $\mathrm{Var}(\widehat{f}_n^{CV}) = \frac{1}{n}\mathrm{Var}\left[f(T) - f_r(T)\right]$. Hopefully, if the reduced-order model f_r is a good approximation of the complete model f, the variance $\mathrm{Var}\left[f(T) - f_r(T)\right]$ can be significantly lower than the variance of the original problem $\mathrm{Var}\left[f(T)\right]$ and the sampling error is drastically reduced.

Another variance reduction technique is the so-called *importance sampling* method. It is used in cases when it is known that the main contribution to the mean $\mathbb{E}\left[f(T)\right]$ are due to particular values of T. The original MC method uses samples $(t_k)_{1\leq k\leq n}$ that are evaluated using the original probability distribution $\mathbb{P}$ of the random variable T. The importance sampling method rather uses samples $(t_k)_{1\leq k\leq n}$ drawn according to a biased probability distribution $\mathbb{P}_{IS}$ which tends to make the most important realizations of T more probable. The empirical estimator is given by

$$\widehat{f}_n^{IS} := \frac{1}{n}\sum_{k=1}^{n} f(t_k)\left(\frac{d\mathbb{P}}{d\mathbb{P}_{IS}}\right)(t_k).$$

The variance of this estimator is

$$\mathrm{Var}(\widehat{f}_n^{IS}) = \frac{1}{n}\left(\mathbb{E}\left[f(T)^2\left(\frac{d\mathbb{P}}{d\mathbb{P}_{IS}}\right)(T)\right] - \mathbb{E}\left[f(T)\right]^2\right).$$

If $\mathbb{P}_{IS}$ is judiciously chosen, the variance can be drastically reduced. However, this method requires to guess a good biased probability measure $\mathbb{P}_{IS}$ which is not always possible in practice. Here again, this choice can be oriented through the use of an accurate reduced-order model.

Another method is *stratified sampling*. The samples $(t_k)_{1\leq k\leq n}$ are drawn according to the original probability distribution conditionnally to being in some specified "stratas". This kind of method is inspired by techniques used for polls. In other words, we use a partition of the set $\mathcal{T}$ as follows: $\mathcal{T} = \bigcup_{i=1}^{m} \mathcal{Q}_i$ and we assume that the probability $p_i = \mathbb{P}(T \in \mathcal{Q}_i)$ is known. The mean $\mathbb{E}[f(T)]$ can be rewritten

$$\mathbb{E}[f(T)] = \sum_{i=1}^{m} \mathbb{E}[f(T)|T \in \mathcal{Q}_i]\, \mathbb{P}(T \in \mathcal{Q}_i).$$

Each of the m quantities $I_i = \mathbb{E}[f(T)|T \in \mathcal{Q}_i]$ can then be estimated using an original Monte-Carlo method with n_i samplings:

$$\widehat{I}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} f(t_j),$$

where the samples $(t_j)_{1\leq j\leq n_i}$ are independently and identically distributed with respect to the conditional probability law of the random vector $\mathbb{E}[T|T \in \mathcal{Q}_i]$. The empirical estimator is given by

$$\widehat{f}_n^{SS} := \sum_{i=1}^{m} \widehat{I}_i p_i.$$

Its variance is $\mathrm{Var}(\widehat{f}_n^{SS}) = \sum_{i=1}^{m} p_i^2 \frac{\sigma_i^2}{n_i}$ where $\sigma_i^2 = \mathrm{Var}(f(T)|T \in \mathcal{Q}_i)$. The choice of the number of samplings n_i can be done freely under the constraint $\sum_{i=1}^{m} n_i = n$. In the case of a *proportional stratification*, the choice is $n_i = p_i n$ and the variance of the empirical estimator is then lower that the variance of the estimator of the original MC method. Indeed,

$$\mathrm{Var}(\widehat{f}_n^{MC}) = \frac{1}{n}\mathrm{Var}(f(T)) \geq \frac{1}{n} \sum_{i=1}^{m} p_i \sigma_i^2 = \mathrm{Var}(\widehat{f}_n^{SS}),$$

since for all $1 \leq i \leq m$, $\sigma_i^2 \leq \mathrm{Var}(f(T))$.

5.2.3 Fiability methods

The aim of fiability methods is to evaluate failure probabilities, namely rare event probabilities of the form $\mathbb{P}[Y > s]$ where s is a given threshold value. Examples of such methods are the *First* and *Second Order Reliability Method* (FORM and SORM). They consist in approximating the *failure domain* $G = \{t \in \mathcal{T},\ y = f(t) > s\}$ by a simpler domain C whose probability $\mathbb{P}(T \in C)$ can be analytically estimated.

Let us assume here for the sake of simplicity that T is a vector of standard independent identically distributed Gaussian random variables. Fiability methods can be easily extended to more general distribution laws by considering copula theory [153]. The first step of the method is to look for the most likely failure point $t^* \in G$, which is called the *conception point*. Due to the particular properties of the gaussian probability distribution, it holds that this point is the point of the failure domain G

that is closest to the origin. It is determined via the resolution of the optimization problem

$$t^* = \underset{t \in G}{\operatorname{argmin}} \|t\|^2.$$

A large number of optimization algorithms have been proposed in the literature to handle this problem. For simple failure domains, the convergence is quickly achieved. Unfortunately, in general, the geometry of the domain G is quite complicated. Often, when a simple reduced-order model is at hand, the latter is used to give an approximation of the failure domain in the neighborhood of the conception point.

If the conception point is assumed to be unique and the frontier of the failure domain is regular enough in the neighborhood of this point, *Breitung's formula* gives the following asymptotic result

$$\mathbb{P}(Y > s) \approx_{\beta \to \infty} E(-\beta) \Pi_{i=1}^{p-1} \frac{1}{\sqrt{1 + \beta \kappa_i}},$$

where $\beta = \|t^*\|$, E is the cumulative distribution of the one-dimensional gaussian probability distribution and κ_i are the principal curvatures of the failure domain in the neighborhood of X^*. It must be pointed out that β is an increasing function of the threshold parameter s.

The FORM and SORM methods then consist in approximating the failure domain G either by a hyperplane or a paraboloid tangent to G at the point t^*. For the FORM method,

$$\mathbb{P}(Y > s) \approx E(-\beta),$$

whereas for the SORM method,

$$\mathbb{P}(Y > s) \approx E(-\beta) \Pi_{i=1}^{N-1} \frac{1}{\sqrt{1 + \beta \kappa_i}}.$$

In the case of the FORM method, the computational cost is essentially determined by the optimization algorithm used to find the conception point t^*, since the remaining computations do not involve evaluations of the function f. However, in the SORM method, the computational cost is higher since the curvatures κ_i have to be computed and this implies calculating the hessian of the function f at the point t^*.

5.2.4 Reduced-order models in UQ

We present here the most classical non-statistical methods used in UQ to derive efficient reduced-order models for the solution $u(x, T)$ of

$$\mathcal{A}(u(x, T); x, T) = b(x, T). \tag{5.8}$$

Sensitivity analysis methods

The *perturbation* method [117] consists in constructing the expansion of the solution $u(x, T)$ in the neighborhood of the mean of the random variables $\mu = \mathbb{E}[T]$.

$$u(x, T) = u_0(x) + \sum_{i=1}^{p} (T_i - \mu_{T_i}) u_{,i}(x) + \sum_{i,j=1}^{p} \frac{1}{2} (T_i - \mu_{T_i})(T_j - \mu_{T_j}) u_{,ij}(x) + \cdots \tag{5.9}$$

where $u_0(x) = u(x, \mu)$, $u_{,i}(x) = \frac{\partial u}{\partial t_i}(x, \mu)$, $u_{,ij}(x) = \frac{\partial^2 u}{\partial t_i \partial t_j}(x, \mu)$, ...

By writing similar expansions for the operator $\mathcal{A}(\cdot; x, T)$ and the right-hand-side $b(x, T)$ and by inserting these expansions in the initial equation (5.8), one obtains that the coefficients in the expansion of u (5.9) are solutions of the following problems:

$$\begin{aligned} \mathcal{A}_0(u_0) &= b_0, \\ \mathcal{A}_0(u_{,i}) &= b_{,i} - \mathcal{A}_{,i}(u_0), \\ \mathcal{A}_0(u_{,ij}) &= b_{,ij} - \mathcal{A}_{,i}(u_{,j}) - \mathcal{A}_{,j}(u_{,i}) - \mathcal{A}_{,ij}(u_0) \cdots \end{aligned}$$

All these problems are deterministic problems with the same deterministic operator $\mathcal{A}_0 = \mathcal{A}(\cdot; x, \mu)$. Perturbation methods are essentially used in order to perform sensitivity analysis of the solution u with respect to the random parameters $T = (T_1, \cdots, T_p)$ around their mean values. Unfortunately, these methods are often limited to a small order expansion. In practice, perturbation methods are used to compute statistical moments up to the second or third order. Thus, their application is limited to the case when the random parameters which come into play do not significantly vary.

The *Neumann decomposition* method [10] is based on the following decomposition

$$\mathcal{A}(\cdot; x, T) = \mathcal{A}_0(\cdot; x) + \widetilde{\mathcal{A}}(\cdot; x, T) = \mathcal{A}_0(\mathcal{I} + \mathcal{A}_0^{-1}\widetilde{\mathcal{A}}(\cdot; T)),$$

where $\mathcal{A}_0$ is a deterministic operator, $\mathcal{A}_0^{-1}$ its inverse and $\mathcal{I}$ the identity operator. Under some assumptions, the inverse of the operator $\mathcal{A}$ can then be written as

$$\mathcal{A}^{-1}(\cdot; x, T) = \sum_{i=0}^{\infty} (-1)^i (\mathcal{A}_0^{-1}\widetilde{\mathcal{A}}(\cdot; x, T))^i \mathcal{A}_0^{-1}$$

so that the solution u of the initial problem (5.8) may be rewritten as

$$u(x, T) = \sum_{i=0}^{\infty} (-1)^i u_i(x, T)$$

where

$$\begin{aligned} \mathcal{A}_0(u_0(x, T); x) &= B(x, T), \\ \mathcal{A}_0(u_i(x, T); x) &= \widetilde{\mathcal{A}}(u_{i-1}(x, T); x, T). \end{aligned}$$

All the functions u_i are solutions to deterministic equations with a stochastic right-hand side with a unique deterministic operator $\mathcal{A}_0$. However, these computations are very expensive, since they are prone to the curse of dimensionality (they lead to high-dimensional parametrized problems). Similarly to perturbation methods, they can only be used to compute low-order statistical moments of the solution u.

Both the perturbation method and the Neumann decomposition method are usually used to do sensitivity analysis of the function u on different random variables and are used to compute statistical moments of low order in these variables.

Spectral stochastic methods

The principle of *spectral stochastic* methods [92] is to decompose the function u as

$$u(x,t) = \sum_{i\in\mathbb{N}} u_i(x)\Phi_i(t),$$

where $(\Phi_i(t))_{i\in\mathbb{N}}$ is a family of functions fixed *a priori*, and to compute the unknown coefficients $u_i(x)$.

The *Galerkin* method is one of the most widely used spectral stochastic method. For $M \in \mathbb{N}^p$, the family $(\Phi_i(T(\omega)))_{1\leq i\leq M}$ is often chosen to be an orthonormal family of functions of $L^2(\Omega,\mathbb{P})$. We consider the weak formulation of the initial problem:

$$\mathbb{E}\left[\mathcal{A}\left(\sum_{k=0}^{M} u_k(x)\Phi_k(T); x, X\right)\Psi(T)\right] = \mathbb{E}\left[b(x,T)\Psi(T)\right],$$

for $\Psi(T) \in L^2(\Omega,\mathbb{P})$, and we evaluate this equation for the test functions $\Psi = \Phi_i$ for $1 \leq i \leq M$, which leads to a set of M coupled deterministic equations to determine the coefficients $u_i(x)$.

We should remark that this method is *intrusive* in the sense defined in Section 5.2.1.

However, there exists other spectral stochastic methods that are *non-intrusive*. One of them is the so-called *stochastic collocation* method. This technique relies on the choice of a set of *collocation points* $\{t_k\}_{1\leq k\leq M}$, and an associated interpolation basis $(\Phi_k)_{1\leq k\leq M}$ for this set of points, for example Lagrange polynomials [11]. In the expansion

$$u(x,T) = \sum_{k=1}^{M} u_k(x)\Phi_k(T),$$

the coefficients $u_k(x)$ are then evaluated as the solutions of the initial deterministic problem

$$\mathcal{A}(u_k(x); x, t_k) = b(x, t_k),$$

with $t_k \in \mathcal{T}$ being the collocation point. Several techniques exist in the literature in order to properly choose the set of collocation points, such as Smolyak grid [180] for instance. We refer to Section 5.5.2 for more details on sparse grid methods.

Let us mention a last non-intrusive spectral stochastic method, the *non-intrusive projection* method. The function u is developed as

$$u(x,T) = \sum_{k=1}^{M} u_k(x)\Phi_k(T),$$

where $(\Phi_k(T))_{k=1}^{M}$ is an orthogonal family of $L^2(\Omega,\mathbb{P})$. The coefficients

$$u_k(x) = \frac{\mathbb{E}[u(x,T)\Phi_k(T)]}{\mathbb{E}[\Phi_k(T)^2]},$$

are then evaluated using a direct numerical integration technique, such as the Monte-Carlo method. However, to achieve accurate representations of the function u, a high

number of integration points are needed, like in the case of collocation methods, and thus computations may be very expensive, unless special sparse adaptive integration grids are used. Of course, the accuracy of such a representation will also highly depends on the regularity of the function u.

The table below summarizes different properties of the UQ methods introduced above. Unfortunately, they are all prone to the curse of dimensionality, except Monte-Carlo methods, for which efficient reduced-order models are crucial to improve the rate of convergence of the algorithm. It is to be noted though that the goals of these methods are different. Monte-Carlo and reliability methods aim at computing esperances and probabilities of random scalar output quantities, whereas the other methods aim at computing a complete approximation of the function $(x, T) \mapsto u(x, T)$.

	Determinitic/Stochastic	Intrusive/Non-intrusive	Main application	Rigorous error estimates
Monte-Carlo	Stochastic	Non-intrusive	Statistical moments	Yes
Reliability Methods	Stochastic	Intrusive	Rare events	No
Perturbation	Deterministic	Intrusive	ROM	No
Neumann	Deterministic	Intrusive	ROM	No
L2 projection	Deterministic	Non-intrusive	ROM	No
Collocation	Deterministic	Non-intrusive	ROM	No
Galerkin	Deterministic	Intrusive	ROM	Yes

	Need of a ROM	Regularity of the function	Prone to the curse of dimensionality
Monte-Carlo	Yes	No	No
Reliability Methods	Yes	Yes	Yes
Perturbation	No	Yes	Yes
Neumann	No	Yes	Yes
L2 projection	No	Yes	Yes
Collocation	No	Yes	Yes
Galerkin	No	No	Yes

Figure 5.2: Classical methods in UQ

5.3 Greedy algorithms and tensor product representations for high-dimensional problems

In this section, we present the approach we adopt in this thesis to deal with the uncertainty quantification problem presented in the introduction. The method is called Proper Generalized Decomposition (PGD) and is very closely linked to the so-called greedy algorithms, introduced in nonlinear approximation theory. We first present greedy algorithms in a very general context, then show how they can be used in practice for the treatment of high-dimensional problems.

5.3.1 Greedy algorithms

We present here a short and non-exhaustive review of some greedy algorithms in a general framework. We refer to [12, 75, 174] for more details.

Let V be a real Hilbert space endowed with the inner product $\langle \cdot, \cdot \rangle_V$ and associated norm $\| \cdot \|_V$. We say that a set $\mathcal{D}$ of functions (elements) from V is a *dictionary* if each element $g \in \mathcal{D}$ is such that $\|g\|_V = 1$ and $\overline{\text{Span}(\mathcal{D})} = V$. For the sake of simplicity, we will assume here that the dictionary is symmetric, i.e. that $\{-g,\ g \in \mathcal{D}\} \subset \mathcal{D}$.

A generic mathematical problem is the *best n-term approximation* problem, i.e. to find the best approximation of a function $u \in V$ by any linear combination of at most n elements of the dictionary. In other words, we would like to find n elements $g_1, \cdots, g_n \in \mathcal{D}$ such that

$$(g_1, \cdots, g_n) \in \underset{(d_1,\cdots,d_n)\in\mathcal{D}}{\text{argmin}} \|u - P_{d_1,\cdots,d_n} u\|_V,$$

where $P_{d_1,\cdots,d_n}$ is the orthogonal projector on $\text{Span}\{d_1, \cdots, d_n\}$ with respect to the scalar product of V.

Finding the best n-term approximation is a very intricate task and cannot be achieved in an explicit way. The problem is actually to find a *good* approximation of any function $u \in V$ as a linear combination of n elements of the dictionary.

In linear approximation, the set of n elements $g_1, \cdots, g_n \in \mathcal{D}$ are often fixed a priori and any function $u \in V$ is approximated by its projection $P_{g_1,\cdots,g_n} u$ on the linear space $\text{Span}(g_1, \cdots, g_n)$.

On the opposite, the key idea behind nonlinear approximation is that the elements used in the approximation are not fixed a priori but depend on the function to be approximated. Greedy algorithms provide a constructive way to find an approximation of a function $u \in V$ as a linear combination of carefully chosen elements of the dictionary. The principle of a greedy algorithm is to look iteratively for the best element in the dictionary.

Here, we make the assumption that for any $u \in V$, there exists an element (not necessarily unique) $g \in \mathcal{D}$ such that

$$g \in \text{argmax}_{d\in\mathcal{D}} \langle u, d \rangle_V. \tag{5.10}$$

It is easy to check that if g is a solution to (5.10), then

$$(g, \langle u, g \rangle_V) \in \underset{(d,\lambda)\in\mathcal{D}\times\mathbb{R}}{\text{argmin}} \|u - \lambda d\|_V.$$

We present here the most classical versions of the greedy algorithms, which were introduced in [75]. Let us begin with the *Pure Greedy Algorithm* (PGA) and the *Orthogonal Greedy Algorithm* (OGA).

PGA:

1. Set $r_0^p := u$, $u_0^p := 0$ and $n = 1$. Choose $\varepsilon > 0$.

2. Find $g_n^p \in \mathcal{D}$ such that

$$g_n^p \in \text{argmax}_{g\in\mathcal{D}}\langle r_{n-1}^p, g\rangle_V.$$

3. Define $u_n^p := u_{n-1}^p + \langle r_{n-1}^p, g_n^p\rangle_V g_n^p$ and $r_n^p := r_{n-1}^p - \langle r_{n-1}^p, g_n^p\rangle_V g_n^p$.

4. If $\|r_n^p\| =\leq \varepsilon\|u_n^p\|$, then stop. Otherwise, $n = n + 1$ and return to step 2.

OGA:

1. Set $r_0^o := u$, $u_0^o := 0$ and $n = 1$. Choose $\varepsilon > 0$.

2. Find $g_n^o \in \mathcal{D}$ such that

$$g_n^o \in \text{argmax}_{g\in\mathcal{D}}\langle r_{n-1}^o, g\rangle.$$

3. Define $H_n^o := \text{Span}\{g_i^o,\ 1 \leq i \leq n\}$, $u_n^o := P_{H_n^o}(u)$ and $r_n^o := u - P_{H_n^o}(u)$.

4. If $\|r_n^o\| \leq \varepsilon\|u_n^o\|$, then stop. Otherwise, $n = n + 1$ and return to step 2.

These algorithms are proved to converge whatever the element $u \in V$.

Theorem 5.3.1. *For any dictionary $\mathcal{D}$ and any $u \in V$, it holds that (taking $\varepsilon = 0$ in the above algorithms)*

PGA $\|r_n^p\| = \|u - u_n^p\| \underset{n\to\infty}{\longrightarrow} 0$;

OGA $\|r_n^o\| = \|u - u_n^o\| \underset{n\to\infty}{\longrightarrow} 0$.

To derive convergence rates for these two algorithms, one has to assume additional properties on the function $u \in V$. For a general dictionary $\mathcal{D}$, we define the class of functions

$$\mathcal{A}_1^0(\mathcal{D}, M) := \left\{ u \in V,\ u = \sum_{k\in\Lambda} c_k v_k,\ v_k \in \mathcal{D},\ \#\Lambda < \infty, \sum_{k\in\Lambda} |c_k| \leq M \right\},$$

and we define $\mathcal{A}_1(\mathcal{D}, M)$ to be the closure in V of $\mathcal{A}_1^0(\mathcal{D}, M)$. Furthermore,

$$\mathcal{A}_1(\mathcal{D}) := \bigcup_{M>0} \mathcal{A}_1(\mathcal{D}, M).$$

Then, the following theorem holds:

Theorem 5.3.2. *Let $\mathcal{D}$ be an arbitrary dictionary in V. Then, for each $u \in \mathcal{A}_1(\mathcal{D}, M)$, we have for all $n \in \mathbb{N}^*$,*

$$\|r_n^p\| = \|u - u_n^p\| \leq Mn^{-11/62},$$
$$\|r_n^o\| = \|u - u_n^o\| \leq Mn^{-1/2}.$$

Actually, the first convergence rate which was proved for the Pure Greedy Algorithm was

$$\|r_n^p\| = \|u - u_n^p\| \leq Mn^{-1/6},$$

and the factor $\frac{1}{6}$ was later improved to $\frac{11}{62}$.

5.3.2 Greedy algorithms for high-dimensional problems

Let us now present the greedy algorithm as it is used in the Proper Generalized Decomposition for the treatment of high-dimensional PDEs. This method was introduced by Ladevèze [123] in the context of time-space decomposition, by Chinesta [5] for the computation of high-dimensional Fokker-Planck equations for polymers, and by Nouy [155] in the context of UQ.

Let $a : \mathcal{X} \times \mathcal{T} \to \mathbb{R}$ be a measurable function such that there exists $\alpha, \beta > 0$ such that

$$\forall (x,t) \in \mathcal{X} \times \mathcal{T},\ \alpha \leq a(x,t) \leq \beta. \tag{5.11}$$

We will focus on two prototypical problems of interest:

$$\left\{ \begin{array}{l} \text{find } u \in L^2(\mathcal{T}, H_0^1(\mathcal{X})) \text{ such that} \\ -\text{div}_x(a \nabla_x u) = f \text{ in } \mathcal{D}'(\mathcal{X} \times \mathcal{T}), \end{array} \right. \tag{5.12}$$

and

$$\left\{ \begin{array}{l} \text{find } u \in H_0^1(\mathcal{X} \times \mathcal{T}) \text{ such that} \\ -\text{div}_{x,t}(a \nabla_{x,t} u) = f \text{ in } \mathcal{D}'(\mathcal{X} \times \mathcal{T}), \end{array} \right. \tag{5.13}$$

with $f \in L^2(\mathcal{T} \times \Xi)$. Problem (5.12) can be seen as a diffusion problem with uncertainty, t denoting the set of random parameters. Problem (5.13) is a multivariate Poisson problem.

More generally, let V be a Hilbert space of multivariate functions $u(x, t_1, \cdots, t_p)$ and let $V_x, V_1, \cdots, V_p$ be Hilbert spaces of single-variate functions which depend respectively on $x, t_1, \cdots, t_p$. The key principle of the PGD relies on the use of the following particular choice of dictionary:

$$\mathcal{D} := \left\{ r \otimes s^{(1)} \otimes \cdots \otimes s^{(p)} \mid r \in V_x,\ s^{(1)} \in V_1,\ \cdots,\ s^{(p)} \in V_p,\ \|r \otimes s^{(1)} \otimes \cdots \otimes s^{(p)}\|_V = 1 \right\},$$

where the tensor product notation means

$$r \otimes s^{(1)} \otimes \cdots \otimes s^{(p)} : \left\{ \begin{array}{ccc} \mathcal{X} \times \mathcal{T}_1 \times \cdots \times \mathcal{T}_p & \longrightarrow & \mathbb{R} \\ (x, t_1, \cdots, t_p) & \mapsto & r(x) s^{(1)}(t_1) \cdots s^{(p)}(t_p). \end{array} \right.$$

Let us denote by

$$\Sigma := \{ r \otimes s^{(1)} \otimes \cdots \otimes s^{(p)},\ r \in V_x,\ s^{(1)} \in V_1,\ \cdots, s^{(p)} \in V_p \}.$$

Under the assumptions

(A1) $\Sigma \subset V$;

(A2) $\overline{\text{Span}\Sigma}^{\|\cdot\|_V} = V$;

(A3) Σ is weakly closed in V,

$\mathcal{D}$ is a well-defined dictionary of V. Actually, the third condition is a sufficient condition to ensure that maximization problems of the form (5.10) have at least one solution.

Performing a greedy algorithm with this particular dictionary provides a separated representation of a given function $u \in V$ of the form

$$\begin{aligned} u(x, t_1, \cdots, t_p) &\approx \sum_{k=1}^{n} r_k(x) s_k^{(1)}(t_1) \cdots s_k^{(p)}(t_p) \\ &= \sum_{k=1}^{n} r_k \otimes s_k^{(1)} \otimes \cdots \otimes s_k^{(p)}(x, t_1, \cdots, t_p). \end{aligned}$$

Both problems (5.12) and (5.13) can be written as minimization problems

$$u = \operatorname*{argmin}_{v \in V} \mathcal{E}(v), \tag{5.14}$$

where $\mathcal{E}(v) := \|v - u\|_V^2$. For problem (5.12), $V = L^2(\mathcal{T}, H_0^1(\mathcal{X}))$ and

$$\forall v \in V,\ \|v\|_V^2 = \int_{\mathcal{X} \times \mathcal{T}} a(x,t) |\nabla_x v(x,t)|^2 \, dx \, dt.$$

For the multivariate Poisson problem (5.13), $V = H_0^1(\mathcal{X} \times \mathcal{T})$ and

$$\forall v \in V,\ \|v\|_V^2 = \int_{\mathcal{X} \times \Xi} a(x,t) |\nabla_{x,t} v(x,t)|^2 \, dx \, dt.$$

The PGD algorithm for the resolution of a high-dimensional problem of the general form (5.14) reads: for all $n \in \mathbb{N}^*$, compute iteratively $(r_n, s_n^{(1)}, \cdots s_n^{(p)}) \in V_x \times V_1 \times \cdots \times V_p$ such that

$$(r_n, s_n^{(1)}, \cdots s_n^{(p)}) \in \operatorname*{argmin}_{(r, s^{(1)}, \cdots, s^{(p)}) \in V_x \times V_1 \times \cdots \times V_p} \mathcal{E}\left(\sum_{k=1}^{n-1} r_k \otimes s_k^{(1)} \otimes \cdots \otimes s_k^{(p)} + r \otimes s^{(1)} \otimes \cdots \otimes s^{(p)} \right). \tag{5.15}$$

In other words, this algorithm exactly amounts to performing a Pure Greedy Algorithm in the Hilbert space V for the approximation of u with the particular dictionary $\mathcal{D}$.

In the case of two Hilbert spaces V_x and V_t, the algorithm reads: for all $n \in \mathbb{N}^*$, compute iteratively $(r_n, s_n) \in V_x \times V_t$ such that

$$(r_n, s_n) \in \operatorname*{argmin}_{(r,s) \in V_x \times V_t} \mathcal{E}\left(\sum_{k=1}^{n-1} r_k \otimes s_k + r \otimes s \right). \tag{5.16}$$

For each $n \in \mathbb{N}$, we use the notation $u_n := \sum_{k=1}^n r_k \otimes s_k^{(1)} \otimes \cdots \otimes s_k^{(p)}$, or $u_n := \sum_{k=1}^n r_k \otimes s_k$ in the case of two Hilbert spaces.

In practice, at each iteration $n \in \mathbb{N}^*$, the computation of $(r_n, s_n^{(1)}, \cdots s_n^{(p)})$ is done by solving the Euler equations associated to the minimization problem (5.15)

through a fixed-point procedure. For instance, the Euler-Lagrange equations associated with problem (5.16) for problem (5.12) read:

$$\begin{cases} -\left(\int_{\mathcal{T}} |s_n|^2\right) \mathrm{div}_x(a\nabla_x r_n) = \int_{\mathcal{T}} \left(f + \mathrm{div}_x(a\nabla_x u_{n-1})\right) s_n, \\ \left(\int_{\mathcal{X}} a|\nabla_x r_n|^2\right) s_n = \int_{\mathcal{X}} \left(f + \mathrm{div}_x(a\nabla_x u_{n-1})\right) r_n, \end{cases}$$

and those for problem (5.13) read

$$\begin{cases} -\mathrm{div}_x\left[\left(\int_{\mathcal{T}} a|s_n|^2\right)\nabla_x r_n\right] + \left(\int_{\mathcal{T}} a|\nabla_t s_n|^2\right) r_n = \int_{\mathcal{T}} \left(f + \mathrm{div}_{x,t}(a\nabla_{x,t} u_{n-1})\right) s_n, \\ \left(\int_{\mathcal{X}} a|\nabla_x r_n|^2\right) s_n - \mathrm{div}_t\left[\left(\int_{\mathcal{X}} a|r_n|^2\right)\nabla_t s_n\right] = \int_{\mathcal{X}} \left(f + \mathrm{div}_{x,t}(a\nabla_{x,t} u_{n-1})\right) r_n. \end{cases} \tag{5.17}$$

The fixed-point algorithm to solve the Euler equations (5.17) for instance consists in calculating iteratively pairs $(r_n^m, s_n^m) \in V_x \times V_t$ for each $m \in \mathbb{N}$ by the following procedure

$$\begin{cases} -\mathrm{div}_x\left[\left(\int_{\mathcal{T}} a|s_n^{m-1}|^2\right)\nabla_x r_n^m\right] + \left(\int_{\mathcal{T}} a|\nabla_t s_n^{m-1}|^2\right) r_n^m = \int_{\mathcal{T}} \left(f + \mathrm{div}_{x,t}(a\nabla_{x,t} u_{n-1})\right) s_n^{m-1}, \\ \left(\int_{\mathcal{X}} a|\nabla_x r_n^m|^2\right) s_n^m - \mathrm{div}_t\left[\left(\int_{\mathcal{X}} a|r_n^m|^2\right)\nabla_t s_n^m\right] = \int_{\mathcal{X}} \left(f + \mathrm{div}_{x,t}(a\nabla_{x,t} u_{n-1})\right) r_n^m. \end{cases}$$

The initial guess (r_n^0, s_n^0) is usually chosen randomly.

An orthogonal version of this greedy algorithm can be defined as follows:

1. set $u_0^o = 0$ and $n = 1$;
2. compute $(r_n, s_n^{(1)}, \cdots s_n^{(p)}) \in V_x \times V_1 \times \cdots \times V_p$ such that
$$(r_n, s_n^{(1)}, \cdots s_n^{(p)}) \in \underset{(r,s^{(1)},\cdots,s^{(p)}) \in V_x \times V_1 \times \cdots \times V_p}{\mathrm{argmin}} \mathcal{E}\left(u_{n-1}^o + r \otimes s^{(1)} \otimes \cdots \otimes s^{(p)}\right). \tag{5.18}$$
3. compute $(\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n$ such that
$$(\alpha_1, \cdots, \alpha_n) \in \underset{(\beta_1,\cdots,\beta_n) \in \mathbb{R}^n}{\mathrm{argmin}} \mathcal{E}\left(\sum_{k=1}^n \beta_k r_k \otimes s_k^{(1)} \otimes \cdots s_k^{(p)}\right);$$
4. set $u_n^o = \sum_{k=1}^n \alpha_k r_k \otimes s_k^{(1)} \otimes \cdots s_k^{(p)}$ and $n = n + 1$.

5.3.3 Convergence results

Why does such a representation enable one to avoid the curse of dimensionality? Using the same notation as in the introduction, when the Hilbert spaces are discretized, i.e. when $V_x = \mathrm{Span}\{\phi_i,\ 1 \le i \le N_x\}$ and $V_l = \mathrm{Span}\{\psi_j^{(l)},\ 1 \le j \le N_l\}$ for $1 \le l \le p$, computing an approximation of u under the form (5.14) amounts to

solving n problems of size $N_x + N_1 + \cdots + N_p$. In the case when $N_1 = \cdots = N_p = N$, then each iteration of the algorithms amounts to solving n problems of size $N_x + pN$. Here the dimension of the discretized problems solved at each iteration scales linearly with p and the greedy algorithm can then be used for computing the solution u of problems which cannot be tackled by standard Galerkin methods. However, the price to pay is that computing the pair $(r_n, s_n) \in V_x \times V_t$ at each iteration n amounts to solving a low-dimensional nonlinear problem (even if the original high-dimensional problem is linear).

The greedy algorithm (5.15) was proved to converge for problem (5.13) in [130] with $V_x = H_0^1(\mathcal{X})$, $V_j = H_0^1(\mathcal{T}_j)$ for all $1 \leq j \leq p$ and $a = 1$, and the analysis done in this paper immediately extends to problem (5.12) with $V_x = H_0^1(\mathcal{X})$ and $V_j = L^2(\mathcal{T}_j)$ for $1 \leq j \leq p$. Similar results were proved in an even more general context [86], which encompasses the cases of (5.12) and (5.13) for an arbitrary diffusion coefficient a satisfying the uniform ellipticity assumption (5.11).

Actually, in these cases, the set Σ satisfies assumptions (A1)-(A3) and the convergence results of the Pure Greedy Algorithm and Orthogonal Greedy algorithm introduced in Section 5.3.1 hold.

Thus, for all $u \in V$,

$$\|u_n - u\|_V \underset{n\to\infty}{\longrightarrow} 0 \tag{5.19}$$

and

$$\|u_n^o - u\|_V \underset{n\to\infty}{\longrightarrow} 0. \tag{5.20}$$

Besides, for all $u \in \mathcal{A}_1(\mathcal{D})$, where

$$\mathcal{A}_1(\mathcal{D}) := \left\{ u \in V,\ u = \sum_{k=1}^{+\infty} r_k \otimes s_n^{(1)} \otimes \cdots \otimes s_k^{(p)},\ \sum_{k=1}^{+\infty} \|r_k \otimes s_n^{(1)} \otimes \cdots \otimes s_k^{(p)}\|_V < +\infty \right\},$$

there exists a constant $C > 0$ such that

$$\|u - u_n\|_V \leq C n^{-11/62}, \tag{5.21}$$

and

$$\|u - u_n^o\|_V \leq C n^{-1/2}. \tag{5.22}$$

In our particular setting, the set $\mathcal{A}_1(\mathcal{D})$ is called the projective space, and for all $u \in \mathcal{A}_1(\mathcal{D})$,

$$\|u\|_p := \inf \left\{ \sum_{k=1}^{+\infty} \|r_k \otimes s_n^{(1)} \otimes \cdots \otimes s_k^{(p)}\|_V,\ \mid u = \sum_{k=1}^{+\infty} r_k \otimes s_n^{(1)} \otimes \cdots \otimes s_k^{(p)} \right\}$$

is called the projective norm of u.

5.3.4 Characterization of the set $\mathcal{A}_1(\mathcal{D})$

A natural question is the following: is it possible to give a simple characterization of the set $\mathcal{A}_1(\mathcal{D})$ when the Hilbert spaces V, V_x, V_1, ..., V_p are standard Sobolev

spaces? Simple characterizations of the full set $\mathcal{A}_1(\mathcal{D})$ are not known so far. However, it is possible to identify subsets of $\mathcal{A}_1(\mathcal{D})$ which can be characterized either in terms of standard Sobolev spaces or in terms of mixed Sobolev spaces. These characterizations are given in [86] for arbitrary diffusion coefficients a, but, for the sake of simplicity, we detail the results of problem (5.13) with $a = 1$.

Then, the following theorem holds:

Theorem 5.3.3. *Let us assume that, $\mathcal{X}$, $\mathcal{T}_1$, ..., $\mathcal{T}_p \subset \mathbb{R}^d$. On the one hand, provided that $m > 1 + (p+1)d/2$,*

$$H^m(\mathcal{X} \times \mathcal{T}) \cap H_0^1(\mathcal{X} \times \mathcal{T}) \subset \mathcal{A}_1(\mathcal{D}). \tag{5.23}$$

On the other hand, if $d = 1$,

$$H^{2,mix}(\mathcal{X} \times \mathcal{T}) \cap H_0^1(\mathcal{X} \times \mathcal{T}) \subset \mathcal{A}_1(\mathcal{D}), \tag{5.24}$$

where

$$H^{2,mix}(\mathcal{X}\times\mathcal{T}) := \left\{ \begin{array}{c} \phi \in L^2(\mathcal{X} \times \mathcal{T}),\ \partial_\alpha \phi \in L^2(\mathcal{X} \times \mathcal{T}), \\ \forall \alpha = (\alpha_x, \alpha_1, \cdots, \alpha_p) \in \mathbb{N}^{p+1},\ |\alpha|_\infty = \max(\alpha_x, \max_{1\leq j\leq p} \alpha_j) \leq 2 \end{array} \right\}.$$

Besides, if $d = 2, 3$,

$$H^{4,mix}(\mathcal{X} \times \mathcal{T}) \cap H_0^1(\mathcal{X} \times \mathcal{T}) \subset \mathcal{A}_1(\mathcal{D}), \tag{5.25}$$

where

$$H^{4,mix}(\mathcal{X}\times\mathcal{T}) := \left\{ \begin{array}{c} \phi \in L^2(\mathcal{X} \times \mathcal{T}),\ \partial_\alpha \phi \in L^2(\mathcal{X} \times \mathcal{T}), \\ \forall \alpha = (\alpha_x, \alpha_1, \cdots, \alpha_p) \in \mathbb{N}^{p+1},\ |\alpha|_\infty = \max(\alpha_x, \max_{1\leq j\leq p} \alpha_j) \leq 4 \end{array} \right\}.$$

Characterization (5.23) had already been proved in [130] for the case $d = 1$. All these characterizations are useful in themselves since there is no simple link between standard Sobolev spaces and anisotropic Sobolev spaces $H^{k,mix}(\mathcal{X} \times \mathcal{T})$. One of the nice features of (5.24) and (5.25) is that these results do not depend on p.

Let us introduce the definition of tensor product of spaces. Let $\langle \cdot, \cdot \rangle_\otimes$ be the scalar product defined on $\mathrm{Span}(\Sigma)$ by

$$\begin{array}{c} \forall (r, s^{(1)}, \cdots, s^{(p)}), (\widetilde{r}, \widetilde{s}^{(1)}, \cdots, \widetilde{s}^{(p)}) \in V_x \times V_1 \times \cdots \times V_p, \\ \langle r \otimes s^{(1)}, \cdots \otimes s^{(p)}, \widetilde{r} \otimes \widetilde{s}^{(1)} \otimes \cdots \otimes \widetilde{s}^{(p)} \rangle_\otimes = \langle r, \widetilde{r} \rangle_{V_x} \langle s^{(1)}, \widetilde{s}^{(1)} \rangle_{V_1} \cdots \langle s^{(p)}, \widetilde{s}^{(p)} \rangle_{V_p}, \end{array}$$

and let $\| \cdot \|_\otimes$ be the associated norm. In particular,

$$\forall (r, s^{(1)}, \cdots, s^{(p)}) \in V_x \times V_1 \times \cdots \times V_p,\ \| r \otimes s^{(1)} \otimes \cdots \otimes s^{(p)} \|_\otimes = \| r \|_{V_x} \| s^{(1)} \|_{V_1} \cdots \| s^{(p)} \|_{V_p},$$

and $\| \cdot \|_\otimes$ is called a *cross-norm*.

Then, $\left(\overline{\mathrm{Span}(\Sigma)}^{\| \cdot \|_\otimes}, \langle \cdot, \cdot \rangle_\otimes \right)$ defines a Hilbert space which is denoted by $V_x \otimes V_1 \otimes \cdots \otimes V_p$ and is called the tensor product of the spaces V_x, V_1, ..., V_p. Note that if V is a Hilbert space such that conditions (A1)-(A3) are fulfilled, then $V_x \otimes V_1 \otimes \cdots \otimes V_p \subset V$ but the inclusion may be strict.

In the case of problem (5.12),

$$V = L^2(\mathcal{T}, H_0^1(\mathcal{X})) = H_0^1(\mathcal{X}) \otimes L^2(\mathcal{T}_1) \otimes \cdots \otimes L^2(\mathcal{T}_p) = V_x \otimes V_1 \otimes \cdots \otimes V_p.$$

However, in the case of problem (5.13),

$$V_x \otimes V_1 \otimes V_p = H_0^1(\mathcal{X}) \otimes H_0^1(\mathcal{T}_1) \otimes \cdots \otimes H_0^1(\mathcal{T}_p) \subsetneq V = H_0^1(\mathcal{X} \times \mathcal{T}_1 \times \cdots \times \mathcal{T}_p).$$

We refer to Chapter 6 for a proof of this latter fact.

5.3.5 The Singular Value Decomposition case and the general linear case

There are two different cases which should be highlighted regarding properties of the greedy algorithms introduced above.

Case 1: **The Singular Value Decomposition case**

Let us consider the case when $\mathcal{E}(v) = \|v - u\|_V^2$ where $V = V_x \otimes V_t$ (tensor product of only two Hilbert spaces) and the norm $\|\cdot\|_V$ is a cross-norm in the sense that for all $(r, s) \in V_x \times V_t$, $\|r \otimes s\|_V = \|r\|_{V_x}\|s\|_{V_t}$.

This is typically the situation of (5.12) with $p = 1$. Then, the pairs $(r_n, s_n) \in V_x \times V_t$ computed at each iteration of the greedy algorithm (5.15) satisfy the following orthogonality property [130], which is an immediate consequence of the Euler equations associated to the minimization problem (5.16):

$$\forall n \neq n',\ \langle s_n, s_{n'}\rangle_{V_t} = \langle r_n, r_{n'}\rangle_{V_x} = 0. \tag{5.26}$$

This orthogonality property implies that the approximation $u_n = \sum_{k=1}^n r_k \otimes s_k$ of u given by the greedy algorithm at iteration n is a best rank-n approximation of u in the sense that

$$\left\| u - \sum_{k=1}^n r_k \otimes s_k \right\|_V = \inf_{(\widetilde{r}_k, \widetilde{s}_k) \in V_x \times V_t,\ 1 \leq k \leq n} \left\| u - \sum_{k=1}^n \widetilde{r}_k \otimes \widetilde{s}_k \right\|_V. \tag{5.27}$$

Besides, the Pure Greedy and Orthogonal Greedy algorithms are equivalent to one another and the decomposition of u under the form

$$u = \sum_{k=1}^{\infty} r_k \otimes s_k$$

with orthogonality properties (5.26) is unique. Note that in the case when V_x and V_t are finite-dimensional subspaces of $L^2(\mathcal{X})$ and $L^2(\mathcal{T})$, the greedy algorithm produces the Singular Value Decomposition of the matrix u.

Let $\lambda_k = \|r_k \otimes s_k\|_V$ for all $1 \leq k \leq \infty$. The sequence $(\lambda_k)_{k \in \mathbb{N}^*}$ is non-increasing, and the decay rate of the algorithm is directly related to the decay rate of the sequence $(\lambda_k)_{k \in \mathbb{N}^*}$ in the sense that

$$\|u - u_n\|_V^2 = \sum_{k=n+1}^{+\infty} \|r_k \otimes s_k\|_V^2 = \sum_{k=n+1}^{\infty} \lambda_k^2.$$

Case 2: **The linear case**

Now let us consider the case when $\mathcal{E}(v)$ is still a quadratic functional of the form $\mathcal{E}(v) = \|v - u\|_V^2$ but when the number of Hilbert spaces is greater than 3 or when the norm $\|\cdot\|_V$ is not a cross-norm. Even if the convergence properties of greedy algorithms (5.19), (5.20), (5.21) and (5.22) still hold, the orthogonality property (5.26) is no more true in general. This implies that the rank-n decomposition given by the greedy algorithm is not optimal in general

in the sense that (5.27) does not hold anymore. The Pure and Orthogonal greedy algorithms are not equivalent to one another and the decay rate of $\|u - u_n\|_V^2$ cannot be easily linked to the sequence of the norms of each tensor product functions appearing in the expansion given by the algorithms.

Let us emphasize that the Poisson problem (5.13) is a prototypical example of situation where the norm $\|\cdot\|_V$ is not a cross-norm. Actually, it holds that $V_x \otimes V_t = H_0^1(\mathcal{X}) \otimes H_0^1(\mathcal{T}) \subsetneq V = H_0^1(\mathcal{X} \times \mathcal{T})$.

When we began this thesis work, the only convergence results which were rigorously proved for the greedy algorithm (5.15) concerned cases when

$$\forall v \in V,\ \mathcal{E}(v) = \|v - u\|_V^2, \tag{5.28}$$

for some $u \in V$ and some norm $\|\cdot\|_V$. A prototypical case of problem which can be rewritten as a minimization problem of an energy functional of the form (5.28) is

$$\begin{cases} \text{find } u \in V \text{ such that} \\ \forall v \in V,\ a(u, v) = b(v), \end{cases} \tag{5.29}$$

where $a : V \times V \to \mathbb{R}$ is a symmetric continuous coercive bilinear form on $V \times V$ and $b : V \to \mathbb{R}$ a continuous linear form on V. In this case, the energy functional $\mathcal{E}$ can be defined as $\mathcal{E}(v) = \|v - u\|_V^2$ where

$$\forall v \in V,\ \|v\|_V^2 = \frac{1}{2} a(v, v).$$

The main contributions of this thesis work are the following. In a joint work with Eric Cancès and Tony Lelièvre, we have extended these convergence results to the case when $\mathcal{E}$ is a general nonlinear strongly convex energy functional. These results are detailed in Section 5.4.

5.4 Contributions of this thesis work

5.4.1 Nonlinear convex problems

A natural mathematical question is the following: does the algorithm (5.15) presented in the preceding section still converge when the energy functional $\mathcal{E}(v)$ is no more assumed to be a quadratic energy functional? The answer to this question is positive and it was proved by Eric Cancès, Tony Lelièvre and myself in [39]. The results and proofs gathered in this article are presented in Chapter 6.

These are an extension of the results proved in [130] to the case when $\mathcal{E}$ is a general strongly convex (possibly non quadratic) energy functional.

Let us introduce the following assumptions:

(A4) the energy functional $\mathcal{E}$ is differentiable and strongly convex, i.e. there exists $\alpha > 0$ such that

$$\forall v, w \in V,\ \mathcal{E}(v) \geq \mathcal{E}(w) + \langle \nabla \mathcal{E}(w), v - w \rangle_V + \frac{\alpha}{2} \|v - w\|_V^2;$$

(A5) the gradient of $\mathcal{E}$ is Lipschitz on bounded sets: for each bounded subset $K \subset V$, there exists a constant $L_K > 0$ such that

$$\forall v, w \in K,\ \|\nabla\mathcal{E}(v) - \nabla\mathcal{E}(w)\|_V \leq L_K \|v - w\|_V.$$

Theorem 5.4.1. *Assume that conditions (A1)-(A5) are satisfied, then the iterations of the greedy algorithm (5.15) are well-defined, in the sense that there exists a minimizer $(r_n, s_n^{(1)}, \cdots, s_n^{(p)}) \in V_x \times V_1 \times \cdots \times V_p$ to (5.15). Moreover, it satisfies $r_n \otimes s_n^{(1)} \otimes s_n^{(p)} \neq 0$ if and only if $u_{n-1} \neq u$. Besides, the sequence $(u_n)_{n\in\mathbb{N}^*}$ strongly converges in V towards the solution u of (5.14).*

We were not able to derive convergence rates for this algorithm in the general case. However, we prove in [39] that the algorithm converges exponentially fast in the finite-dimensional case.

Theorem 5.4.2. *When the spaces $V_x, V_1, \cdots, V_p$ are finite dimensional, the convergence rate of the algorithm is exponential, i.e. there exists $C, \sigma > 0$ such that for all $n \in \mathbb{N}^*$,*

$$\|u_n - u\|_V \leq Ce^{-\sigma n}.$$

It should be noted that in the above theorem, the constant C can be estimated by $\|u\|_V$ but the constant σ heavily depends on the dimensions of the spaces $V_x, V_1, \cdots, V_p$ and the number of variables. Indeed, if we assume that $\dim V_x = \dim V_1 = \cdots = \dim V_p = N$, then the constant $1 - \sigma$ scales like $N^{-(p+1)}$ as N goes to infinity.

In practice, when the Euler equations associated with the minimization problem (5.15) are solved, one can never be certain that the solution $(r_n, s_n^{(1)}, \cdots, s_n^{(p)})$ is the global minimum of (5.15). We prove in [39] that, in the case of two variables, and under the additional assumption

(A6) there exists $\beta, \gamma > 0$ such that for all $(r, s) \in V_x \times V_t$,

$$\beta\|r\|_{V_x}\|s\|_{V_t} \leq \|r \otimes s\|_V \leq \gamma\|r\|_{V_x}\|s\|_{V_t},$$

it is not necessary to reach the global minimum of (5.16) to ensure the convergence of the greedy algorithm. However, our proof cannot be straightforwardly extended to problems involving three or more spaces.

Theorem 5.4.3. *Assume that we consider only two Hilbert spaces V_x and V_t and that conditions (A1)-(A6) are satisfied. Then, if at each iteration $n \in \mathbb{N}$, the pair $(r_n, s_n) \in V_x \times V_x$ is chosen to be a **local** minimum of (5.16) such that $\mathcal{E}(u_n) < \mathcal{E}(u_{n-1})$, then $(u_n)_{n\in\mathbb{N}^*}$ still converges strongly in V towards the solution u of (5.14). Besides, if the Hilbert spaces V_x and V_t are finite dimensional, the rate of convergence of the algorithm is still exponential in n, i.e. there exists $C, \sigma > 0$ such that*

$$\|u_n - u\|_V \leq Ce^{-\sigma n}.$$

From these results, a natural strategy can be adopted to deal with an obstacle problem with uncertainty with a large number of random parameters (keeping in mind the original problem presented in the introduction): the idea is to use a penalized formulation of the obstacle problem to apply the Progressive Generalized Decomposition algorithm. We implemented the PGD algorithm on a particular toy obstacle problem and observed numerically that the usual fixed-point procedure used for the minimization of quadratic functionals does not converge for the resolution of the Euler equations associated to the minimization problem (5.15) when the penalization parameter is too large. A numerical procedure to find a suitable pair (r_n, s_n) ensuring the convergence of the greedy algorithm is proposed in Chapter 6 to circumvent this difficulty.

Let us mention that the results we proved in [39] were later extended by Nouy and Falco [156] to the case when Banach spaces (instead of Hilbert spaces) are considered, as well as for the orthogonal version of the greedy algorithm.

5.5 Appendix: Other methods for high-dimensional problems

Of course, greedy algorithms and PGD are not the only methods proposed in the literature to try to circumvent the curse of dimensionality. We present here a short and non-exhaustive review of other numerical methods used to deal with the resolution of high-dimensional differential equations.

5.5.1 Other tensor product representation formats

We focus here on the case when the Hilbert space V of multivariate functions can be written as a tensor product of Hilbert spaces of univariate functions $V = \bigotimes_{i=1}^{p} V_i$. For all $1 \leq i \leq p$, an element $u_i \in V_i$ may be viewed as a function $u_i : x_i \in \mathcal{X}_i \mapsto u_i(x_i)$ and an element $u \in V$ may be viewed as a multivariate function $(x_1, \cdots, x_p) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \mapsto u(x_1, \cdots, x_p)$.

The concepts that are reviewed here are explained with more details in [103, 120, 164].

Before introducing the main tensor formats that are encountered in the literature, let us recall the notion of *embedded manifold* [61].

Definition 5.5.1.

- *Let N and M be topological manifolds of respective dimensions $n \leq m$. A topological embedding of N in M is a continuous map $i : N \to M$ which carries N homeomorphically onto its image $i(N)$.*
- *If M is a manifold and $\mathcal{M} \subset M$, we say that $\mathcal{M}$ is an embedded manifold of M if there is a manifold N and an embedding $i : N \to M$ such that $\mathcal{M} = i(N)$.*

We will consider here the application of tensor formats to the resolution of minimization a given function $\mathcal{E} : V \to \mathbb{R}$. If $\mathcal{M}$ (which will correspond to a fixed set

of tensors with a given format) is an embedded manifold of V, approximation on $\mathcal{M}$ can be performed by the Dirac-Frenkel variational principle [144]. Denoting by $T_U\mathcal{M}$ the tangent space of $\mathcal{M}$ at $U \in \mathcal{M}$, the minimization of the functional $\mathcal{E}$ leads to the following equation

$$\left\{ \begin{array}{l} \text{find } U \in \mathcal{M} \text{ such that} \\ \forall V \in T_U\mathcal{M},\ \langle \mathcal{E}'(U), V\rangle_V = 0. \end{array} \right.$$

The numerical treatment of such an equation then requires that an efficient parametrization of the tangent space $T_U\mathcal{M}$ is available.

Let us now present the tensor formats which are mainly used in order to represent a function $u \in V$. The most classical method is the so-called *canonical decomposition* of the tensor u. This method uses a representation by r elementary products of single-variate functions:

$$(x_1, \cdots, x_p) \mapsto u(x_1, \cdots, x_p) = \sum_{k=1}^{r} \left(\bigotimes_{i=1}^{p} u_{i,k} \right) (x_1, \cdots, x_p). \tag{5.30}$$

In the case when $p = 2$, this format can be rewritten as

$$(x_1, x_2) \mapsto u(x_1, x_2) = \sum_{k=1}^{r} u_{1,k} \otimes u_{2,k}(x_1, x_2).$$

The number of terms r in the expression above is called the *canonical rank* of the function u. This decomposition can also be found in the literature under the names CANDECOMP or PARAFAC [120].

In the case when $\dim(V_i) = n$ for all $1 \leq i \leq p$, the complexity of the canonical representation is equal to rpn, which makes the canonical format a very popular choice for the treatment of high-dimensional problems [16]. This is indeed the choice that is made in the PGD approach we have explained in the previous section with $r = 1$.

However, the set of rank-r-tensors

$$\mathcal{C}_r := \left\{ u \in V,\ u(x_1, \cdots, x_p) = \sum_{k=1}^{r} \left(\bigotimes_{i=1}^{p} u_{i,k} \right) (x_1, \cdots, x_p),\ \forall 1 \leq k \leq r,\ \forall 1 \leq i \leq p,\ u_{i,k} \in V_i \right\},$$

is not a weakly closed subset of V as soon as $d \geq 3$ and $r \geq 2$ [69]. This implies that there may not exist a best approximation, i.e. there may not exist a minimizer to the problem

$$\inf_{\widetilde{u} \in \mathcal{C}_r} \|u - \widetilde{u}\|_V.$$

Moreover, $\mathcal{C}_r$ is not an embedded manifold, which makes difficult the identification of a tangent space, which is needed in practice for the resolution of high-dimensional partial differential equations.

Other tensor formats exist in the literature to avoid the shortcomings linked to the canonical representation of a tensor. Among them, the *Tucker decomposition* uses the following representation of a tensor u:

$$(x_1, \cdots x_p) \mapsto u(x_1, \cdots, x_p) = \sum_{k_1=1}^{r_1} \cdots \sum_{k_p=1}^{r_p} c_{k_1,\cdots,k_p} \left(\bigotimes_{i=1}^{p} u_{i,k_i} \right) (x_1, \cdots, x_p),$$

where for all $1 \leq i \leq p$, $r_i \in \mathbb{N}^*$, $u_{i,k_i} \in V_i$ for all $1 \leq k_i \leq r_i$ and $c_{k_1,\cdots,k_p} \in \mathbb{R}$. In the case $p = 2$ this decomposition can be rewritten as

$$(x_1, x_2) \mapsto u(x_1, x_2) = \sum_{k=1}^{r_1} \sum_{k'=1}^{r_2} c_{k,k'} u_{1,k} \otimes u_{2,k'}(x_1, x_2),$$

Actually, in the Tucker format, the function u is decomposed over all the possible tensor products between the functions $(u_{i,k_i})_{1 \leq k_i \leq r_i}$ for all $1 \leq i \leq p$.

The vector $\mathbf{r}_T := (r_1, \cdots, r_p) \in (\mathbb{N}^*)^p$ is called the Tucker rank of u. For any function $u \in V$ which has a finite-rank Tucker decomposition, the Tucker rank $\mathbf{r}_T$ is unique. Besides, $\mathcal{T}_{\mathbf{r}_T}$, which is the set of tensors of Tucker rank $\mathbf{r}_T$, is weakly closed in V, which ensures the existence of best approximations, and is an embedded manifold [118], and its tangent space is well-characterized. Thus, it possesses all the desirable theoretical properties but, unfortunately, its complexity still scales exponentially with the number of Hilbert spaces p. Actually, if $\mathbf{r}_T = (r, \cdots, r)$ and if $\dim V_i = n$ for all $1 \leq i \leq p$, the complexity scales like $\mathcal{O}(r^p + nrp)$, which limits the applicability of the Tucker format for very large p.

The *tensor train (TT) decomposition* [158] enables one to get rid of this exponential complexity. In this representation, the function u is decomposed as

$$\begin{aligned}(x_1, \cdots, x_p) &\mapsto u(x_1, \cdots, x_p) \\ &= \sum_{k_1=1}^{r_1} \cdots \sum_{k_{p-1}=1}^{r_{p-1}} U_1(x_1, k_1) U_2(k_1, x_2, k_2) \cdots U_{p-1}(k_{p-2}, x_{p-1}, k_{p-1}) U_p(k_{p-1}, x_p).\end{aligned}$$

The TT rank of u is then defined as $\mathbf{r}_{TT} = (r_1, \cdots, r_{p-1}) \in (\mathbb{N}^*)^{p-1}$. In the above expression, the function u is represented by matrix-valued functions $U_1(x_1), \cdots, U_p(x_p)$ where

$$\begin{aligned} x_1 &\mapsto U_1(x_1) \in \mathbb{R}^{1 \times r_1}, \\ x_2 &\mapsto U_2(x_2) \in \mathbb{R}^{r_1 \times r_2}, \\ &\cdots \\ x_{p-1} &\mapsto U_{p-1}(x_{p-1}) \in \mathbb{R}^{r_{p-2} \times r_{p-1}}, \\ x_p &\mapsto U_p(x_p) \in \mathbb{R}^{r_{p-1} \times 1}. \end{aligned}$$

Thus, the function u can be written as a product of matrices

$$u(x_1, \cdots, x_p) = U_1(x_1) \cdots U_p(x_p).$$

Such a decomposition is also called a *matrix product state*. Again, the set of functions of TT rank at most $\mathbf{r}_{TT}$ is a weakly closed subset of V and an embedded manifold which possesses a stable local parametrization of its tangent space [113]. Besides, its storage complexity scales like $\mathcal{O}(r^2 np)$ which enables one to get rid of the exponential dependence in the dimension of the Tucker format. That is why the TT format is a popular way for treating high-dimensional problems.

The *hierarchical Tucker (HT) format* introduced in [104] is a generalization of the TT format, which uses a hierarchical splitting, described by a binary dimension partition tree.

Let us highlight here the link between these different tensor formats and methods used in the field of quantum chemistry [164]. The canonical representation with $r = 1$ corresponds to the Hartree-Fock model, where the electronic problem is defined over the set of Slater determinants, which are antisymmetrized versions of tensor product functions. The Tucker format is applied in the multi-configurational self-consistent field approach and the TT format is closely linked to the Density Matrix Renormalization Group algorithm.

The most common algorithms used to compute in practice local best approximations of a given tensor in the Tucker, TT or HT format usually perform a series of successive Singular Value Decompositions [103]. For the Tucker format, the most common algorithms are higher order orthogonal iteration (HOOI) [68], Newton-Grassman approach [82] or Higher Order Singular Value Decomposition (HOSVD) [67]. Similar algorithms were recently proposed independently by Oseledets [158] for the TT format and Grasedyck [96] for the HT format.

For the treatment of high-dimensional optimization problems, the main methods at hand are either the use of greedy algorithms (as exposed in Section 5.3) or Alternating Least Square approaches [112]. The latter is a generalization of the fixed-point algorithm presented in Section 5.3 to more general tensor formats, which consists in (i) choosing a given variate x_i with $1 \leq i \leq p$, (ii) computing the optimal (in a certain sense) single-variate functions $U(x_i)$ while keeping fixed the functions depending on the other variates, and, (iii) repeating sequentially the procedure for all $1 \leq i \leq p$.

5.5.2 Sparse grids

The sparse grid method is a numerical discretization technique for multivariate problems. This approach was first introduced by Smolyak [169] then developed by Zenger [184]. A very complete introduction to sparse grids can be found in [32]. We will only give here a brief overview of this method. Sparse grid methods are also known as hyperbolic cross points or splitting interpolations.

Let $\mathcal{X} = [0, 1]$. The sparse grid method relies on the use of multilevel bases which are constructed as follows. In the classical approach, the hierarchical basis functions are constructed with the standard hat function

$$\phi(\xi) := \begin{cases} 1 - |\xi| & \text{if } \xi \in [-1, 1], \\ 0 & \text{otherwise.} \end{cases}$$

Then, a set of uniform grids of level m and mesh width $h_m = 2^{-m}$ are introduced. The grid points $\xi_{m,i}$ are defined as

$$\xi_{m,i} := i2^{-m}, \quad 0 \leq i \leq 2^m.$$

To these grid points is associated a family of basis functions $(\phi_{m,i})_{1 \leq i \leq 2^m - 1}$ defined by

$$\phi_{m,i}(x) := \phi\left(\frac{x - x_{m,i}}{2^{-m}}\right).$$

This basis is the standard basis of the set of $\mathbb{P}_1$ Lagrange finite element functions with mesh size 2^{-m} that are equal to zero on the boundary of $\mathcal{X}$. It is usually called the *nodal basis* or *Lagrange basis*. Let us denote by

$$V_m := \text{Span}\left\{\phi_{m,i},\ 1 \leq i \leq 2^m - 1\right\}$$

the space of $\mathbb{P}_1$ finite element functions with mesh size 2^{-m}. Let us also introduce the hierarchical increment spaces W_m, defined by

$$W_m := \text{Span}\left\{\phi_{m,i},\ i \in I_m\right\},$$

where

$$I_m := \left\{i \in \mathbb{N},\ 1 \leq i \leq 2^m - 1,\ i \text{ odd}\right\}.$$

These increment spaces then satisfy the relation

$$V_m = \bigoplus_{k \leq m} W_k,$$

where the notation $\bigoplus$ means that the sum is direct. The natural basis corresponding to this decomposition, i.e. $(\phi_{i,k})_{k \leq m,\, i \in I_k}$ is called the *hierarchical basis* of V_m and any continuous piecewise linear function $u \in V_m$ can be uniquely decomposed as

$$u = \sum_{k=1}^{m} \sum_{i \in I_k} u_{k,i} \phi_{k,i},$$

where $u_{k,i} \in \mathbb{R}$ for all $1 \leq k \leq m$ and $i \in I_k$.

Before describing how the above discretization can be generalized to high-dimensional spaces, let us introduce some notation. If $\mathbf{i} = (i_1, \cdots, i_p) \in \mathbb{N}^p$ and $\mathbf{k} = (k_1, \cdots, k_p) \in \mathbb{N}^p$ are multi-indices, the notation

$$\mathbf{i} \leq \mathbf{k}$$

means

$$\forall 1 \leq j \leq p,\ i_j \leq k_j.$$

Besides, $2^{\mathbf{i}}$ denotes the multi-index $(2^{i_1}, \cdots, 2^{i_p}) \in \mathbb{N}^p$ and $\mathbf{1}$ the multi-index $(1, \cdots, 1) \in \mathbb{N}^p$.

A multidimensional hierarchical basis of $\mathcal{X}^p = [0,1]^p$ can be constructed by using the p-dimensional tensorization of the one-dimensional basis $(\phi_{k,i})_{1 \leq k \leq m,\, i \in I_k}$. In other words, if $\mathbf{m} = (m_1, \cdots, m_p) \in \mathbb{N}^p$ is a multi-index denoting the level of the discretization in each dimension, we introduce the set of grid points $\mathbf{x}_{\mathbf{m},\mathbf{i}}$ defined as

$$\begin{aligned} &\mathbf{x}_{\mathbf{m},\mathbf{i}} = (x_{m_1,i_1}, \cdots, x_{m_p,i_p}) \text{ where } \mathbf{i} = (i_1, \cdots, i_p) \in \mathbb{N}^p \\ &\text{with } \mathbf{1} \leq \mathbf{i} \leq 2^{\mathbf{m}}. \end{aligned}$$

Then, for each grid point $\mathbf{x}_{\mathbf{m},\mathbf{i}}$, we introduce the associated p-dimensional basis function $\phi_{\mathbf{m},\mathbf{i}}$ which is defined as the tensor product function

$$\phi_{\mathbf{m},\mathbf{i}}(x_1, \cdots, x_p) := \Pi_{j=1}^{p} \phi_{m_j,i_j}(x_j).$$

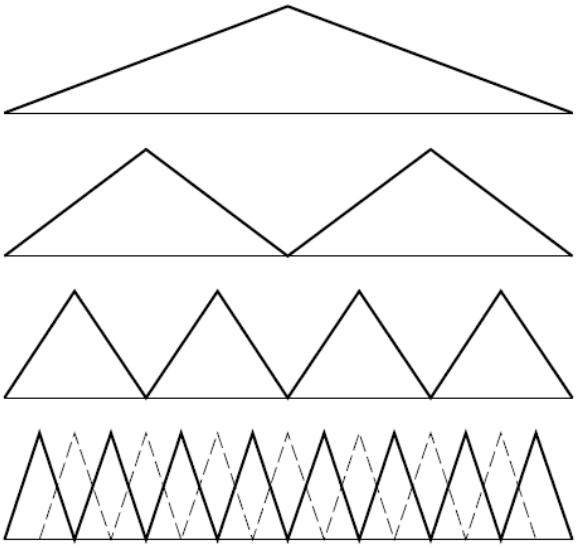

Figure 5.3: Piecewise linear hierarchical basis (solid line) and nodal point basis (dashed line) [32]

The family $(\phi_{\mathbf{m},\mathbf{i}})_{\mathbf{1}\leq\mathbf{i}\leq 2^{\mathbf{m}}-\mathbf{1}}$ forms a nodal basis of the set of continuous piecewise p-linear functions equal to zero on the boundary of $\mathcal{X}^p$

$$V_{\mathbf{m}} := \text{Span}\left\{\phi_{\mathbf{m},\mathbf{i}},\ \mathbf{1}\leq\mathbf{i}\leq 2^{\mathbf{m}}-\mathbf{1}\right\}. \tag{5.31}$$

Similarly to the one-dimensional case, the hierarchical increments $W_{\mathbf{m}}$ are defined by

$$W_{\mathbf{m}} := \text{Span}\left\{\phi_{\mathbf{m},\mathbf{i}},\ \mathbf{i}\in\mathbf{I}_{\mathbf{m}}\right\}$$

where

$$\mathbf{I}_{\mathbf{m}} := \left\{\mathbf{i}\in\mathbb{N}^p,\ \mathbf{1}\leq\mathbf{i}\leq 2^{\mathbf{m}}-\mathbf{1},\ i_j \text{ odd for all } 1\leq j\leq p\right\}.$$

Thus, we still have the identity

$$V_{\mathbf{m}} = \bigoplus_{\mathbf{k}\leq\mathbf{m}} W_{\mathbf{k}},$$

so that any function $u\in V_{\mathbf{m}}$ can be uniquely decomposed as

$$u(\mathbf{x}) = \sum_{\mathbf{1}\leq\mathbf{k}\leq\mathbf{m}}\ \sum_{\mathbf{i}\in\mathbf{I}_{\mathbf{k}}} u_{\mathbf{k},\mathbf{i}}\phi_{\mathbf{k},\mathbf{i}}(\mathbf{x}),$$

with hierarchical coefficients $u_{\mathbf{k},\mathbf{i}}\in\mathbb{R}$.

Let us introduce the following norms for multi-indices $\mathbf{k}=(k_1,\cdots,k_p)\in\mathbb{N}^p$,

$$|\mathbf{k}|_1 := \sum_{j=1}^{p}|k_j| \quad \text{and} \quad |\mathbf{k}|_\infty := \max_{1\leq j\leq p}|k_j|.$$

The idea of sparse grids is to rely on this hierarchical finite element basis but to keep only a small number of grid points, in order to optimize the ratio between the accuracy of the resulting approximation and the number of degrees of freedom. This rate can be optimized by considering the *sparse grid* spaces $\widehat{V}_n$ of level n defined by

$$\widehat{V}_n := \bigoplus_{|\mathbf{k}|_1 \leq n+p-1} W_{\mathbf{k}}.$$

The corresponding *full grid* space is equal to

$$V_n := \bigoplus_{|\mathbf{k}|_\infty \leq n} W_{\mathbf{k}}.$$

Actually, denoting by $\mathbf{m} = n(1, \cdots, 1)$, it holds that $V_n = V_{\mathbf{m}}$ where $V_{\mathbf{m}}$ is defined by (5.31). The full grid space V_n actually corresponds to the discretization space associated with a standard $\mathbb{P}_1$ finite element discretization based on a uniform discretization of mesh size $h_n = 2^{-n}$. The dimension of the space V_n scales like $\mathcal{O}(h_n^{-p})$.

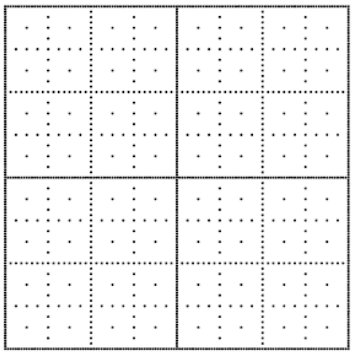

Figure 5.4: Regular sparse grid in dimension 2 [32]

The number of degrees of freedom of the sparse grid space $\widehat{V}_n$ is equal to

$$\begin{aligned} \dim \widehat{V}_n &= \sum_{i=0}^{n-1} 2^i \begin{pmatrix} p-1+i \\ p-1 \end{pmatrix} \\ &= \mathcal{O}\left(h_n^{-1} |\log_2(h_n)|^{p-1}\right). \end{aligned}$$

Thus, the dimension of the sparse grid space $\widehat{V}_n$ is significantly lower than the dimension of the full grid space V_n.

It remains to compare the approximation property of the two spaces V_n and $\widehat{V}_n$. To do this, we consider functions belonging to the anisotropic Sobolev space

$$H^{2,mix}(\mathcal{X}^p) := \left\{ u \in L^2(\mathcal{X}^p),\ \partial^\alpha u \in L^2(\mathcal{X}^p),\ \alpha \in \mathbb{N}^p,\ |\alpha|_\infty \leq 2 \right\},$$

and denote by Π_{V_n} (respectively $\Pi_{\widehat{V}_n}$) the $L^2(\mathcal{X}^p)$-orthogonal projector of $L^2(\mathcal{X}^p)$ onto V_n (respectively onto $\widehat{V}_n$). It holds that for all $u \in H^{2,mix}(\mathcal{X}^p) \cap H_0^1(\mathcal{X}^p)$, the approximation error of the function u on the sparse grid space is

$$\left\|u - \Pi_{\widehat{V}_n} u\right\|_{L^2(\mathcal{X}^p)} = \mathcal{O}\left(h_n^2 |\log_2 h_n|^{p-1}\right),$$

whereas the approximation accuracy on the full grid space is

$$\|u - \Pi_{V_n} u\|_{L^2(\mathcal{X}^p)} = \mathcal{O}\left(h_n^2\right).$$

This shows that the use of sparse grids is a linear approximation method which enables us to decrease drastically the dimension of a finite element space while keeping a reasonably good accuracy of the numerical solution if the exact solution is regular enough. Strangely enough, the appropriate Sobolev spaces to consider are the same spaces as those highlighted by Figueroa et al. to ensure good convergence rates of the greedy algorithms (see Section 5.3.4). Other sparse grids optimal with respect to other Sobolev norms can be found in [98]. Some recent developments concern the use of adaptive sparse grids which are chosen in order to optimize the error made by the approximation of a particular function u [31]. These methods fall back into the scope of nonlinear approximation.

5.5.3 Reduced basis

The *reduced basis* (RB) method [30] aims at approximating the solution of parameter-dependent partial differential equations. It relies on the computation of the solution for a small number of well-chosen values of the parameter in a preliminary "off-line" stage. These functions then form the Galerkin basis of a discretization space used to solve the differential equation for any value of the parameter in an "online" stage.

More precisely, let $\mathcal{T} \subset \mathbb{R}^p$ be a closed compact parameter domain and $\mathcal{X} \subset \mathbb{R}^d$ be the spatial domain. For each $t \in \mathcal{T}$, the solution $u(\cdot, t)$ is assumed to belong to a Hilbert space $V_x \subset L^2(\mathcal{X})$. The weak formulation of the parametric differential equation reads

$$a(u(\cdot, t), v; t) = l(v), \quad \forall v \in V_x, \tag{5.32}$$

where the form l is continuous on V_x and the bilinear form $a(\cdot, \cdot; t)$ is symmetric, continuous and coercive on V_x, uniformly with respect to the parameter $t \in \mathcal{T}$.

The reduced basis method consists in precomputing solutions $u(\cdot, t_i)$ for well-chosen parameters $(t_i)_{1 \leq i \leq n}$ with $n \in \mathbb{N}^*$ and then use the discretization space $V_n := \mathrm{Span}\{u(\cdot, t_i),\ 1 \leq i \leq n\}$ through a Galerkin procedure. In other words, for each $t \in \mathcal{T}$, an approximation $u_n(\cdot, t) \in V_n$ of the exact solution $u(\cdot, t)$ is computed as the solution of

$$a(u_n(\cdot, t), v_n; t) = l(v_n), \quad \forall v_n \in V_n. \tag{5.33}$$

From Cea's lemma, it holds

$$\|u(\cdot, t) - u_n(\cdot, t)\|_{V_x} \leq C \inf_{v_n \in V_n} \|u(\cdot, t) - v_n\|_{V_x},$$

where C is a positive constant independent of $n \in \mathbb{N}$ and $t \in \mathcal{T}$.

Thus, the RB method produces an accurate approximation of $u(\cdot, t)$ for any $t \in \mathcal{T}$ provided that the set $F = \{u(\cdot, t),\ t \in \mathcal{T}\}$ is well-approximated by the finite-dimensional space V_n.

The *Kolmogorov n-width* gives a good indication on how well a subset $F \subset V_x$ can be approximated by a n-dimensional linear subspace. It is defined as

$$d_n(F) := \inf_{Y_n \subset V_x;\, \dim Y_n = n} \sup_{u \in F} \inf_{v_n \in Y_n} \|u - v_n\|_{V_x}.$$

In the case when $d_n(F)$ decays rapidly with increasing n, the reduced basis method is likely to provide a good approximation of the solution $u(\cdot, t)$ for any $t \in \mathcal{T}$.

The difficulty now relies on finding an appropriate set of parameters $(t_i)_{1 \leq i \leq n}$ such that $\sup_{u \in F} \inf_{v_n \in V_n} \|u - v_n\|_{V_x}$, with $V_n := \mathrm{Span}\{u(\cdot, t_i),\ 1 \leq i \leq n\}$ is close to the Kolmogorov n-width of the set F. Greedy algorithms stand as the state-of-the-art technique to find such a subset with good approximation properties in practice. Such an algorithm reads

1. compute $t_1 \in \operatorname{argmax}_{t \in \mathcal{T}} \|u(\cdot, t)\|_{V_x}$;
2. for $n \geq 2$, assume that $t_1, \cdots, t_{n-1}$ are defined, let $V_{n-1} := \mathrm{Span}\{u(t_i),\ 1 \leq i \leq n-1\}$. Then,

$$t_n \in \operatorname*{argmax}_{t \in \mathcal{T}} \|u(\cdot, t) - u_{n-1}(\cdot, t)\|_{V_x},$$

where $u_{n-1}(\cdot, t)$ is the solution to the Galerkin problem

$$\begin{cases} \text{find } u_{n-1}(\cdot, t) \in V_{n-1} \text{ such that} \\ a(u_{n-1}(\cdot, t), v_{n-1}; t) = l(v_{n-1}), \quad \forall v_{n-1} \in V_{n-1}. \end{cases}$$

In practice, when F is given as the set of solutions of the parametrized equation (5.32), quantities of the form $\sup_{t \in \mathcal{T}} \|u(\cdot, t)\|_{V_x}$ or $\sup_{t \in \mathcal{T}} \|u(\cdot, t) - u_{n-1}(\cdot, t)\|_{V_x}$ cannot be easily computed. Instead, a posteriori error estimators are used in order to estimate this quantities. For a parametrized problem of the form (5.32), a standard estimator can be defined by

$$\Delta_{n-1}(t) := \|r_{n-1}(\cdot, t)\|_{V_x'},$$

where $r_{n-1}(\cdot, t)$ is the residual, i.e.

$$\forall v \in V_x,\ r_{n-1}(v; t) = l(v) - a\left(u_{n-1}(\cdot, t), v; t\right).$$

Then, the algorithm becomes: for all $n \in \mathbb{N}$, $t_n \in \operatorname*{argmax}_{t \in \mathcal{T}} \Delta_{n-1}(t)$. In practice, to compute a solution to this maximization problem, the set of parameters $\mathcal{T}$ must be replace by a discrete trial set $\widetilde{\mathcal{T}} \subset \mathcal{T}$ sufficiently large to represent efficiently the full set of parameters $\mathcal{T}$. Actually, the exploration of the trial set $\widetilde{\mathcal{T}}$ may lead to a quite expensive algorithm.

Reduced Basis methods have been used in the context of UQ by Boyaval and al. [25, 26]. Let us also mention that RB methods can be used for the discretization

of high-dimensional nonlinear problems through the use of a special interpolation technique, called the *magic points* [145].

Convergence rates for the full RB algorithm have not been proved yet. However, some results have been obtained in [17, 30] in the following particular setting: F being a general compact subset of V_x, we consider the algorithm

- $u_1 \in \underset{u \in F}{\operatorname{argmax}} \|u\|_{V_x}$;
- for $n \geq 2$, assume that $u_1, \cdots, u_{n-1}$ are defined, and let $V_{n-1} := \operatorname{Span}\{u_i,\ 1 \leq i \leq n-1\}$. Then,
$$u_n \in \underset{u \in F}{\operatorname{argmax}} \|u - P_{n-1}u\|_{V_x},$$
where P_{n-1} is the V_x-orthogonal projector onto V_{n-1}.

Let us denote by
$$\sigma_n(F) := \sup_{u \in F} \inf_{v_n \in V_n} \|u - v_n\|_{V_x}.$$
A natural mathematical question is then the following: how well does $\sigma_n(F)$ compare with the best approximation error defined by the Kolmogorov n-width $d_n(F)$? Several answers to this question are given in [17, 30].

In particular, it is proved that, if F is a compact subset of the Hilbert space V_x, then
$$\sigma_n(F) \leq \frac{2^{n+1}}{\sqrt{3}} d_n(F). \tag{5.34}$$
Besides, an explicit example of compact set F is given for which $\sigma_n(F) \geq c2^n d_n(F)$ where c is a constant independent of $n \in \mathbb{N}$, so that the estimate (5.34) is optimal in the general case. However, except when $d_n(F)$ decays exponentially fast, estimate (5.34) is not satisfactory since it does not even imply that $\sigma_n(F) \underset{n \to \infty}{\longrightarrow} 0$.

Better estimates can be achieved provided that some additional assumptions are made on the decay rate of $d_n(F)$. Typically, the authors of [17] prove that, if the set F is compact and if the Kolmogorov n-width $d_n(F)$ decays polynomially in n, then so does $\sigma_n(F)$ and the decay rate remains the same. More precisely,

Theorem 5.5.1. *Let F be a compact subset of V_x. Suppose that there exists $M, \alpha > 0$ such that $d_0(F) \leq M$ and*
$$d_n(F) \leq Mn^{-\alpha},\ n \in \mathbb{N}.$$
Then,
$$\sigma_n(F) \leq CMn^{-\alpha}, n \in \mathbb{N},$$
for some constant $C > 0$ independent on $n \in \mathbb{N}$.

A similar result can be obtained when $d_n(F)$ is assumed to decay as e^{-an^α} for some $a, \alpha > 0$.

Theorem 5.5.2. *Let F be a compact subset of V_x. Suppose that*

$$d_n(F) \leq Me^{-an^\alpha},\ n \in \mathbb{N},$$

for some $M, a, \alpha > 0$. Then, setting $\beta := \frac{\alpha}{\alpha+1}$, one has

$$\sigma_n(F) \leq CMe^{-cn^\beta},\ n \in \mathbb{N}, \tag{5.35}$$

for some $c, C > 0$ independent on $n \in \mathbb{N}$.

Note that (5.35) is asymptotically sharper than estimate (5.34) as long as $\alpha > 1$, whereas the converse holds when $\alpha < 1$. When $\alpha = 1$, the comparison depends on the value of the parameters a and c.

5.5.4 High dimensional sparse polynomial approximations

The approach presented in thi section was recently proposed in [59], in order to deal with approximation of stochastic or parametrized PDEs with a possibly infinite number of parameters. The prototypical problem we consider is the following:

$$\begin{cases} \text{For all } y = (y_j)_{j\in\mathbb{N}} \in U = [-1,1]^{\mathbb{N}}, \text{ find } u(\cdot, y) \in H_0^1(\mathcal{X}) \text{ such that} \\ -\mathrm{div}_x(a\nabla_x u(\cdot, y)) = f \text{ in } \mathcal{D}'(\mathcal{X}). \end{cases} \tag{5.36}$$

where $f \in H^{-1}(\mathcal{X})$ and $a = a(x, y)$ is a variable diffusion coefficient depending on $x \in \mathcal{X}$ and on a vector y of parameters in an affine manner:

$$a = a(x, y) = \overline{a}(x) + \sum_{j\in\mathbb{N}} y_j\psi_j(x),\ x \in \mathcal{X},\ y = (y_j)_{j\in\mathbb{N}} \in U = [-1, 1]^{\mathbb{N}},$$

where $(\psi_j)_{j\in\mathbb{N}}$ is a family of functions. In the sequel, we will denote by $V_x = H_0^1(\mathcal{X})$, endowed with the scalar product

$$\forall v, w \in V_x,\ \langle v, w\rangle_{V_x} = \int_{\mathcal{X}} \nabla v \cdot \nabla w.$$

A uniform ellipticity condition is assumed on the diffusion coefficient a, i.e. there exists $\alpha, \beta > 0$ such that

$$\forall x \in \mathcal{X},\ \forall y \in U,\ 0 < \alpha \leq a(x, y) \leq \beta. \tag{5.37}$$

Besides, it is assumed that the sequence $(\|\psi_j\|_{L^\infty(\mathcal{X})})_{j\in\mathbb{N}}$ belongs to $l^p(\mathbb{N})$ for some $0 < p < 1$.

The objective of the high dimensional sparse polynomial approximation is to compute in a reasonable amount of time a good approximation of $u(y)$ for any value of $y \in U$.

Let $\mathcal{F}$ be the set of all finitely supported sequences $\nu = (\nu_j)_{j\in\mathbb{N}}$ of integers. Under the above assumptions, it can be proved [58] that at any $y \in U$, the function $y \mapsto u(\cdot, y)$ admits a partial derivative $\partial_y^\nu u(\cdot, y) \in V_x$ for any $\nu \in \mathcal{F}$. A way to study sparse polynomial approximation of the function u is to study the convergence of the Taylor series

$$u(y) = \sum_{\nu\in\mathcal{F}} t_\nu y^\nu,$$

where

$$y^\nu := \Pi_{j\in\mathbb{N}} y_j^{\nu_j}.$$

The Taylor coefficients $t_\nu = t_\nu(x) \in V_x$ are

$$t_\nu := \frac{1}{\nu!}\partial^\nu u|_{y=0} \quad \text{with} \quad \nu! := \Pi_{j\in\mathbb{N}}\nu_j!.$$

The objective is then to identify a set $\Lambda \subset \mathcal{F}$ with $\text{card}(\Lambda) \leq N$ and such that u is well-approximated in the space

$$V_\Lambda = \left\{ \sum_{\nu\in\Lambda} c_\nu y^\nu \mid \forall \nu \in \Lambda,\ c_\nu \in V_x \right\},$$

for example by the Taylor expansion $u_\Lambda := \sum_{\nu\in\Lambda} t_\nu y^\nu$.

In the literature, a priori choices of sets Λ have been proposed to design linear approximation methods, for instance sparse grids with the following conditions $\sum_{j\in\mathbb{N}} \alpha_j\nu_j \leq A(N)$ or $\Pi_{j\in\mathbb{N}}(1+\beta_j\nu_j) \leq B(N)$. Instead, in the approach proposed by Cohen et al., the set Λ is chosen in order to be optimally adapted to u.

It holds that for all $y \in U$,

$$\|u(y) - u_\Lambda(y)\|_{V_x} \leq \left\| \sum_{\nu\notin\Lambda} t_\nu y^\nu \right\|_{V_x} \leq \sum_{\nu\notin\Lambda} \|t_\nu\|_{V_x}.$$

A natural strategy is to consider the best N-term approximation of $u(y)$ in the $l^1(\mathcal{F})$ norm, that is to take for Λ the set of indices $\nu \in \mathcal{F}$ corresponding to the largest N values of $\|t_\nu\|_{V_x}$. It can be proved that provided that the sequence $(\|t_\nu\|_{V_x})_{\nu\in\mathcal{F}}$ is in $l^p(\mathcal{F})$ for some $0 < p < 1$, for this particular choice of Λ,

$$\sum_{\nu\notin\Lambda} \|t_\nu\|_{V_x} \leq CN^{-s} \quad \text{with } s := \frac{1}{p} - 1.$$

A natural question then arises: what are necessary conditions ensuring that the sequence $(\|t_\nu\|_{V_x})_{\nu\in\mathcal{F}}$ is in $l^p(\mathcal{F})$? The following theorem was proved in [58].

Theorem 5.5.3. *Under the uniform ellipticity assumption (5.37), for any* $0 < p < 1$,

$$(\|\psi_j\|_{L^\infty(\mathcal{X})})_{j\in\mathbb{N}} \in l^p(\mathbb{N}) \Rightarrow (\|t_\nu\|_{V_x})_{\nu\in\mathcal{F}} \in l^p(\mathcal{F}).$$

Thus, the Taylor expansion of u inherits the sparsity properties of the sequence $(\|\psi_j\|_{L^\infty(\mathcal{X})})_{j\in\mathbb{N}}$. Besides, if the set $\mathcal{K} := \{u(y), y \in U\}$ is a compact subset of V_x and if we denote by

$$E_\Lambda := \text{Span}\{t_\nu; \nu \in \Lambda\},$$

with Λ corresponding to the largest N values of $\|t_\nu\|_{V_x}$, then it holds that

$$d_N(\mathcal{K}) \leq \max \|u(y) - u_\Lambda(y)\|_{V_x} \leq CN^{-s}.$$

The curse of dimensionality is avoided since, in this error bound, there is no dependence on the (possibly infinite) number of parameters which come into play

in equation (5.36). Such approximation rates cannot be proved for usual a priori choices of Λ.

Different strategies, in particular iterative adaptive strategies [57] have been proposed to construct a sequence of sets $(\Lambda_N)_{N\in\mathbb{N}^*}$ for which there exists $C > 0$ such that

$$\forall N \in \mathbb{N}^*,\ \mathrm{card}(\Lambda_N) \leq N \ \text{ and } \ \max \|u(y) - u_{\Lambda_N}(y)\|_{V_x} \leq CN^{-s}.$$

5.5.5 Compressed Sensing

Compressed sensing (CS), also known as compressive sampling, is a family of methods in the field of signal reconstruction originally developed by the works of Candès, Romberg and Tao [45, 46, 47] and Donoho [76].

The CS theory aims at providing efficient algorithms to reconstruct *sparse* signals. Let us precise this notion. Let V_x and V_t be two Hilbert spaces (which will later refer to Hilbert spaces of functions depending respectively on the variables $x \in \mathcal{X}$ and $t \in \mathcal{T}$) and let $u \in V_x \otimes V_t$. Let $(\psi_i)_{i\in\mathbb{N}^*}$ be an orthonormal basis of the space V_t, so that u can be decomposed as

$$u = \sum_{i=1}^{+\infty} c_i \otimes \psi_i,$$

where for all $i \in \mathbb{N}^*$, $c_i \in V_x$. For the sake of simplicity, we will assume that u can be expanded as

$$u = \sum_{i=1}^{N} c_i \otimes \psi_i, \tag{5.38}$$

where N is a possibly very large integer. Let us denote by $\mathbf{c} = (c_i)_{1\leq i\leq N} \in V_x^N$.

For an arbitrary integer $s \in \mathbb{N}$, the signal u of the form (5.38) is said to be *s-sparse* in the basis $(\psi_i)_{i\in\mathbb{N}^*}$ if

$$\mathrm{Card}\left(\{1 \leq i \leq N,\ \|c_i\|_{V_x} \neq 0\}\right) \leq s.$$

More generally, a signal u is said to be sparse, if it is s-sparse with $s \ll N$.

In practice, since N is potentially very large, we cannot have access to the full sequence $(c_i)_{1\leq i\leq N}$ which would completely characterize the function u. The basis of the CS theory is to assume that only a few pieces of information about the signal u is known, namely only n elements of V_x, $\mathbf{y} = (y_1, \cdots, y_n) \in V_x^n$ with n being very small compared to N. The vector $\mathbf{y}$, also called the *information vector* , can be written as

$$\mathbf{y} = \Phi\mathbf{c},$$

where $\Phi \in \mathbb{R}^{n\times N}$ is called the *representing matrix*.

The CS theory aims at solving the following inverse problem: find the signal vector $\mathbf{c}$ knowing the measurement vector y and the sensing matrix Φ. To circumvent the ill-posedness of this problem, keeping in mind the fact that the signal u is sparse, CS rather focuses on solving

$$\mathbf{c} \in \underset{\widetilde{\mathbf{c}}\in V_x^N}{\mathrm{argmin}} \|\widetilde{\mathbf{c}}\|_{l_0} \quad \text{subject to} \quad \Phi\widetilde{\mathbf{c}} = \mathbf{y}, \tag{5.39}$$

where $\|\widetilde{\mathbf{c}}\|_{l_0} = \text{Card}\,\{1 \leq i \leq N \mid \widetilde{c}_i \neq 0\}$.

It is explained in [79] how the CS formalism can be used in the field of high-dimensional parametrized partial differential equations.

Let us assume that the signal $u \in V_x \times V_t$ is actually the solution of a parametrized PDE for almost all $t \in \mathcal{T}$,

$$\mathcal{A}(u(x,t);x,t) = b(x,t) \text{ in } \mathcal{D}'(\mathcal{X}),$$

(recall the UQ problem (5.6) for instance), and that a computational code allowing to find the solution $u(\cdot,t)$ for any value $t \in \mathcal{T}$, is available. However, in practice, this code is often expensive to run, so that we can only have access to solutions $u(\cdot,t_1)$, ..., $u(\cdot,t_n)$ for a small number n of values of t.

From this information, we would like to compute an approximation of the full solution u under the form

$$u(x,t) = \sum_{i=1}^{N} c_i(x)\psi_i(t),$$

where $(\psi_i)_{1\leq i\leq N}$ is an orthonormal family of functions of V_t fixed a priori, and where the coefficients $c_i \in V_x$ are to be computed for all $1 \leq i \leq N$ (see for instance the spectral methods presented in Section 5.2.4). Actually, setting

$$\Phi = \begin{pmatrix} \psi_1(t_1) & \cdots & \psi_N(t_1) \\ \vdots & \ddots & \vdots \\ \psi_1(t_n) & \cdots & \psi_N(t_n) \end{pmatrix},$$

and denoting by $y_i = u(\cdot,t_i)$ for all $1 \leq i \leq n$, reconstructing the whole function u is exactly the same kind of problem as the CS problem.

We present here different algorithms proposed for dealing with standard CS problems for signal reconstruction. For the sake of simplicity, we will assume here that $V_x = \mathbb{R}$, but these algorithms can be easily generalized to arbitrary Hilbert spaces V_x.

Let us recall that we aim at solving the inverse problem: find $\mathbf{c}^*$ such that

$$\mathbf{c} \in \underset{\widetilde{\mathbf{c}}\in\mathbb{R}^N}{\text{argmin}} \|\widetilde{\mathbf{c}}\|_{l_0} \quad \text{subject to} \quad \Phi\widetilde{\mathbf{c}} = \mathbf{y}. \tag{5.40}$$

Compressed sensing theory relies on the assumption that the sensing matrix Φ obeys the so-called *restricted isometry property* (RIP) which will be introduced below. For each $s \in \mathbb{N}^*$, we define the isometry constant δ_s of a matrix Φ as the smallest real number such that

$$\forall \mathbf{c} \in \mathbb{R}^N,\ (1-\delta_s)\|\mathbf{c}\|_{l_2}^2 \leq \|\Phi\mathbf{c}\|_{L_2}^2 \leq (1+\delta_s)\|\mathbf{c}\|_{l_2}^2.$$

A matrix Φ is said to obey the RIP of order s if δ_s is smaller than 1. When this property holds, Φ approximately preserves the Euclidean length of s-spares signals, which in turn implies that s-sparse vectors cannot be in the null space of Φ. This is useful as, otherwise, there would be no hope in reconstructing these vectors. An

equivalent description of the RIP is to say that all subsets of s columns of Φ are in fact nearly orthogonal.

It holds that as soon as $\delta_{2s} < 1$, problem (5.39) has a unique s-sparse solution. Unfortunately, computing the solution is a hard combinatorial problem, for the resolution of (5.39) is numerically unstable and requires an exhaustive enumeration of all the $\begin{pmatrix} N \\ s \end{pmatrix}$ possible locations of the non-zero entries of $\mathbf{c}$.

Surprisingly, an optimization problem based on the l_1 norm: find $\mathbf{c}^* \in \mathbb{R}^N$ such that

$$\mathbf{c}^* \in \underset{\widetilde{\mathbf{c}} \in \mathbb{R}^N}{\operatorname{argmin}} \|\widetilde{\mathbf{c}}\|_{l_1} \quad \text{subject to} \quad \Phi\widetilde{\mathbf{c}} = \mathbf{y}. \tag{5.41}$$

can recover exactly s-sparse signals and closely approximate compressible signals provided that δ_{2s} is small enough. More precisely [48],

Theorem 5.5.4. *Assume that $\delta_{2s} < \sqrt{2}-1$. Then, the solution $\boldsymbol{c}^*$ of problem (5.41) satisfies for some constant C_0,*

$$\|\boldsymbol{c}^* - \boldsymbol{c}\|_{l_1} \leq C_0 \|\boldsymbol{c} - \boldsymbol{c}_s\|_{l_1},$$

and

$$\|\boldsymbol{c}^* - \boldsymbol{c}\|_{l_2} \leq C_0 s^{-1/2} \|\boldsymbol{c} - \boldsymbol{c}_s\|_{l_1},$$

where $\boldsymbol{c}$ denotes the solution of (5.40) and $\boldsymbol{c}_s$ the vector $\boldsymbol{c}$ with all but the s largest entries set to zero.

The above theorem implies that, if the sensing matrix A satisfies the RIP condition, and if $\mathbf{c}$ is s-sparse, then the recovery by problem (5.41) is exact.

This is good news since problem (5.41) is convex and can be solved using linear programming via an interior point method [45, 76]. These algorithms, known as *Basis Pursuit*, thus provide a very accurate approximation of a sparse signal x, even in the case when data is corrupted with noise, but their main drawback is that they are very slow to converge.

Other numerical methods to solve (5.39) are based on greedy algorithms and were developed from the original work of Tropp and Gilbert [176] on the so-called *Orthogonal Matching Pursuit* algorithm. Different improved versions of this algorithm have been developed later, such as the *stagewise Orthogonal Matching Pursuit* [77] or the *regularized Orthogonal Pursuit* algorithm [152]. All these algorithms run faster than traditional algorithms based on the convex optimization problem (5.41), but they do not guarantee error bounds on the approximation of the solution as good as those guaranteed by Theorem 5.5.4.

However, two recent and similar greedy algorithms, the *Compressive Sampling Matching Pursuit* (CoSaMP) [151] and the *Subspace Pursuit* [63] algorithms were proved to guarantee error bounds as accurate as those given by convex optimization methods, while keeping the good running time properties of the greedy algorithms. For example, the CoSaMP algorithm runs in a time that is $\mathcal{O}(N \log N \log s)$ where s is the desired sparsity of the reconstructed signal.

This approach, proposed very recently by Doostan [78] in the context of parametrized differential equations, seems to be a promising non-intrusive algorithm to deal with high-dimensional problems.

Chapter 6

Convergence of a greedy algorithm for high-dimensional convex nonlinear problems

The results of this chapter were gathered in an article published in *Mathematical Models and Methods in Applied Sciences*. We prove that greedy algorithms can be used for the minimization of nonlinear strongly convex energy functionals and can thus offer interesting perspectives for high-dimensional nonlinear problems. Besides, the rate of convergence of the methods is exponential in the finite-dimensional case. We illustrate these convergence results on the resolution of an obstacle problem with uncertainty, using a penalized formulation.

Convergence of a greedy algorithm for high-dimensional convex nonlinear problems

Eric Cancès[1] Virginie Ehrlacher[1] Tony Lelièvre[1]

Abstract

In this article, we present a greedy algorithm based on a tensor product decomposition, whose aim is to compute the global minimum of a strongly convex energy functional. We prove the convergence of our method provided that the gradient of the energy is Lipschitz on bounded sets. The main interest of this method is that it can be used for high-dimensional nonlinear convex problems. We illustrate this method on a prototypical example for uncertainty propagation on the obstacle problem.

6.1 Introduction

The main motivation for this work comes from two important and challenging problems in contemporary scientific computing:

- the uncertainty quantification for some nonlinear models in mechanics, and more precisely, for contact problems;
- the computation of some high-dimensional functions in molecular dynamics.

Concerning the first domain of application which is the main focus of this work, there is now a wide literature on the subject, ranging from specific questions related to the modeling of the noise sources (in particular of their correlation), to dedicated methods for the study of events with very small probabilities (reliability). The focus of this paper is rather on the development of methods to compute efficiently a *reduced model* which rapidly gives the output of interest as a function of the random variable which enters the input parameters, in the context of contact problems in continuum mechanics. Such a model can then be used to evaluate the distribution of the outputs (for a given distribution of the input parameters), or to reduce the variance in a Monte Carlo computation for example. Many methods have been proposed in the literature to attack this problem [154, 92]: stochastic collocation methods, Galerkin methods, perturbation methods, etc. To be more specific, let us assume that the noise on the parameters of the model can be modeled by a possibly large number of random variables $T = (T_1, \dots, T_p) \in \mathbb{R}^p$, so that the quantity of interest (say the displacement field) $u(t, x)$ is a function of $(p + d)$ variables, where d is the dimension of the physical space. The question is then how to approximate a function on

[1]Université Paris Est, CERMICS, Projet MICMAC, Ecole des Ponts ParisTech - INRIA, 6 & 8 avenue Blaise Pascal, 77455 Marne-la-Vallée Cedex 2, France, (cances@cermics.enpc.fr, ehrlachv@cermics.enpc.fr, lelievre@cermics.enpc.fr)

such a high-dimensional space. The natural idea at the basis of many methods is to look for the solution to this problem as a linear combination of tensor products:

$$u(t,x) = \sum_{i=1}^{l} \sum_{j=1}^{m} U^{ij}\, \phi_i(t)\psi_j(x),$$

where $(\phi_i)_{1\leq i\leq l}$ and $(\psi_j)_{1\leq j\leq m}$ are bases of vector spaces of dimension l and m respectively which are fixed a priori, and where $(U^{ij})_{1\leq i\leq l, 1\leq j\leq m}$ are real numbers to be computed. This method leads to the resolution of a problem in a vector space of dimension $N = lm$ which may be very large. This difficulty becomes all the more pregnant if p is really large, so that the solution should be typically approximated as a sum:

$$u(t,x) = \sum_{i_1=1}^{l} ... \sum_{i_p=1}^{l} \sum_{j=1}^{m} U^{i_1,...,i_p,j}\, \phi_{i_1}^1(t_1) \dots \phi_{i_p}^p(t_p)\psi_j(x). \tag{6.1}$$

In this case, $N = l^p m$ will be too large for a classical discretization method. The method we are studying is a way to circumvent this difficulty.

The second application we have in mind is the computation of the solution to a high-dimensional Poisson equation arising in molecular dynamics, called the committor function [81]. Mathematically, this function gives the probability for a stochastic process to reach a given region (say $A \subset \mathbb{R}^d$) before another one (say $B \subset \mathbb{R}^d$). Using Feynman-Kac formula, it can be shown that this function satisfies a Poisson equation in a weighted Sobolev space, with Dirichlet boundary conditions (namely 1 on A and 0 on B). Typically, the stochastic process lives in a high-dimensional space (d is large), so that computing this function is a challenge.

In both cases, the difficulty comes from the high-dimensionality of the function to approximate. The principle of the method we are interested in is: (i) to rewrite the original problem as a minimization problem:

$$u \in \operatorname*{argmin}_{v\in V} \mathcal{E}(v) \tag{6.2}$$

where $\mathcal{E}$ is a functional defined on some Hilbert space V and (ii) to expand the solution in tensor products of lower-dimensional functions

$$u_n(t,x) = \sum_{k=1}^{n} r_k(t)s_k(x). \tag{6.3}$$

In practice, for each k, the functions r_k and s_k are computed as linear combinations of the functions of the bases $(\phi_i)_{1\leq i\leq l}$ and $(\psi_j)_{1\leq j\leq m}$ so that

$$r_k(t) = \sum_{i=1}^{l} R_k^i \phi_i(t), \tag{6.4}$$

and

$$s_k(x) = \sum_{j=1}^{m} S_k^j \psi_j(x), \tag{6.5}$$

where for each $k \in \mathbb{N}^*$, $R_k = (R_k^i)_{1\leq i\leq l} \in \mathbb{R}^l$ and $S_k = (S_k^j)_{1\leq j\leq m} \in \mathbb{R}^m$. In the end, computing the approximation (6.3) leads to a problem of dimension $\widetilde{N} = n(l+m)$ which, provided that n remains small enough, will hopefully be lower than the dimension of the

problem obtained with the classical approach $N = lm$ when the size of the bases l and m are large.

The reduction of dimension is even more significant when we are in the case of equation (6.1). Indeed, the approximation (6.3) can be adapted in this case in the following form:

$$u_n(t,x) = \sum_{k=1}^{n} r_k^1(t_1) \cdots r_k^p(t_p) s_k(x).$$

In this case, the overall dimension of the problem will be $\widetilde{N} = n(pl+m)$ instead of $N = l^p m$ in the classical approach.

Such a representation of a function as a sum of tensor products to avoid the curse of dimensionality has already been introduced in the literature. One approach consists in using the so-called sparse tensor product representation [169, 179, 32]. If the solution u we wish to approximate is sufficiently regular, one does not need to use fine discretizations in each direction. This idea can be used for example in Galerkin-like discretizations. However, this method loses its efficiency in the case when the solution u is not regular enough or when the mesh considered is complicated.

We adopt another approach in this article. The principle of our method is to determine *sequentially* the pairs of functions (r_k, s_k) which intervene in the approximation (6.3) through the following minimization problem:

$$(r_n, s_n) \in \underset{(r,s)\in V_t \times V_x}{\operatorname{argmin}} \mathcal{E}\left(\sum_{k=1}^{n-1} r_k(t)s_k(x) + r(t)s(x)\right), \tag{6.6}$$

where V_t and V_x in (6.6) denote respectively Hilbert spaces of functions depending only on the variable t or only on the variable x.

To rewrite the two problems mentioned above as minimization problems on Hilbert spaces, we penalize the constraints, namely the presence of the obstacle for the contact problem, or the Dirichlet boundary conditions for the high-dimensional Poisson problem.

The method described above has been introduced by Chinesta [5] for solving high-dimensional Fokker-Planck equations, by Nouy [157] in the context of uncertainty quantification in mechanics, and is very much related to so-called greedy algorithms [174, 130] used in nonlinear approximation theory. The main contributions of this work are the following:

- the convergence of the greedy algorithm (6.3)-(6.6) to the unique solution of (6.2) is proved, under the key assumptions that $\mathcal{E}$ is strongly convex and that the gradient of $\mathcal{E}$ is Lispchitz on bounded sets;
- an exponential rate of convergence is obtained in the finite dimensional case;
- an adequate procedure to solve the minimization subproblem (6.6) is proposed and tested on an academic test case.

This paper can be seen as an extension of previous works on greedy algorithms [174, 130] which concentrate on the linear case, namely when $\mathcal{E}(v) = \frac{1}{2}\|v\|_V^2 - L(v)$, where $\|\cdot\|_V$ is the norm of the Hilbert space V, and L a continuous linear form on V.

We would like to stress that even if all the results and proofs are provided in the context of tensor products of two functions, our results can be easily generalized to the case of tensor products of more than two functions such as (6.1) *except for the results in*

Section 6.5. We have chosen not to present the results in this general setting for the sake of clarity.

The paper is organized as follows. In Section 6.2, we introduce the general setting for the problem we consider and state the main result of this paper, namely the convergence of the greedy algorithm. Section 6.2 also presents more precisely the two specific examples of application we have in mind. Section 6.3 is devoted to the proof of the convergence. In Section 6.4, an exponential rate of convergence is proved, in the finite dimensional setting (i.e. when V_t and V_x are finite dimensional spaces). Section 6.5 shows that, under specific additional assumptions which are typically satisfied in the context of uncertainty quantification, the convergence results also hold if (r_n, s_n) in (6.6) is only a local minimum. Finally, Section 6.6 is devoted to a discussion of the numerical implementation, as well as to the presentation of test cases on a toy model.

6.2 Presentation of the problem and the convergence result

In this paper, we are interested in the convergence of a greedy algorithm for the minimization of high-dimensional nonlinear convex problems.

We first introduce the general theoretical setting in which we prove the convergence, then describe two prototypical examples to which our analysis can be applied.

6.2.1 General theoretical setting

Throughout this article, p and d denote some positive integers, and $\mathcal{T}$ and $\mathcal{X}$ some open sets of $\mathbb{R}^p$ and $\mathbb{R}^d$ respectively.

Let V_t and V_x be Hilbert spaces of real-valued functions respectively defined over $\mathcal{T}$ and $\mathcal{X}$ (typically L^2 or Sobolev spaces). Let $\|.\|_t$ and $\|.\|_x$ be the norms of V_t and V_x.

We introduce the following tensor product for all $(r, s) \in V_t \times V_x$,

$$r \otimes s : \left\{ \begin{array}{ccc} \mathcal{T} \times \mathcal{X} & \to & \mathbb{R} \\ (t, x) & \mapsto & r(t)s(x) \end{array} \right. , \tag{6.7}$$

which defines a real-valued function on $\mathcal{T} \times \mathcal{X}$.

We also denote by $\Sigma = \{r \otimes s \mid (r, s) \in V_t \times V_x\}$.

Let V be a Hilbert space of real-valued functions defined on $\mathcal{T} \times \mathcal{X}$. The scalar product of V is denoted by $\langle ., . \rangle$ and the associated norm by $\|.\|_V$.

Let $\mathcal{E}$ be a differentiable real-valued functional defined on V. For all $v \in V$, we denote by $\mathcal{E}'(v)$ the gradient of $\mathcal{E}$ at v.

We make the following assumptions:

($A1$) Span(Σ) is a dense subset of V for $\|.\|_V$;

($A2$) for all sequences of Σ bounded in V, there exists a subsequence which weakly converges in V towards an element of Σ;

($A3$) the functional $\mathcal{E}$ is strongly convex for $\|.\|_V$, i.e. there exists a constant $\alpha \in \mathbb{R}_+^*$ for which

$$\forall v, w \in V,\ \mathcal{E}(v) \geq \mathcal{E}(w) + \langle \mathcal{E}'(w), v - w \rangle + \frac{\alpha}{2} \|v - w\|_V^2. \tag{6.8}$$

The functional $\mathcal{E}$ is also said to be α-convex;

(*A*4) the gradient of $\mathcal{E}$ is Lipschitz on bounded sets: for each bounded subset K of V, there exists a nonnegative constant $L_K \in \mathbb{R}_+$ such that

$$\forall v, w \in V, \quad \|\mathcal{E}'(v) - \mathcal{E}'(w)\|_V \leq L_K \|v - w\|_V. \tag{6.9}$$

The unique global minimizer of $\mathcal{E}$ on V is denoted by u. Its existence and uniqueness are ensured by the α-convexity of the functional $\mathcal{E}$:

$$u = \operatorname*{argmin}_{v \in V} \mathcal{E}(v).$$

We are going to study the following algorithm: the sequence $((r_n, s_n))_{n \in \mathbb{N}^*} \in (V_t \times V_x)^{\mathbb{N}^*}$ is defined recursively by

$$(r_n, s_n) \in \operatorname*{argmin}_{(r,s) \in V_t \times V_x} \mathcal{E}\left(\sum_{k=1}^{n-1} r_k \otimes s_k + r \otimes s\right). \tag{6.10}$$

Throughout this article, we will denote for all $n \in \mathbb{N}^*$,

$$u_n = \sum_{k=1}^{n} r_k \otimes s_k. \tag{6.11}$$

Our main result is the following theorem, whose proof is given in Section 6.3.

Theorem 6.2.1. *Under the assumptions* (*A*1), (*A*2), (*A*3) *and* (*A*4), *the iterations of the algorithm are well-defined, in the sense that (6.10) has at least one minimizer* (r_n, s_n). *Moreover, the sequence* $(u_n)_{n \in \mathbb{N}}$ *strongly converges in* V *towards* u.

Remark 6.2.1. For each $n \in \mathbb{N}^*$, the minimizer of (6.10) is not unique in general. In particular, notice that the function $V_t \times V_x \ni (r, s) \mapsto \mathcal{E}\left(\sum_{k=1}^{n-1} r_k \otimes s_k + r \otimes s\right)$ is not convex.

Remark 6.2.2. Theorem 6.2.1 could be generalized to the case of tensor products of more than two Hilbert spaces.

Indeed, let $q \in \mathbb{N}$ with $q \geq 3$. Let $p_1, \cdots, p_q$ be q positive integers. Let $\mathcal{T}_1, \cdots, \mathcal{T}_q$ be q open subsets of $\mathbb{R}^{p_1}, \cdots, \mathbb{R}^{p_q}$ respectively. We consider q Hilbert spaces, $V_1, \cdots, V_q$ of real-valued functions defined respectively on $\mathcal{T}_1, \cdots, \mathcal{T}_q$. Let V be a Hilbert space of real-valued functions defined on $\mathcal{T}_1 \times \cdots \times \mathcal{T}_q$. Let $\mathcal{E}$ be a real-valued differentiable functional defined on V. We denote by $\Sigma = \left\{r^{(1)} \otimes \cdots \otimes r^{(q)} \mid \left(r^{(1)}, \cdots, r^{(q)}\right) \in V_1 \times \cdots \times V_q\right\}$. Our algorithm can then easily be adapted provided that assumptions (*A*1), (*A*2), (*A*3) and (*A*4) are satisfied: $\left(r_n^{(1)}, \cdots, r_n^{(q)}\right) \in V_1 \times \cdots \times V_q$ are defined recursively by

$$\left(r_n^{(1)}, \cdots, r_n^{(q)}\right) \in \operatorname*{argmin}_{\left(r^{(1)}, \cdots, r^{(q)}\right) \in V_1 \times \cdots \times V_q} \mathcal{E}\left(\sum_{k=1}^{n-1} r_k^{(1)} \otimes \cdots \otimes r_k^{(q)} + r^{(1)} \otimes \cdots \otimes r^{(q)}\right).$$

Our convergence result also holds in this case. But for the sake of simplicity, we will limit our analysis to the case of only two Hilbert spaces.

Remark 6.2.3. Let $(.,.)$ be the scalar product defined on $\mathrm{Span}(\Sigma)$ as: for all $(r_1, r_2, s_1, s_2) \in V_t^2 \times V_x^2$,

$$(r_1 \otimes s_1, r_2 \otimes s_2) = \langle r_1, r_2 \rangle_{V_t} \langle s_1, s_2 \rangle_{V_x},$$

where $\langle ., . \rangle_{V_t}$ and $\langle ., . \rangle_{V_x}$ denote the scalar products of V_t and V_x respectively. Let $\|.\|$ be the cross-norm associated to the scalar product $(.,.)$. The tensor space of V_t and V_x, denoted as $V_t \otimes V_x$ is then defined as the closure of $\mathrm{Span}(\Sigma)$ for the product norm $\|.\|$,

$$V_t \otimes V_x = \overline{\mathrm{Span}(\Sigma)}^{\|.\|}.$$

Let us point out that the Hilbert space V is not necessarily equal to $V_t \otimes V_x$, the tensor space of V_t and V_x associated to the tensor product (6.7). Indeed, an example where our analysis can be applied and where $V \neq V_t \otimes V_x$ is given in Section 6.2.2 (see Remark 6.2.5.). However, the following inclusion relationship holds: $V_t \otimes V_x \subset V$.

Remark 6.2.4. If V_t and V_x are discretized in finite-dimensional spaces of dimension l and m, our algorithm consists in solving several problems in dimension $l + m$ instead of solving one problem of dimension lm. Thus, the method can be implementable even for very high-dimensional problems, contrarily to classical approximation methods.

6.2.2 Prototypical problems

To prove that the general theoretical setting we described in Section 6.2.1 is satisfied on the prototypical problems we present in this section, we need the following lemma, which is well-known in distribution theory [165].

Lemma 6.2.1. *Let $U \in \mathcal{D}'(\mathcal{T} \times \mathcal{X})$ be a distribution such that for any functions $(\phi, \psi) \in \mathcal{C}_c^\infty(\mathcal{T}) \times \mathcal{C}_c^\infty(\mathcal{X})$,*

$$(U, \phi \otimes \psi)_{(\mathcal{D}'(\mathcal{T}\times\mathcal{X}), \mathcal{D}(\mathcal{T}\times\mathcal{X}))} = 0.$$

Then $U = 0$ in $\mathcal{D}'(\mathcal{T} \times \mathcal{X})$. Moreover, for any two sequences of distributions $R_n \in \mathcal{D}'(\mathcal{T})$ and $S_n \in \mathcal{D}'(\mathcal{X})$ such that $\lim_{n\to\infty} R_n = R$ in $\mathcal{D}'(\mathcal{T})$ and $\lim_{n\to\infty} S_n = S$ in $\mathcal{D}'(\mathcal{X})$, $\lim_{n\to\infty} R_n \otimes S_n = R \otimes S$ in $\mathcal{D}'(\mathcal{T} \times \mathcal{X})$.

Uncertainty propagation on obstacle problems

An example of application of our algorithm is the study of uncertainty propagation on obstacle problems. We assume that uncertainty can be modeled by a set of p random variables T_1, T_2, ..., T_p, and that the random vector $T = (T_1, ..., T_p)$ takes its values in $\mathcal{T}$.

We consider also that the physical problem is defined over the domain $\mathcal{X}$, which is supposed to be a bounded subset of $\mathbb{R}^d$. If H is a Hilbert space of functions defined on $\mathcal{X}$, we denote by

$$L_T^2(\mathcal{T}, H) = \left\{ v : \mathcal{T} \to H \mid \mathbb{E}\left[\|v(T)\|_H^2 \right] < +\infty \right\},$$

where $\mathbb{E}$ denotes the expectation with respect to the probability law of T, and $\|.\|_H$ denotes the norm of H. We endow $L_T^2(\mathcal{T}, H)$ with the scalar product defined by $\langle v, w \rangle_{L_T^2(\mathcal{T},H)} = \mathbb{E}\left[\langle v(T), w(T) \rangle_H \right]$.

A formulation of the obstacle problem with uncertainty is the following [101]. Let $g \in L_T^2(\mathcal{T}, H_0^1(\mathcal{X}))$ and $f \in L_T^2(\mathcal{T}, H^{-1}(\mathcal{X}))$. A membrane is stretched over the domain $\mathcal{X}$ and is deflected by some random force having pointwise density $f(T, x)$ for $x \in \mathcal{X}$. At the boundary $\partial\mathcal{X}$, the membrane is fixed and in the interior of $\mathcal{X}$ the deflection is assumed to

be bounded from below by the function $g(T,x)$ (a random obstacle). Then the deflection $z = z(T,x)$ is solution of the following obstacle problem with uncertainty (see Fig. 6.1):

$$\left\{ \begin{array}{c} -\Delta_x z(t,x) \geq f(t,x) \\ z(t,x) \geq g(x,t) \\ (\Delta_x z(t,x) + f(t,x))\,(z(t,x) - g(t,x)) = 0 \end{array} \right\} \quad \begin{array}{l} \text{for a.a. } (t,x) \in \mathcal{T} \times \mathcal{X}, \\ \\ \text{for a.a. } (t,x) \in \mathcal{T} \times \partial\mathcal{X}. \end{array} \tag{6.12}$$

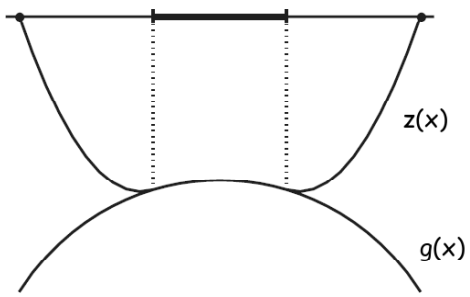

Figure 6.1: Obstacle problem.

An equivalent formulation of this problem is the following. Let us denote

$$\mathcal{K}_g = \left\{ v \in L^2_T(\mathcal{T}, H^1_0(\mathcal{X})) \mid \text{for a.a. } (t,x) \in \mathcal{T} \times \mathcal{X},\ v(t,x) \geq g(t,x) \right\}.$$

Solving the obstacle problem (6.12) consists in solving the minimization problem

$$\inf_{v \in \mathcal{K}_g} \mathcal{J}(v), \tag{6.13}$$

where $\mathcal{J}(v) = \mathbb{E}\left[\frac{1}{2} \int_{\mathcal{X}} |\nabla_x v(T,x)|^2 dx - \langle f(T,.), v(T,.) \rangle_{H^{-1}(\mathcal{X}), H^1_0(\mathcal{X})} \right]$.

One of the main difficulties of this kind of problems is their very high nonlinearity. Many methods have been proposed to approximate the solution of these problems in the case without uncertainty [95, 90, 94, 101]. Among them, penalization methods [101, 95] are among the most widely used. They consist in approximating the solution of a given obstacle problem by a sequence of solutions of penalized problems defined on the entire Hilbert space.

Let ρ be a parameter in $\mathbb{R}_+$. Such a penalized problem associated with problem (6.13) may be defined as

$$\inf_{v \in L^2_T(\mathcal{T}, H^1_0(\mathcal{X}))} \mathcal{J}_\rho(v), \tag{6.14}$$

where $\mathcal{J}_\rho(v) = \mathcal{J}(v) + \mathbb{E}\left[\frac{\rho}{2} \int_{\mathcal{X}} [g(T,x) - v(T,x)]^2_+ dx \right]$.

Here and below, we denote by $[a]_+$ the positive part of the real number a, i.e. $[a]_+ = 0$ if $a \leq 0$ and $[a]_+ = a$ if $a \geq 0$.

When ρ goes to infinity, the solution z_ρ of problem (6.14) strongly converges to the solution z of problem (6.13). The goal of the algorithm we described in the previous section is to calculate the solution $u = z_\rho$ of this regularized problem for a given value of the parameter ρ.

Let us check that the general theoretical setting we described in Section 6.2.1 can be applied in this case.

Let us consider $V = L^2_T(\mathcal{T}, H^1_0(\mathcal{X}))$, $V_t = L^2_T(\mathcal{T}, \mathbb{R})$, $V_x = H^1_0(\mathcal{X})$ and $\mathcal{E}(v) = \mathcal{J}_\rho(v)$ for $v \in V$. We have $\Sigma = \{r \otimes s \mid (r,s) \in V_t \times V_x\}$. We endow $H^1_0(\mathcal{X})$ with the scalar product defined by $\langle s_1, s_2 \rangle_{H^1_0(\mathcal{X})} = \int_{\mathcal{X}} \nabla s_1(x).\nabla s_2(x) dx$. In this case, we have $V = V_t \otimes V_x$ and as a consequence assumption $(A1)$ is obviously satisfied.

Besides, assumption $(A2)$ is satisfied as well. If $((r_n, s_n))_{n\in\mathbb{N}} \in (V_t \times V_x)^{\mathbb{N}}$ is such that $(\|r_n \otimes s_n\|_V)_{n\in\mathbb{N}}$ is bounded, it is possible to extract a subsequence which weakly converges in V towards an element $w \in V$. Besides, there exists a non-negative constant $C \in \mathbb{R}_+$ such that for all $n \in \mathbb{N}$,

$$\begin{aligned}
\|r_n \otimes s_n\|^2_V &= \mathbb{E}\left[\int_{\mathcal{X}} |\nabla_x (r_n \otimes s_n)(T,x)|^2 dx\right] \\
&= \mathbb{E}\left[|r_n(T)|^2\right] \int_{\mathcal{X}} |\nabla_x s_n(x)|^2 dx \\
&= \|r_n\|^2_{V_t} \|s_n\|^2_{V_x} \\
&\leq C.
\end{aligned}$$

We can then choose $((r_n^*, s_n^*))_{n\in\mathbb{N}} \in (V_t \times V_x)^{\mathbb{N}}$ such that $r_n^* \otimes s_n^* = r_n \otimes s_n$ and $\|r_n^*\|_{L^2_T(\mathcal{T},\mathbb{R})} = 1$. The sequences $(r_n^*)_{n\in\mathbb{N}}$ and $(s_n^*)_{n\in\mathbb{N}}$ are then bounded in $L^2_T(\mathcal{T},\mathbb{R})$ and $H^1_0(\mathcal{X})$ respectively and we can extract subsequences which weakly converge in $L^2_T(\mathcal{T},\mathbb{R})$ and $H^1_0(\mathcal{X})$ towards $r^\infty \in L^2_T(\mathcal{T},\mathbb{R})$ and $s^\infty \in H^1_0(\mathcal{X})$ respectively. As the weak convergences in $L^2_T(\mathcal{T},\mathbb{R})$ and $H^1_0(\mathcal{X})$ imply the convergences in the distributional sense, the sequence $r_n^* \otimes s_n^* = r_n \otimes s_n$ necessarily converges towards $r^\infty \otimes s^\infty$ in $\mathcal{D}'(\mathcal{T}\times\mathcal{X})$ by Lemma 6.2.1. As the weak convergence in V also implies the convergence in the sense of the distributions, we obtain, by uniqueness of the limit, $w = r^\infty \otimes s^\infty \in \Sigma$. Hence assumption $(A2)$ is satisfied.

The functional $\mathcal{E}$ is differentiable and 1-convex. Indeed, for all $v \in V$,

$$\mathcal{E}(v) = \frac{1}{2}\|v\|^2_V + \left(\mathbb{E}\left[\langle f(T,.), v(T,.)\rangle_{H^{-1}(\mathcal{X}),H^1_0(\mathcal{X})} + \frac{\rho}{2}\int_{\mathcal{X}} [g(T,x) - v(T,x)]^2_+ dx\right]\right),$$

is the sum of a 1-convex function ($V \ni v \mapsto \frac{1}{2}\|v\|^2_V$) and of a convex function ($V \ni v \mapsto \mathbb{E}\left[\langle f(T,.), v(T,.)\rangle_{H^{-1}(\mathcal{X}),H^1_0(\mathcal{X})} + \frac{\rho}{2}\int_{\mathcal{X}} [g(T,x) - v(T,x)]^2_+ dx\right]$). The functional $\mathcal{E}$ therefore obeys property (6.8) with $\alpha = 1$. Hence, assumption $(A3)$ is satisfied.

Let us finally check that the gradient of $\mathcal{E}$ is Lipschitz. For all $v, w, y \in V$,

$$\begin{aligned}
\left|\langle \mathcal{E}'(v) - \mathcal{E}'(w), y\rangle\right| &\leq \left|\mathbb{E}\left[\int_{\mathcal{X}} \nabla_x(v(T,x) - w(T,x)).\nabla_x y(T,x) dx\right]\right| \\
&\quad + \rho\left|\mathbb{E}\left[\int_{\mathcal{X}} ([g(T,x) - v(T,x)]_+ - [g(T,x) - w(T,x)]_+) y(T,x) dx\right]\right| \\
&\leq \|v - w\|_V \|y\|_V \\
&\quad + \rho\mathbb{E}\left[\int_{\mathcal{X}} \left|[g(T,x) - v(T,x)]_+ - [g(T,x) - w(T,x)]_+\right| |y(T,x)| \, dx\right].
\end{aligned}$$

For $a, b \in \mathbb{R}$, it can easily be seen that $|[a]_+ - [b]_+| \leq |a - b|$. This implies

$$\begin{aligned}\left|\langle \mathcal{E}'(v) - \mathcal{E}'(w), y\rangle\right| &\leq \|v - w\|_V \|y\|_V + \rho \mathbb{E}\left[\int_{\mathcal{X}} |v(T,x) - w(T,x)|\, |y(T,x)| dx\right] \\ &\leq \|v - w\|_V \|y\|_V \\ &\quad + \rho \left(\mathbb{E}\left[\int_{\mathcal{X}} |v(T,x) - w(T,x)]|^2 dx\right]\right)^{1/2} \left(\mathbb{E}\left[\int_{\mathcal{X}} |y(T,x)]|^2 dx\right]\right)^{1/2}.\end{aligned}$$

The Poincaré inegality in $H_0^1(\mathcal{X})$ implies that there exists a nonnegative constant $D \in \mathbb{R}_+$ such that for all $h \in V$,

$$\left|\mathbb{E}\left[\int_{\mathcal{X}} |h(T,x)]|^2 dx\right]\right|^{1/2} \leq D\|h\|_V.$$

This yields

$$\left|\langle \mathcal{E}'(v) - \mathcal{E}'(w), y\rangle\right| \leq (1 + \rho D^2)\|v - w\|_V \|y\|_V,$$

hence,

$$\|\mathcal{E}'(v) - \mathcal{E}'(w)\|_V \leq (1 + \rho D^2)\|v - w\|_V.$$

The functional $\mathcal{E}$ then obeys property (6.9) with a constant $L = 1 + \rho D^2$ independent of the bounded set considered.

Thus, our obstacle problem (6.14) falls into the general theoretical setting introduced in Section 6.2.1.

There exist several variants of the obstacle problem which could be tackled with our algorithm. We refer to [101] or [94] for such examples.

High-dimensional Poisson equation

Our algorithm may also be used to calculate the solution of other problems than obstacle problems. Other examples are high-dimensional nonlinear Poisson equations. A specific application where such high dimensional Poisson equations arise is the calculation of the so-called committor function in molecular dynamics [81], which is an important quantity to compute reaction rates or to derive some effective dynamics for example.

Let $q \in \mathbb{N}^*$. The committor is the solution to the following problem:

$$z = \operatorname*{argmin}_{v \in W} \frac{1}{2} \int_{\mathbb{R}^q \setminus (\overline{A} \cup \overline{B})} |\nabla v(y)|^2 \exp(-U(y))\, dy$$

where q is typically large, A and B are disjoint smooth open sets of $\mathbb{R}^q$, $U : \mathbb{R}^q \to \mathbb{R}$ is a given potential function such that $\int_{\mathbb{R}^q} \exp(-U) < \infty$ and

$$W = \left\{ \begin{array}{c} v \in L^2_{loc}(\mathbb{R}^q), \\ \int_{\mathbb{R}^q \setminus (\overline{A} \cup \overline{B})} |\nabla v(y)|^2 \exp(-U(y))\, dy < \infty, \\ v = 1 \text{ on } \overline{A} \text{ and } v = 0 \text{ on } \overline{B} \end{array} \right\}.$$

For $y \in \mathbb{R}^q \setminus (\overline{A} \cup \overline{B})$, $z(y)$ can be interpreted as the probability that the stochastic process Q_t^y solution to

$$Q_t^y = y - \int_0^t \nabla U(Q_s^y)\, ds + \sqrt{2}\, W_t$$

reaches $\overline{A}$ before $\overline{B}$. Here, W_t denotes a q-dimensional Brownian motion.

Let $p, d \in \mathbb{N}^*$ such that $q = p + d$. In this example, we consider the case when $C = \mathbb{R}^q \setminus \left(\overline{A} \cup \overline{B}\right)$ is bounded. Let $\mathcal{T}$ and $\mathcal{X}$ be open convex bounded subsets of $\mathbb{R}^p$ and $\mathbb{R}^d$ respectively such that $\overline{C} \subset \Omega := \mathcal{T} \times \mathcal{X}$ and such that $\mu\left((A \cup B) \cap \Omega\right) \neq 0$ where μ denotes the Lebesgue measure. We also assume that $U \in \mathcal{C}^\infty(\mathbb{R}^q)$. In this case, the initial problem can be rewritten as a minimization problem set on

$$\widetilde{W} = \left\{v \in H^1(\mathcal{T} \times \mathcal{X}) |\, v = 1 \text{ on } A \cap \Omega \text{ and } v = 0 \text{ on } B \cap \Omega\right\},$$

instead of W. Indeed, as $U \in \mathcal{C}^\infty(\mathbb{R}^q, \mathbb{R})$ and Ω is bounded, there exists constants $\gamma, \kappa > 0$ such that for all $y \in \Omega$, $\gamma \leq \exp(-U(y)) \leq \kappa$. And thus, we have $v \in W$ if and only if $v|_\Omega \in \widetilde{W}$, $v|_{\overline{A}\setminus\Omega} = 1$ and $v|_{\overline{B}\setminus\Omega} = 0$.

The penalized version of the committor problem then reads

$$u = \underset{v \in H^1(\mathcal{T}\times\mathcal{X})}{\operatorname{argmin}} \mathcal{E}(v), \tag{6.15}$$

where

$$\mathcal{E}(v) = \frac{1}{2}\int_\Omega |\nabla v(y)|^2 \exp(-U(y))\, dy + \frac{\rho}{2}\left(\int_{A\cap\Omega} |v(y) - 1|^2\, dy + \int_{B\cap\Omega} |v(y)|^2\, dy\right),$$

for some $\rho > 0$.

Let us check that the general theoretical setting described in Section 6.2.1 is relevant for this problem.

In this case, we consider $V = H^1(\mathcal{T} \times \mathcal{X})$, $V_t = H^1(\mathcal{T})$ and $V_x = H^1(\mathcal{X})$. The inner products that are defined over these Hilbert spaces are the following. For all $v_1, v_2 \in V$, $r_1, r_2 \in V_t$, $s_1, s_2 \in V_x$,

$$\langle v_1, v_2\rangle_V = \int_\mathcal{T}\int_\mathcal{X} \left(\nabla v_1(t,x).\nabla v_2(t,x) + v_1(t,x)v_2(t,x)\right)\, dt\, dx,$$

$$\langle r_1, r_2\rangle_{V_t} = \int_\mathcal{T} \left(\nabla r_1(t).\nabla r_2(t) + r_1(t)r_2(t)\right)\, dt,$$

$$\langle s_1, s_2\rangle_{V_x} = \int_\mathcal{X} \left(\nabla s_1(x).\nabla s_2(x) + s_1(x)s_2(x)\right)\, dx.$$

Remark 6.2.5. *Let us point out that in this case, $V \neq V_t \otimes V_x$. Indeed, for all $(r, s) \in V_t \times V_x$, the V-norm of the tensor product $r \otimes s$ reads*

$$\|r \otimes s\|_V^2 = \|r\|_{L^2(\mathcal{T})}^2 \|\nabla s\|_{L^2(\mathcal{X})}^2 + \|\nabla r\|_{L^2(\mathcal{T})}^2 \|s\|_{L^2(\mathcal{X})}^2 + \|r\|_{L^2(\mathcal{T})}^2 \|s\|_{L^2(\mathcal{X})}^2,$$

which is not a cross-norm, equivalent to the norm induced by $\|.\|_{V_t}$ and $\|.\|_{V_x}$ over $V_t \otimes V_x$, which is

$$\|r \otimes s\|_{V_t\otimes V_x} = \|r\|_{V_t}\|s\|_{V_x}.$$

Indeed, let us consider $\mathcal{T} = \mathcal{X} = (0,1)$, $r_l(t) = \frac{1}{l}\sin(l^2\pi t)$ and $s_l(x) = \frac{1}{l}\sin(l^2\pi x)$ for $(t,x) \in (0,1)^2$ and $l \in \mathbb{N}^*$. The sequence $\left(\|r_l \otimes s_l\|_V\right)_{n\in\mathbb{N}^*}$ is bounded, but the sequence $\left(\|r_l\|_{V_t}\|s_l\|_{V_x}\right)_{l\in\mathbb{N}^*}$ is not.

Assumption $(A1)$ holds true, since $\widetilde{\Sigma} = \left\{r \otimes s \mid (r,s) \in \mathcal{C}^\infty\left(\overline{\mathcal{T}}\right) \times \mathcal{C}^\infty\left(\overline{\mathcal{X}}\right)\right\}$ is such that $\widetilde{\Sigma} \subset \Sigma$ and $\operatorname{Span}\left(\widetilde{\Sigma}\right)$ is dense in $H^1(\mathcal{T} \times \mathcal{X})$. Hence, $\operatorname{Span}(\Sigma)$ is also dense in V.

Let us prove that assumption $(A2)$ also holds true. If $((r_n, s_n))_{n\in\mathbb{N}} \in (V_t \times V_x)^{\mathbb{N}}$ is such that $(\|r_n \otimes s_n\|_V)_{n\in\mathbb{N}}$ is bounded, we can extract a subsequence of $(r_n \otimes s_n)_{n\in\mathbb{N}^*}$ which weakly converges in V towards an element $w \in V$. Besides, there exists a nonnegative constant $C \in \mathbb{R}_+$ such that for all $n \in \mathbb{N}$,

$$\begin{aligned}\|r_n \otimes s_n\|_V^2 &= \int_{\mathcal{T}\times\mathcal{X}} \left(|\nabla r_n(t)|^2|s_n(x)|^2 + |r_n(t)|^2|\nabla s_n(x)|^2 + |r_n(t)|^2|s_n(x)|^2\right) dt\, dx \\ &= \|\nabla r_n\|_{L^2(\mathcal{T})}^2\|s_n\|_{L^2(\mathcal{X})}^2 + \|r_n\|_{L^2(\mathcal{T})}^2\|\nabla s_n\|_{L^2(\mathcal{X})}^2 + \|r_n\|_{L^2(\mathcal{T})}^2\|s_n\|_{L^2(\mathcal{X})}^2 \\ &\leq C.\end{aligned}$$

We can then choose $((r_n^*, s_n^*))_{n\in\mathbb{N}} \in (V_t \times V_x)^{\mathbb{N}}$ such that $r_n^* \otimes s_n^* = r_n \otimes s_n$ and such that $\|r_n^*\|_{L^2(\mathcal{T})} = 1$. The sequences $(r_n^*)_{n\in\mathbb{N}}$ and $(s_n^*)_{n\in\mathbb{N}}$ are then bounded in $L^2(\mathcal{T})$ and $H^1(\mathcal{X})$ and we can extract subsequences which weakly converge in $L^2(\mathcal{T})$ and $H^1(\mathcal{X})$ respectively towards r^∞ and s^∞. As the weak convergences in $L^2(\mathcal{T})$ and $H^1(\mathcal{X})$ imply the convergences in the distributional sense, $r_n^* \otimes s_n^* = r_n \otimes s_n$ necessarily converges towards $r^\infty \otimes s^\infty$ in the distributional sense by Lemma 6.2.1. As the weak convergence in V also implies the convergence in the sense of the distributions, by uniqueness of the limit, $w = r^\infty \otimes s^\infty$. Let us suppose $w \neq 0$. In that case, we have $r^\infty \neq 0$ and $s^\infty \neq 0$. Besides, we have

$$\|w\|_V^2 = \|r^\infty\|_{L^2(\mathcal{T})}^2\|\nabla s^\infty\|_{L^2(\mathcal{X})}^2 + \|r^\infty\|_{H^1(\mathcal{T})}^2\|s^\infty\|_{L^2(\mathcal{X})}^2,$$

hence

$$\|r^\infty\|_{H^1(\mathcal{T})}^2 \leq \frac{\|w\|_V^2}{\|s^\infty\|_{L^2(\mathcal{X})}^2}.$$

As a consequence $\|r^\infty\|_{H^1(\mathcal{T})}$ is finite and $r^\infty \in H^1(\mathcal{T})$. Hence $w = r^\infty \otimes s^\infty \in \Sigma$. If $w = 0$, then obviously $w \in \Sigma$. Hence, assumption $(A2)$ holds true.

The functional $\mathcal{E}$ is differentiable and strongly convex. To prove this, it is sufficient to prove that there exists a constant $\alpha \in \mathbb{R}_+^*$ such that for all $v, w \in V$, $\langle \mathcal{E}'(v) - \mathcal{E}'(w), v - w\rangle \geq \alpha\|v - w\|_V^2$. Indeed, there exists $\gamma > 0$ such that for all $y \in \mathbb{R}^q$, $\exp(-U(y)) \geq \gamma$. Thus, there exists a constant $\delta > 0$ such that, for all $v, w \in V$,

$$\langle \mathcal{E}'(v) - \mathcal{E}'(w), v - w\rangle \geq \delta \left(\int_\Omega |\nabla(v - w)|^2 + \int_{A\cap\Omega} |v - w|^2 + \int_{B\cap\Omega} |v - w|^2\right).$$

To prove that the functional $\mathcal{E}$ is strongly convex, it is sufficient to have the following inequality: there exists a constant $C_\Omega \in \mathbb{R}_+^*$ such that for all $v \in H^1(\Omega)$,

$$\int_\Omega |\nabla v|^2 + \int_{(A\cup B)\cap\Omega} |v|^2 \geq C_\Omega\|v\|_{H^1(\Omega)}^2. \tag{6.16}$$

As $\mathcal{T}$ and $\mathcal{X}$ are bounded open convex subsets of $\mathbb{R}^p$ and $\mathbb{R}^d$ respectively, Ω is then a bounded open convex subset of $\mathbb{R}^q$ such that $\mu((A \cup B) \cap \Omega) \neq 0$ and inequality (6.16) is a well-known Poincare-like inequality.

Hence, assumption $(A3)$ is satisfied.

Let us check that the gradient of $\mathcal{E}$ is Lipschitz. For all $v, w, z \in V$,

$$\begin{aligned}
\left|\langle \mathcal{E}'(v) - \mathcal{E}'(w), z\rangle\right| &\leq \left|\int_\Omega \nabla(v(t,x) - w(t,x)).\nabla z(t,x) \exp(-U(t,x))\, dt\, dx\right| \\
&\quad + \rho\left|\int_{(A\cup B)\cap\Omega} (v(t,x) - w(t,x)) z(t,x)\, dt\, dx\right| \\
&\leq \|\exp(-U)\|_{L^\infty(\Omega)} \|\nabla(v-w)\|_{L^2(\Omega)} \|\nabla z\|_{L^2(\Omega)} \\
&\quad + \rho \|v-w\|_{L^2((A\cup B)\cap\Omega)} \|z\|_{L^2((A\cup B)\cap\Omega)} \\
&\leq \|\exp(-U)\|_{L^\infty(\Omega)} \|\nabla(v-w)\|_{L^2(\Omega)} \|\nabla z\|_{L^2(\Omega)} \\
&\quad + \rho \|v-w\|_{L^2(\Omega)} \|z\|_{L^2(\Omega)} \\
&\leq \|v-w\|_V \|z\|_V (\|\exp(-U)\|_{L^\infty(\Omega)} + \rho).
\end{aligned}$$

Hence

$$\|\mathcal{E}'(v) - \mathcal{E}'(w)\|_V \leq (\|\exp(-U)\|_{L^\infty(\Omega)} + \rho)\|v-w\|_V.$$

The functional $\mathcal{E}$ therefore obeys property (6.9) with a constant $L = \|\exp(-U)\|_{L^\infty(\Omega)} + \rho$ independent of the bounded subset considered.

Thus, the committor problem falls into the general theoretical setting introduced in Section 6.2.1.

6.3 Proof of Theorem 6.2.1

6.3.1 The iterations are well-defined

We begin by proving that the iterations of the algorithm are well-defined. For this, we will need the following lemma.

Lemma 6.3.1. *Let w be a function in V. Then there exists a pair $(r,s) \in V_t \times V_x$ such that $\mathcal{E}(w + r\otimes s) < \mathcal{E}(w)$ if and only if $\mathcal{E}'(w) \neq 0$.*

Proof. Let $w \in V$ and let us suppose that $\mathcal{E}(w + r\otimes s) \geq \mathcal{E}(w)$ for all $(r,s) \in V_t \times V_x$. For a given pair (r,s), for all $\varepsilon \in \mathbb{R}$,

$$\mathcal{E}(w + \varepsilon r\otimes s) - \mathcal{E}(w) \geq 0.$$

As a consequence, we have the following by letting ε go to 0: $\langle \mathcal{E}'(w), r\otimes s\rangle = 0$. This holds for all $(r,s) \in V_t \times V_x$. Hence, for all $z \in \mathrm{Span}(\Sigma)$, we also have $\langle \mathcal{E}'(w), z\rangle = 0$, and the density of $\mathrm{Span}(\Sigma)$ in V, which is ensured by assumption $(A1)$, yields

$$\mathcal{E}'(w) = 0.$$

.

Conversely, let us assume that $\mathcal{E}'(w) = 0$. Then, as $\mathcal{E}$ is α-convex, w is necessarily the global minimizer of $\mathcal{E}$ and, in particular, we have for all $(r,s) \in V_t \times V_x$,

$$\mathcal{E}(w + r\otimes s) \geq \mathcal{E}(w).$$

This concludes the proof. □

Using this lemma, the following result can be proved:

Proposition 6.3.1. *For all $n \in \mathbb{N}^*$, there exists a solution $(r_n, s_n) \in V_t \times V_x$ to the minimization problem (6.10). Moreover, $r_n \otimes s_n \neq 0$ if and only if $u_{n-1} \neq u$, where u_n is defined by (6.11).*

Proof. Firstly, let us prove the existence of a minimizer for problem (6.10).

Let $n \in \mathbb{N}^*$. For all $(r, s) \in V_t \times V_x$, $\mathcal{E}(u_{n-1}+r\otimes s) \geq \mathcal{E}(u)$. So $m = \inf_{(r,s)\in V_t\times V_x} \mathcal{E}(u_{n-1}+ r \otimes s)$ exists in $\mathbb{R}$.

We then consider a minimizing sequence $(r^{(l)}, s^{(l)})_{l\in\mathbb{N}} \in (V_t \times V_x)^{\mathbb{N}}$ such that

$$\lim_{l\to\infty} \mathcal{E}(u_{n-1} + r^{(l)} \otimes s^{(l)}) = m.$$

Using (6.8) and the fact that $\mathcal{E}'(u) = 0$, we have

$$\mathcal{E}(u_{n-1} + r^{(l)} \otimes s^{(l)}) - \mathcal{E}(u) \geq \frac{\alpha}{2}||u_{n-1} + r^{(l)} \otimes s^{(l)} - u||_V^2.$$

Then the sequence $(r^{(l)} \otimes s^{(l)})_{l\in\mathbb{N}}$ is bounded in V because $(\mathcal{E}(u_{n-1} + r^{(l)} \otimes s^{(l)}))_{l\in\mathbb{N}}$ is convergent and consequently bounded.

As assumption $(A2)$ is satisfied, we can then extract a subsequence (which we still denote $(r^{(l)} \otimes s^{(l)})_{l\in\mathbb{N}}$) which weakly converges in V towards an element of Σ. In other words, there exist $r^\infty \in V_t$ and $s^\infty \in V_x$ such that $(r^{(l)} \otimes s^{(l)})_{l\in\mathbb{N}}$ weakly converges in V towards $r^\infty \otimes s^\infty$.

Furthermore, as the functional $\mathcal{E}$ is convex and continuous on V,

$$\mathcal{E}(u_{n-1} + r^\infty \otimes s^\infty) \leq \lim_{l\to\infty} \mathcal{E}(u_{n-1} + r^{(l)} \otimes s^{(l)}) = m.$$

Hence $\mathcal{E}(u_{n-1} + r^\infty \otimes s^\infty) = m$ so that (r^∞, s^∞) is a minimizer of problem (6.10).

Let us prove now that $r^\infty \otimes s^\infty \neq 0$ if and only if $u_{n-1} \neq u$.

If $u_{n-1} = u$, we have $\mathcal{E}(u + r \otimes s) > \mathcal{E}(u)$ for all $(r, s) \in V_t \times V_x$ such that $r \otimes s \neq 0$ as $\mathcal{E}$ is strictly convex. So a minimizer $r^\infty \otimes s^\infty$ of problem (6.10) must necessarily satisfy $r^\infty \otimes s^\infty = 0$.

Conversely, if $u_{n-1} \neq u$, we have $\mathcal{E}'(u_{n-1}) \neq 0$ and from Lemma 6.3.1, there exists a pair $(r, s) \in V_t \times V_x$ such that $\mathcal{E}(u_{n-1}+r\otimes s) < \mathcal{E}(u_{n-1})$. Hence, $\mathcal{E}(u_{n-1}+r^\infty\otimes s^\infty) < \mathcal{E}(u_{n-1})$ and $r^\infty \otimes s^\infty$ cannot be equal to 0. □

Proposition 6.3.2. *For each $n \in \mathbb{N}^*$, a minimizer (r_n, s_n) of problem (6.10) obeys the following Euler equation:*

$$\forall (r, s) \in V_t \times V_x, \quad \left\langle \mathcal{E}'(u_n), r \otimes s_n + r_n \otimes s \right\rangle = 0. \tag{6.17}$$

This result is obtained by considering the first-order conditions of the minimization problem (6.10). This will be useful in the proof of convergence.

6.3.2 Proof of convergence

In this subsection, we present the different steps of the proof.

Lemma 6.3.2. *The series $\sum_{n=1}^\infty \|r_n \otimes s_n\|_V^2$ and the sequence $(\mathcal{E}(u_n))_{n\in\mathbb{N}^*}$ are convergent.*

Proof. Let us set $E_n = \mathcal{E}(u_n) = \mathcal{E}\left(\sum_{k=1}^n r_k \otimes s_k\right)$.

Using (6.10), $E_n \leq \mathcal{E}\left(u_{n-1} + r \otimes s\right)$ for all $(r, s) \in V_t \times V_x$, and in particular, by taking $r \otimes s = 0$, $(E_n)_{n\in\mathbb{N}^*}$ is a non-increasing sequence. Moreover, it is bounded from below. Indeed, for all $n \in \mathbb{N}^*$, we have $E_n \geq \mathcal{E}(u)$. Thus, it is convergent.

This implies that the sequence defined as $W_n = E_{n-1} - E_n$ is nonnegative, converges to 0, and satisfies $\sum_{n=1}^{\infty} W_n < +\infty$.

Besides, the α-convexity of $\mathcal{E}$ yields the following inequality:

$W_n \geq -\langle \mathcal{E}'(u_n), r_n \otimes s_n \rangle + \frac{\alpha}{2}\|r_n \otimes s_n\|_V^2$.

Using the Euler equations (6.17), $\langle \mathcal{E}'(u_n), r_n \otimes s_n \rangle = 0$, and thus, $W_n \geq \frac{\alpha}{2}\|r_n \otimes s_n\|_V^2$. Hence the result. □

Lemma 6.3.3. *The sequence $(u_n)_{n\in\mathbb{N}^*}$ is bounded in V.*

Proof. By α-convexity of the functional $\mathcal{E}$, we have

$$\begin{aligned} \mathcal{E}(0) \quad &\geq \quad \mathcal{E}(u_n) \geq \mathcal{E}(u) + \langle \mathcal{E}'(u), u_n - u \rangle \\ &\quad + \frac{\alpha}{2}\|u - u_n\|_V^2. \end{aligned}$$

Thus $\|u - u_n\|_V^2 \leq \frac{2}{\alpha}(\mathcal{E}(0) - \mathcal{E}(u))$.

Therefore, the sequence $(u_n)_{n\in\mathbb{N}^*}$ is bounded in V. □

The following estimate is essential for the proof of convergence.

Proposition 6.3.3. *There exists a constant $A \in \mathbb{R}_+$ such that, for all $n \in \mathbb{N}^*$ and all $(r, s) \in V_t \times V_x$,*

$$\left|\left\langle \mathcal{E}'(u_{n-1}), r \otimes s \right\rangle\right| \leq A\|r_n \otimes s_n\|_V \|r \otimes s\|_V. \tag{6.18}$$

Proof. Let $M \in \mathbb{R}_+$ be such that for all $n \in \mathbb{N}^*$, $\|u_n\|_V \leq M$. Its existence is ensured by Lemma 6.3.3. Let $N \in \mathbb{R}_+$ be such that for all $n \in \mathbb{N}^*$, $\|r_n \otimes s_n\|_V \leq N$. Let $K = \overline{B}(0, M + 2N + 3)$ be the closed ball of V centered at 0 of radius $M + 2N + 3$. Let L be the Lipschitz constant associated with K in (6.9).

For all $(r, s) \in V_t \times V_x$, we have $\mathcal{E}(u_{n-1} + r \otimes s) - \mathcal{E}(u_{n-1} + r_n \otimes s_n) \geq 0$.

Then, by the convexity of $\mathcal{E}$, we have the following inequality

$$\left\langle \mathcal{E}'(u_{n-1} + r \otimes s), r_n \otimes s_n - r \otimes s \right\rangle \leq \mathcal{E}(u_{n-1} + r_n \otimes s_n) - \mathcal{E}(u_{n-1} + r \otimes s) \leq 0,$$

which leads to

$$\langle \mathcal{E}'(u_{n-1} + r \otimes s), r \otimes s \rangle \geq \langle \mathcal{E}'(u_{n-1} + r \otimes s), r_n \otimes s_n \rangle. \tag{6.19}$$

Let $(r, s) \in V_t \times V_x$ such that $\|r \otimes s\|_V \leq \max(1, \|r_n \otimes s_n\|_V)$. We then have, by using (6.9) and (6.19),

$$\begin{aligned} -\langle \mathcal{E}'(u_{n-1}), r \otimes s \rangle \quad &= \quad -\langle \mathcal{E}'(u_{n-1}), r \otimes s \rangle + \langle \mathcal{E}'(u_{n-1} + r \otimes s), r \otimes s \rangle \\ &\quad -\langle \mathcal{E}'(u_{n-1} + r \otimes s), r \otimes s \rangle \\ &\leq \quad L\|r \otimes s\|_V^2 - \langle \mathcal{E}'(u_{n-1} + r \otimes s), r \otimes s \rangle \\ &= \quad L\|r \otimes s\|_V^2 - \langle \mathcal{E}'(u_{n-1} + r \otimes s), r \otimes s \rangle + \langle \mathcal{E}'(u_{n-1} + r \otimes s), r_n \otimes s_n \rangle \\ &\quad -\langle \mathcal{E}'(u_{n-1} + r \otimes s), r_n \otimes s_n \rangle \\ &\leq \quad L\|r \otimes s\|_V^2 - \langle \mathcal{E}'(u_{n-1} + r \otimes s), r_n \otimes s_n \rangle \\ &= \quad L\|r \otimes s\|_V^2 - \langle \mathcal{E}'(u_{n-1} + r \otimes s), r_n \otimes s_n \rangle + \langle \mathcal{E}'(u_{n-1} + r_n \otimes s_n), r_n \otimes s_n \rangle \\ &\quad -\langle \mathcal{E}'(u_{n-1} + r_n \otimes s_n), r_n \otimes s_n \rangle \\ &\leq \quad L\|r \otimes s\|_V^2 + L\|r \otimes s - r_n \otimes s_n\|_V \|r_n \otimes s_n\|_V \\ &\quad -\langle \mathcal{E}'(u_{n-1} + r_n \otimes s_n), r_n \otimes s_n \rangle \\ &= \quad L\|r \otimes s\|_V^2 + L\|r \otimes s - r_n \otimes s_n\|_V \|r_n \otimes s_n\|_V. \end{aligned}$$

The last line has been obtained by taking into account the fact that $\langle \mathcal{E}'(u_{n-1} + r_n \otimes s_n), r_n \otimes s_n \rangle = 0$ because of the Euler equation (6.17).

Thus, for all $(r, s) \in V_t \times V_x$ such that $\|r \otimes s\|_V \leq \max(1, \|r_n \otimes s_n\|_V)$,

$$\langle \mathcal{E}'(u_{n-1}), r \otimes s \rangle + L\|r \otimes s\|_V^2 + L\|r \otimes s\|_V \|r_n \otimes s_n\|_V + L\|r_n \otimes s_n\|_V^2 \geq 0.$$

As a consequence,

$$|\langle \mathcal{E}'(u_{n-1}), r \otimes s \rangle| \leq L\|r \otimes s\|_V^2 + L\|r \otimes s\|_V \|r_n \otimes s_n\|_V + L\|r_n \otimes s_n\|_V^2.$$

Let $(r, s) \in V_t \times V_x$ such that $\|r \otimes s\|_V = 1$ and $t \in \mathbb{R}$ such that $t \leq \max(1, \|r_n \otimes s_n\|_V)$. Then, we have

$$|\langle \mathcal{E}'(u_{n-1}), tr \otimes s \rangle| \leq Lt^2\|r \otimes s\|_V^2 + Lt\|r \otimes s\|_V \|r_n \otimes s_n\|_V + L\|r_n \otimes s_n\|_V^2.$$

And, by setting $t = \|r_n \otimes s_n\|_V$, we obtain the following inequality for all $(r, s) \in V_t \times V_x$ such that $\|r \otimes s\|_V = 1$,

$$|\langle \mathcal{E}'(u_{n-1}), r \otimes s \rangle| \leq 3L\|r_n \otimes s_n\|_V \|r \otimes s\|_V.$$

Of course, this inequality also holds true for all $(r, s) \in V_t \times V_x$ such that $\|r \otimes s\|_V \neq 1$. Therefore, (6.18) holds with $A = 3L$. □

We now state an elementary result which will be useful in the sequel.

Lemma 6.3.4. *Let $(a_n)_{n \in \mathbb{N}^*}$ be a sommable sequence of $\mathbb{R}_+$. Then, there exists a subsequence of $(na_n)_{n \in \mathbb{N}^*}$ which converges to 0.*

Proof. If such a subsequence could not be extracted, it would imply

$$\exists \varepsilon_0 > 0 \ , \ \exists n_0 \in \mathbb{N}^* \ , \ \forall n \geq n_0 \ , \ na_n \geq \varepsilon_0.$$

Thus, the series $\sum_{n=1}^{\infty} a_n$ would diverge. Hence the contradiction. □

We are now in position to complete the proof of Theorem 6.2.1.

Proof. By Lemma 6.3.2, the sequence $(\mathcal{E}(u_n))_{n \in \mathbb{N}^*}$ is convergent. Let us denote its limit by E. We want to prove that $E = \mathcal{E}(u)$.

Firstly, for all $n \in \mathbb{N}^*$, $\mathcal{E}(u_n) \geq \mathcal{E}(u)$, since u is the global minimizer of the functional $\mathcal{E}$. By letting n go to infinity, we obtain $E \geq \mathcal{E}(u)$.

It remains to prove that $E \leq \mathcal{E}(u)$.

Let us first prove that $(\mathcal{E}'(u_n))_{n \in \mathbb{N}^*}$ weakly converges to 0 in V. Let $M \in \mathbb{R}_+$ such that for all $n \in \mathbb{N}^*$, $\|u_n\|_V \leq M$. Its existence is ensured by Lemma 6.3.3. Let $K = \overline{B}(0, M + 2 + \|u\|_V)$ be the closed ball of V centered at 0 of radius $M + 2 + \|u\|_V$. Let L be the Lipschitz constant associated with K in (6.9). Using (6.9) and the fact that $\mathcal{E}'(u) = 0$, we have $\|\mathcal{E}'(u_n)\|_V \leq L\|u - u_n\|_V$ and as $(u_n)_{n \in \mathbb{N}^*}$ is bounded in V by Lemma 6.3.3, we deduce that $(\mathcal{E}'(u_n))_{n \in \mathbb{N}^*}$ is also bounded in V. We can then extract a subsequence of $(\mathcal{E}'(u_n))_{n \in \mathbb{N}^*}$ which weakly converges in V towards $w \in V$. By using Proposition 3.3 and by letting n go to infinity in (6.18), we deduce that $\langle w, r \otimes s \rangle = 0$ for all $(r, s) \in V_t \times V_x$. Then, as $\mathrm{Span}(\Sigma)$ is dense in V with assumption $(A1)$, necessarily $w = 0$. Thus the sequence $(\mathcal{E}'(u_n))_{n \in \mathbb{N}^*}$ weakly converges to 0 in V.

As $\mathcal{E}$ is convex, we have the following inequality for all $n \in \mathbb{N}^*$,

$$\mathcal{E}(u_n) \leq \mathcal{E}(u) + \langle \mathcal{E}'(u_n), u_n - u \rangle_V. \tag{6.20}$$

Let us prove that we can extract a subsequence of $(\langle \mathcal{E}'(u_n), u_n\rangle)_{n\in\mathbb{N}^*}$ which converges to 0. Let $n \in \mathbb{N}^*$. By using Proposition 6.3.3,

$$\begin{aligned} |\langle \mathcal{E}'(u_n), u_n\rangle| &\leq \sum_{k=1}^{n} |\langle \mathcal{E}'(u_n), r_k \otimes s_k\rangle|, \\ &\leq A\sum_{k=1}^{n} \|r_{n+1}\otimes s_{n+1}\|_V \|r_k \otimes s_k\|_V, \\ &\leq A(n\|r_{n+1}\otimes s_{n+1}\|_V^2)^{1/2} \left(\sum_{k=1}^{n} \|r_k \otimes s_k\|_V^2\right)^{1/2}. \end{aligned}$$

As the sequence $\left(\sum_{k=1}^{n} \|r_k \otimes s_k\|_V^2\right)_{n\in\mathbb{N}^*}$ converges by Lemma 6.3.2, we have $\sum_{k=1}^{n} \|r_k \otimes s_k\|_V^2 \leq \sum_{k=1}^{\infty} \|r_k \otimes s_k\|_V^2 < \infty$. Furthermore, we can also extract a subsequence from $(n\|r_{n+1}\otimes s_{n+1}\|_V^2)_{n\in\mathbb{N}^*}$ which converges to 0 (see Lemma 6.3.4).

We can then extract a subsequence from $(\langle \mathcal{E}'(u_n), u_n\rangle_V)_{n\in\mathbb{N}^*}$ which converges to 0.

By letting n go to infinity in (6.20) with this subsequence, we obtain that $E \leq \mathcal{E}(u)$.

We have thus proved that $E = \mathcal{E}(u)$.

Besides, as the functional $\mathcal{E}$ is α-convex, (6.8) yields the following inequality,

$$\frac{\alpha}{2}\|u - u_n\|_V^2 \leq \mathcal{E}(u_n) - \mathcal{E}(u),$$

which necessarily implies that $\|u - u_n\|_V$ converges to 0 when n goes to infinity, which proves that $(u_n)_{n\in\mathbb{N}^*}$ strongly converges towards u in V. □

6.4 Rate of convergence in the finite-dimensional case

In the case when V_t and V_x are finite-dimensional, we are able to prove that the algorithm converges exponentially fast.

Theorem 6.4.1. *We assume that V_t and V_x are finite-dimensional and that assumptions $(A1)$, $(A2)$, $(A3)$ and $(A4)$ are fulfilled. Then there exist two constants $\tau > 0$ and $\sigma \in (0,1)$ such that for all $n \in \mathbb{N}^*$,*

$$0 \leq \mathcal{E}(u_n) - \mathcal{E}(u) \leq \tau\sigma^n, \tag{6.21}$$

and

$$\|u - u_n\|_V \leq \sqrt{\frac{2\tau}{\alpha}}\sigma^{n/2}. \tag{6.22}$$

Proof. Let us denote by $l = \dim V_t$ and $m = \dim V_x$. Then we can consider that $V_t = \mathbb{R}^l$, $V_x = \mathbb{R}^m$ and $V = \mathbb{R}^{l\times m}$ (which is implied by $(A1)$).

As the spaces are finite-dimensional, all the norms are equivalent, and we can consider without loss of generality that $\|.\|_{V_t}$, $\|.\|_{V_x}$ and $\|.\|_V$ are equal to the Frobenius norms of $\mathbb{R}^l$, $\mathbb{R}^m$ and $\mathbb{R}^{l\times m}$ defined by:

$$\begin{aligned} \|R\|_l^2 &= R^T R, \\ \|S\|_m^2 &= S^T S, \\ \|U\|_{lm}^2 &= \mathrm{Tr}(U^T U). \end{aligned} \tag{6.23}$$

Notice that for all $(R,S)\in\mathbb{R}^l\times\mathbb{R}^m$,

$$\|R\otimes S\|_V=\|RS^T\|_{lm}=\|R\|_l\|S\|_m.$$

Let $(\phi_i)_{1\leq i\leq l}$ and $(\psi_j)_{1\leq j\leq m}$ be orthonormal bases of V_t and V_x respectively. Then, $(\phi_i\otimes\psi_j)_{1\leq i\leq l,1\leq j\leq m}$ forms an orthonormal basis of V.

Our goal is to prove that there exists a constant $\sigma\in(0,1)$ such that for all $n\in\mathbb{N}^*$,

$$\mathcal{E}(u_n)-\mathcal{E}(u)\leq\sigma\left(\mathcal{E}(u_{n-1})-\mathcal{E}(u)\right). \tag{6.24}$$

Let $n\in\mathbb{N}^*$. Let us notice that

$$\mathcal{E}(u_n)-\mathcal{E}(u)=\mathcal{E}(u_n)-\mathcal{E}(u_{n-1})+\mathcal{E}(u_{n-1})-\mathcal{E}(u). \tag{6.25}$$

As for all $n\in\mathbb{N}^*$, $\mathcal{E}(u_n)-\mathcal{E}(u)\geq 0$, it is then sufficient with (6.25) to prove that there exists $\lambda\in(0,1)$ such that

$$\mathcal{E}(u_n)-\mathcal{E}(u_{n-1})\leq-\lambda\left(\mathcal{E}(u_{n-1})-\mathcal{E}(u)\right), \tag{6.26}$$

to have (6.24) with $\sigma=1-\lambda\in(0,1)$.

Let us notice that (6.8) and (6.17) yield

$$\mathcal{E}(u_n)-\mathcal{E}(u_{n-1})\leq-\frac{\alpha}{2}\|r_n\otimes s_n\|_V^2. \tag{6.27}$$

Besides, let $M\in\mathbb{R}_+$ such that for all $n\in\mathbb{N}^*$, $\|u_n\|_V\leq M$. Its existence is ensured by Lemma 6.3.3. Let $K=\overline{B}(0,M+\|u\|_V+2)$ be the closed ball of V centered at 0 of radius $M+\|u\|_V+2$. Let L be the Lipschitz constant of the gradient of $\mathcal{E}$ associated to K in (6.9).

Using (6.9) and the fact that $\mathcal{E}'(u)=0$, we also have,

$$\mathcal{E}(u_{n-1})-\mathcal{E}(u)\leq L\|u-u_{n-1}\|_V^2. \tag{6.28}$$

With (6.27) and (6.28), it is sufficient to prove that there exists a constant $\kappa\in(0,1)$ such that for all $n\in\mathbb{N}^*$,

$$\|r_n\otimes s_n\|_V\geq\kappa\|u-u_{n-1}\|_V, \tag{6.29}$$

in order to have (6.26) and hence (6.24).

Indeed, if (6.29) holds, we then have, using (6.27), (6.29) and (6.28),

$$\begin{aligned}\mathcal{E}(u_n)-\mathcal{E}(u_{n-1}) &\leq -\frac{\alpha}{2}\|r_n\otimes s_n\|_V^2\\ &\leq -\frac{\alpha}{2}\kappa^2\|u-u_{n-1}\|_V^2\\ &\leq -\frac{\alpha}{2L}\kappa^2\left(\mathcal{E}(u_{n-1})-\mathcal{E}(u)\right).\end{aligned}$$

As the α-convexity of $\mathcal{E}$ and the fact that $\mathcal{E}'(u)=0$ yields

$$\mathcal{E}(u_{n-1})-\mathcal{E}(u)\geq\frac{\alpha}{2}\|u-u_{n-1}\|_V^2, \tag{6.30}$$

inequalities (6.30) and (6.28) then imply that $\frac{\alpha}{2}\leq L$ and then (6.26) holds with $\lambda=\frac{\alpha}{2L}\kappa^2\in(0,1)$.

Let us prove inequality (6.29). From Proposition 6.3.3, estimate (6.18) holds true. As $(\phi_i \otimes \psi_j)_{1\leq i\leq l, 1\leq j\leq m}$ forms an orthonormal basis of V, we obtain, using (6.18),

$$\begin{aligned}\|\mathcal{E}'(u_n)\|_V^2 &= \sum_{i=1}^{l}\sum_{j=1}^{m}\langle \mathcal{E}'(u_n), \phi_i\otimes\psi_j\rangle^2 \\ &\leq \sum_{i=1}^{l}\sum_{j=1}^{m} A^2\|r_{n+1}\otimes s_{n+1}\|_V^2\|\phi_i\otimes\psi_j\|_V^2 \\ &= lmA^2\|r_{n+1}\otimes s_{n+1}\|_V^2.\end{aligned}$$

We then have the following estimate:

$$\|\mathcal{E}'(u_n)\|_V \leq \sqrt{lm}A\|r_{n+1}\otimes s_{n+1}\|_V. \tag{6.31}$$

The α-convexity of $\mathcal{E}$ and estimate (6.31) lead to

$$\begin{aligned}\mathcal{E}(u_{n-1}) - \mathcal{E}(u) &\leq -\langle \mathcal{E}'(u_{n-1}), u-u_{n-1}\rangle - \frac{\alpha}{2}\|u-u_{n-1}\|_V^2 \\ &\leq \sqrt{lm}A\|r_n\otimes s_n\|_V\|u-u_{n-1}\|_V - \frac{\alpha}{2}\|u-u_{n-1}\|_V^2.\end{aligned}$$

Besides, by using the fact that $\mathcal{E}(u_{n-1}) - \mathcal{E}(u) \geq 0$, we obtain

$$\|r_n\otimes s_n\|_V \geq \frac{\alpha}{2\sqrt{lm}A}\|u-u_{n-1}\|_V,$$

which is (6.29) with $\kappa = \frac{\alpha}{2\sqrt{lm}A} \in (0,1)$ for A large enough.

Hence the result. □

Remark 6.4.1. *This result can be generalized to the case of tensor products of more than two Hilbert spaces. Indeed, with the notation of Remark 6.2.2, and if we denote $l_1 = dimV_1, \cdots, l_q = dimV_q$, estimate (6.31) becomes*

$$\|\mathcal{E}'(u_n)\|_V \leq \sqrt{l_1\cdots l_q}A\|r_{n+1}\otimes s_{n+1}\|_V,$$

and the proof still holds.

6.5 Case of a local minimum

We are able to extend the results of Theorem 6.2.1 and Theorem 6.4.1 in the case when (r_n, s_n) in (6.10) is only defined as a **local** minimum which ensures the decrease of the energy, more precisely, when (r_n, s_n) is defined recursively as:

$$(r_n, s_n) = \underset{(r,s)\in V_t\times V_x}{\text{local argmin}}\, \mathcal{E}\left(u_{n-1} + r\otimes s\right), \tag{6.32}$$

such that

$$\mathcal{E}\left(u_n\right) < \mathcal{E}\left(u_{n-1}\right), \tag{6.33}$$

where u_n is defined as in (6.11).

To extend these results, we will need an additional assumption (which is naturally fulfilled in the finite dimensional case), see Remark 6.5.2 below:

$(A5)$ There exist $\beta, \gamma \in \mathbb{R}_+$ such that

$$\forall (r,s) \in V_t \times V_x, \quad \beta \|r\|_{V_t} \|s\|_{V_x} \leq \|r \otimes s\|_V \leq \gamma \|r\|_{V_t} \|s\|_{V_x}. \tag{6.34}$$

Theorem 6.5.1. *Let us suppose that the assumptions $(A1)$, $(A2)$, $(A3)$, $(A4)$ and $(A5)$ hold true. Then, the iterations of the algorithm described above are well-defined in the sense that (6.32) has at least one local minimizer (r_n, s_n) which satisfies (6.33). Moreover, the sequence $(u_n)_{n \in \mathbb{N}^*}$ strongly converges in V towards u.*

Proof. The proof is similar to the proof of Theorem 6.2.1 given in Section 6.3 except for Proposition 6.3.3 which gives estimate (6.18):

$$\forall (r,s) \in V_t \times V_x, \quad |\langle \mathcal{E}'(u_n), r \otimes s \rangle| \leq A \|r_{n+1} \otimes s_{n+1}\|_V \|r \otimes s\|_V.$$

This estimate is no longer true, but we have a similar result which will be enough to complete the proof. Indeed, let us prove that there exists a constant $B \in \mathbb{R}_+$ such that

$$\forall n \in \mathbb{N}^*, \ \forall (r,s) \in V_t \times V_x, \quad |\langle \mathcal{E}'(u_n), r \otimes s \rangle| \leq B \|r_n \otimes s_n\|_V \|r \otimes s\|_V. \tag{6.35}$$

Let $M \in \mathbb{R}_+$ such that for all $n \in \mathbb{N}^*$, $\|u_n\|_V \leq M$. Its existence is ensured by Lemma 6.3.3. Let $K = \overline{B}(0, M+2)$ be the closed ball of V centered at 0 and of radius $M+2$. Let L be the Lipschitz constant associated to K in (6.9).

Let $(r,s) \in V_t \times V_x$ and $n \in \mathbb{N}^*$. As (r_n, s_n) is a local minimum of $V_t \times V_x \ni (y,z) \mapsto \mathcal{E}\left(\sum_{k=1}^{n-1} r_k \otimes s_k + y \otimes z\right)$, there exists a constant $\eta \in (0,1)$ such that for all $\varepsilon \in (0, \eta)$, we have

$$\mathcal{E}\left(u_{n-1} + (r_n + \varepsilon r) \otimes (s_n + \varepsilon s)\right) \geq \mathcal{E}\left(u_{n-1} + r_n \otimes s_n\right). \tag{6.36}$$

Moreover, by convexity of the functional $\mathcal{E}$, we have the following inequality

$$\begin{array}{c} \mathcal{E}\left(u_{n-1} + (r_n + \varepsilon r) \otimes (s_n + \varepsilon s)\right) - \mathcal{E}\left(u_{n-1} + r_n \otimes s_n\right) \\ \leq \langle \mathcal{E}'(u_n + \varepsilon(r_n \otimes s + r \otimes s_n) + \varepsilon^2 r \otimes s), \varepsilon(r_n \otimes s + r \otimes s_n) + \varepsilon^2 r \otimes s \rangle. \end{array} \tag{6.37}$$

We deduce from (6.36), (6.37) and property (6.9) that, for all ε small enough so that $\|\varepsilon(r_n \otimes s + r \otimes s_n) + \varepsilon^2 r \otimes s\|_V \leq 1$,

$$\begin{array}{rcl} 0 & \leq & \langle \mathcal{E}'(u_n + \varepsilon(r_n \otimes s + r \otimes s_n) + \varepsilon^2 r \otimes s), \varepsilon(r_n \otimes s + r \otimes s_n) + \varepsilon^2 r \otimes s \rangle, \\ & \leq & \langle \mathcal{E}'(u_n), \varepsilon(r_n \otimes s + r \otimes s_n) + \varepsilon^2 r \otimes s \rangle + L \|\varepsilon(r_n \otimes s + r \otimes s_n) + \varepsilon^2 r \otimes s\|_V^2. \end{array}$$

As (r_n, s_n) is a local minimum of the functional $V_t \times V_x \ni (y,z) \mapsto \mathcal{E}\left(\sum_{k=1}^{n-1} r_k \otimes s_k + y \otimes z\right)$, (r_n, s_n) still obeys the Euler equation (6.17) and thus $\langle \mathcal{E}'(u_n), \varepsilon(r_n \otimes s + r \otimes s_n) \rangle = 0$.

Finally, we have

$$\varepsilon^2 \langle \mathcal{E}'(u_n), r \otimes s \rangle + L\varepsilon^2 \|r_n \otimes s + r \otimes s_n + \varepsilon r \otimes s\|_V^2 \geq 0.$$

Dividing this expression by ε^2 and letting ε go to zero, we obtain

$$\langle \mathcal{E}'(u_n), r \otimes s \rangle + L \|r_n \otimes s + r \otimes s_n\|_V^2 \geq 0,$$

which leads to

$$|\langle \mathcal{E}'(u_n), r \otimes s \rangle| \leq L\left(\|r_n \otimes s + r \otimes s_n\|_V^2 + \|r_n \otimes s - r \otimes s_n\|_V^2\right).$$

All this holds without the additional assumption (6.34) for all $(r,s) \in V_t \times V_x$. To derive estimate (6.35), we use the additional assumption we made on $\|.\|_V$:

$$\begin{aligned}
|\langle \mathcal{E}'(u_n), r \otimes s \rangle|^{1/2} &\leq \sqrt{L} \left(\|r_n \otimes s + r \otimes s_n\|_V^2 + \|r_n \otimes s - r \otimes s_n\|_V^2 \right)^{1/2}, \\
&\leq \sqrt{L} \left(\|r_n \otimes s + r \otimes s_n\|_V + \|r_n \otimes s - r \otimes s_n\|_V \right), \\
&\leq 2\sqrt{L} \left(\|r_n \otimes s\|_V + \|r \otimes s_n\|_V \right), \\
&\leq 2\sqrt{L}\gamma \left(\|r_n\|_{V_t} \|s\|_{V_x} + \|r\|_{V_t} \|s_n\|_{V_x} \right).
\end{aligned}$$

We can then choose $(r_n^*, s_n^*) \in V_t \times V_x$ and $(r^*, s^*) \in V_t \times V_x$ such that $r_n^* \otimes s_n^* = r_n \otimes s_n$ and $r^* \otimes s^* = r \otimes s$ and such that $\|r_n^*\|_{V_t} = \|s_n^*\|_{V_x} \leq \sqrt{\frac{1}{\beta}\|r_n \otimes s_n\|_V}$ and $\|r^*\|_{V_t} = \|s^*\|_{V_x} \leq \sqrt{\frac{1}{\beta}\|r \otimes s\|_V}$. Thus,

$$\begin{aligned}
|\langle \mathcal{E}'(u_n), r \otimes s \rangle|^{1/2} &= |\langle \mathcal{E}'(u_n), r^* \otimes s^* \rangle|^{1/2}, \\
&\leq 2\sqrt{L}\gamma \left(\|r_n^*\|_{V_t} \|s^*\|_{V_x} + \|r^*\|_{V_t} \|s_n^*\|_{V_x} \right), \\
&\leq 4\frac{\sqrt{L}\gamma}{\beta} \|r_n \otimes s_n\|_V^{1/2} \|r \otimes s\|_V^{1/2}.
\end{aligned}$$

And in the end, we obtain estimate (6.35) with $B = 16\frac{L\gamma^2}{\beta^2}$. With this result, it is then possible to conclude as in the proof of Theorem 6.2.1. □

Remark 6.5.1. Problem (6.14) falls into the scope of Theorem 6.5.1. On the other hand, this is not the case for problem (6.15), for which property (6.34) is not true (see Remark 6.2.5). We were not able to prove a similar result in the case when $\|.\|_V$ does not satisfy property (6.34).

Remark 6.5.2. Here are two typical examples for which assumption $(A5)$ holds :

- In the case when $V = V_t \otimes V_x$, property (6.34) holds with $\beta = \gamma = 1$. This holds in uncertainty propagation problems where $V = L_T^2(\mathcal{T}, H)$ with H an Hilbert space of real-valued functions defined on $\mathcal{X}$. Denoting by $V_t = L_T^2(\mathcal{T}, \mathbb{R})$ and $V_x = H$, then $V = V_t \otimes V_x$.
- In other cases, to find an approximation of the global minimum of the energy $\mathcal{E}$, the Hilbert spaces V_t and V_x are usually discretized in finite-dimensional spaces. The problem can then be rewritten as a problem over $V_t = \mathbb{R}^l$, $V_x = \mathbb{R}^m$ with $l, m \in \mathbb{N}^*$, and then V is naturally defined as the Hilbert space $V = \mathbb{R}^{l \times m}$. Then, assumptions $(A1)$, $(A2)$, $(A3)$ and $(A4)$ are automatically satisfied on the discretized spaces. As all the norms are equivalent in finite dimension, the norms on $\mathbb{R}^l$, $\mathbb{R}^m$ and $\mathbb{R}^{l \times m}$ induced by the norms defined over the original Hilbert spaces V_t, V_x and V are equivalent to the Frobenius norms, defined by (6.23). These norms satisfy property (6.34) since for all $(R,S) \in \mathbb{R}^l \times \mathbb{R}^m$, $\|RS^T\|_{lm} = \|R\|_l \|S\|_m$. Hence, the norms induced by the norms defined on the original Hilbert spaces automatically satisfy property (6.34) even if the property is not satisfied in the continuous spaces.

As in Section 6.4, we can prove that the algorithm defined by (6.32) and (6.33) converges exponentially fast in finite dimension.

Theorem 6.5.2. *Let us consider the algorithm defined by (6.32) and (6.33). Let $l, m \in \mathbb{N}^*$. Let $V_t = \mathbb{R}^l$, $V_x = \mathbb{R}^m$ and $V = \mathbb{R}^{l\times m}$. Then there exist two constants $\tau > 0$ and $\sigma \in (0,1)$ such that for all $n \in \mathbb{N}^*$,*

$$0 \leq \mathcal{E}(u_n) - \mathcal{E}(u) \leq \tau\sigma^n, \tag{6.38}$$

and

$$\|u - u_n\|_V \leq \sqrt{\frac{2\tau}{\alpha}}\sigma^{n/2}. \tag{6.39}$$

Proof. As the spaces are finite-dimensional, assumptions $(A1)$, $(A2)$, $(A3)$, $(A4)$ and $(A5)$ are automatically fulfilled (see Remark 6.5.2) and estimate (6.35) holds true. The proof is similar to the proof of Theorem 6.4.1. Indeed, (6.25), (6.27), (6.28) and (6.30) still hold. Then it is sufficient to prove an inequality similar to (6.29) to prove Theorem 6.5.2. However, as (6.35) holds instead of (6.18), inequality (6.31) is replaced by:

$$\|\mathcal{E}'(u_n)\|_V \leq \sqrt{lm}B\|r_n \otimes s_n\|_V^2, \tag{6.40}$$

and consequently, an inequality similar to (6.29) must be obtained in another way.

Let $M \in \mathbb{R}_+$ such that for all $n \in \mathbb{N}^*$, $\|u_n\|_V \leq M$. Its existence is ensured by Lemma 6.3.3. Let $K = \overline{B}(0, M + 2 + \|u\|_V)$ be the closed ball of V centered at 0 of radius $M + 2 + \|u\|_V$. Let L be the Lipschitz constant associated with K in (6.9).

On the one hand, using the convexity of $\mathcal{E}$, (6.9) and the fact that $\mathcal{E}'(u) = 0$, we have

$$\mathcal{E}(u_{n-1}) - \mathcal{E}(u_n) \leq -\langle \mathcal{E}'(u_{n-1}), r_n \otimes s_n\rangle \leq L\|r_n \otimes s_n\|_V\|u - u_{n-1}\|_V. \tag{6.41}$$

On the other hand, (6.30), the convexity of $\mathcal{E}$, (6.17) and (6.40) yield

$$\begin{aligned}
\mathcal{E}(u_{n-1}) - \mathcal{E}(u_n) &= \mathcal{E}(u_{n-1}) - \mathcal{E}(u) + \mathcal{E}(u) - \mathcal{E}(u_n) \\
&\geq \frac{\alpha}{2}\|u - u_{n-1}\|_V^2 + \mathcal{E}(u) - \mathcal{E}(u_n) \\
&\geq \frac{\alpha}{2}\|u - u_{n-1}\|_V^2 + \langle \mathcal{E}'(u_n), u - u_n\rangle \\
&= \frac{\alpha}{2}\|u - u_{n-1}\|_V^2 + \langle \mathcal{E}'(u_n), u - u_{n-1}\rangle \\
&\geq \frac{\alpha}{2}\|u - u_{n-1}\|_V^2 - \sqrt{lm}B\|r_n \otimes s_n\|_V\|u - u_{n-1}\|_V.
\end{aligned}$$

Then, using (6.41), we have (6.29) with $\kappa = \frac{\alpha}{2\left(L+\sqrt{lm}B\right)} \in (0,1)$. We can conclude as in the proof of Theorem 6.4.1. □

Remark 6.5.3. The results given in this section may not stand when we consider more than two Hilbert spaces. Indeed, the scheme of the proof of Theorem 6.5.1 cannot be easily adapted and we do not necessarily have an estimate similar to (6.35).

6.6 Numerical results

In this section, we illustrate the convergence properties of the algorithm introduced in Section 6.2 in a very simple setting, namely a one-dimensional membrane problem with uncertainty, with a basic space discretization method. Additional investigations to demonstrate the applicability and the efficiency of the procedure on high-dimensionnal problems are still required. We however refer to Nouy [157] for illustrations of the interest of the method for problems in high dimensions.

6.6.1 Implementation of the algorithm

Let us recall problem (6.14). Let $f \in L^2_T(\mathcal{T}, H^{-1}(\mathcal{X}))$ and $g \in L^2_T(\mathcal{T}, H^1_0(\mathcal{X}))$. Let us assume that the random variable T has a probability density $p(t)$ for $t \in \mathcal{T}$. In other words,

$$\mathbb{P}(T \in \mathcal{D}) = \int_{\mathcal{D}} p(t)\, dt,$$

where $\mathcal{D}$ is a measurable subset of $\mathcal{T}$.

For a given value of the penalization parameter $\rho \in \mathbb{R}_+$, we wish to calculate an approximation of the minimizer u of the problem

$$\left\{ \begin{array}{l} \text{Find } u \in L^2_T(\mathcal{T}, H^1_0(\mathcal{X})) \text{ such that} \\ \quad u = \operatorname{argmin}_{v \in L^2_T(\mathcal{T}, H^1_0(\mathcal{X}))} \mathcal{E}(v), \end{array} \right. \tag{6.42}$$

where

$$\mathcal{E}(v) = \mathbb{E}\left[\frac{1}{2} \int_{\mathcal{X}} |\nabla_x v(T,x)|^2\, dx - \langle f(T,.), v(T,.) \rangle_{H^{-1}(\mathcal{X}), H^1_0(\mathcal{X})} + \frac{\rho}{2} \int_{\mathcal{X}} [g(T,x) - v(T,x)]_+^2 dx \right].$$

In other words,

$$\begin{aligned} \mathcal{E}(v) &= \frac{1}{2} \int_{\mathcal{T}\times\mathcal{X}} |\nabla_x v(t,x)|^2 p(t)\, dt\, dx - \int_{\mathcal{T}} \langle f(t,.), v(t,.) \rangle_{H^{-1}(\mathcal{X}), H^1_0(\mathcal{X})}\, p(t)\, dt \\ &\quad + \frac{\rho}{2} \int_{\mathcal{T}\times\mathcal{X}} [g(t,x) - v(t,x)]_+^2 p(t)\, dt\, dx. \end{aligned}$$

In this case, the greedy algorithm can be rewritten in the following form. Set $f_0 = f$ and $g_0 = g$ and define recursively $(r_n, s_n) \in L^2_T(\mathcal{T}) \times H^1_0(\mathcal{X})$ as

$$(r_n, s_n) \in \operatorname*{argmin}_{(r,s) \in L^2_T(\mathcal{T}) \times H^1_0(\mathcal{X})} \mathcal{E}_n(r \otimes s),$$

with

$$\begin{aligned} \mathcal{E}_n(r \otimes s) &= \frac{1}{2} \int_{\mathcal{T}\times\mathcal{X}} |\nabla_x (r \otimes s)(t,x)|^2 p(t)\, dt\, dx - \int_{\mathcal{T}} \langle f_{n-1}(t,.), r \otimes s(t,.) \rangle_{H^{-1}(\mathcal{X}), H^1_0(\mathcal{X})}\, p(t)\, dt \\ &\quad + \frac{\rho}{2} \int_{\mathcal{T}\times\mathcal{X}} [g_{n-1}(t,x) - r \otimes s(t,x)]_+^2 p(t)\, dt\, dx, \end{aligned}$$

where

$$\begin{aligned} f_n &= f_{n-1} + \Delta_x(r_n \otimes s_n), \\ g_n &= g_{n-1} - r_n \otimes s_n. \end{aligned}$$

Indeed,

$$\mathcal{E}(u_{n-1} + r \otimes s) = \mathcal{E}(u_{n-1}) - \frac{\rho}{2} \int_{\mathcal{T}\times\mathcal{X}} [g_{n-1}(t,x)]_+^2 p(t)\, dt\, dx + \mathcal{E}_n(r \otimes s),$$

where u_n is defined as in (6.11).

In fact, from Theorem 6.5.1, it is sufficient for (r_n, s_n) to be a *local* minimum of $L^2_T(\mathcal{T}) \times H^1_0(\mathcal{X}) \ni (r,s) \mapsto \mathcal{E}_n(r \otimes s)$ such that:

$$\mathcal{E}_n(r_n \otimes s_n) < \frac{\rho}{2} \int_{\mathcal{T}\times\mathcal{X}} [g_{n-1}(t,x)]^2_+ p(t)\, dt\, dx,$$

which ensures (6.33). Denoting $(\cdot,\cdot)$ the scalar product of $L^2_T(\mathcal{T}, L^2(\mathcal{X}))$, assuming $f \in L^2_T(\mathcal{T}, L^2(\mathcal{X}))$, we can also write for all $v \in L^2_T(\mathcal{T}, H^1_0(\mathcal{X}))$,

$$\mathcal{E}(v) \;=\; \frac{1}{2}\left(\nabla_x v, \nabla_x v\right) - (f,v) + \frac{\rho}{2}\left([g-v]_+, [g-v]_+\right),$$

where

$$\forall v, w \in L^2_T(\mathcal{T}, L^2(\mathcal{X})),\ (u,v) = \int_{\mathcal{T}\times\mathcal{X}} v(t,x) w(t,x) p(t)\, dt\, dx.$$

We write the algorithm in the discrete case, and, for clarity, we restrict ourselves to the case of two open intervals $\mathcal{T}$ and $\mathcal{X}$ of $\mathbb{R}$. More precisely, $\mathcal{T} = (a,b)$ and $\mathcal{X} = (c,d)$, with $a, b, c, d \in \mathbb{R}$, such that $a < b$ and $c < d$. Let us denote $\Omega = \mathcal{T} \times \mathcal{X}$. Let us also assume for simplicity that T follows a uniform law of probability on $\mathcal{T} = (a,b)$ (i.e. $p(t) = \frac{1}{b-a}$ for all $t \in \mathcal{T}$).

Let $l, m \in \mathbb{N}^*$ with $l \geq 2$ the numbers of degrees of freedom in the discretized spaces of V_t and V_x respectively. We introduce a regular subdivision $(t_i)_{1\leq i\leq l}$ of the interval $\mathcal{T} = (a,b)$ (and respectively a regular subdivision $(x_j)_{0\leq j\leq m+1}$ of the interval $\mathcal{X} = (c,d)$):

$$\begin{array}{c} \forall 1 \leq i \leq l,\ t_i = a + (i-1)\frac{b-a}{l-1}, \\ \forall 0 \leq j \leq m+1,\ x_j = c + j\frac{d-c}{m+1}. \end{array}$$

Let $(\phi_i)_{1\leq i\leq l} \subset V_t$ and $(\psi_j)_{1\leq j\leq m} \subset V_x$ be continuous functions such that

$$\phi_i(t_{i'}) = \delta_{ii'},\ \forall 1 \leq i, i' \leq l, \tag{6.43}$$

and

$$\psi_j(x_{j'}) = \delta_{jj'},\ \forall 1 \leq j \leq m,\ 0 \leq j' \leq m+1, \tag{6.44}$$

and let us consider $\widetilde{V}_t = \mathrm{Span}(\phi_i)_{1\leq i\leq l}$, $\widetilde{V}_x = \mathrm{Span}(\psi_j)_{1\leq j\leq m}$ and $\widetilde{V} = \widetilde{V}_t \otimes \widetilde{V}_x$. For example, Lagrange finite elements satisfy properties (6.43) and (6.44).

Our goal is to find an approximation $\widetilde{u} \in \widetilde{V}$ of the function u under the following form

$$\widetilde{u}(t,x) = \sum_{i=1}^{l}\sum_{j=1}^{m} U^{ij} \phi_i(t) \psi_j(x) \tag{6.45}$$

where $U = (U^{ij})_{1\leq i\leq l, 1\leq j\leq m} = (\widetilde{u}(t_i, x_j))_{1\leq i\leq l, 1\leq j\leq m} \in \mathbb{R}^{k\times l}$.

A discretized variational version of problem (6.42) is

$$\left\{ \begin{array}{l} \text{Find } \widetilde{u} \in \widetilde{V} \text{ such that} \\ \widetilde{u} = \mathrm{argmin}_{v\in\widetilde{V}}\, \mathcal{E}(v). \end{array} \right. \tag{6.46}$$

For $v \in \widetilde{V}$, it holds

$$\mathcal{E}(v) = \frac{1}{2}\mathrm{Tr}\left(\Phi V D V^T\right) - \mathrm{Tr}\left(F V^T\right) + \frac{\rho}{2}\int_{\mathcal{T}\times\mathcal{X}} \left[\sum_{i=1}^{l}\sum_{j=1}^{m} V^{ij}\phi_i(t)\otimes\psi_j(x) - g(t,x)\right]^2_+ dt\, dx,$$

where $V = (V^{ij})_{1\leq i\leq l,\, 1\leq j\leq m} = (v(t_i, x_j))_{1\leq i\leq l,\, 1\leq j\leq m}$ and $\Phi = (\Phi^{ii'})_{1\leq i,i'\leq l} \in \mathbb{R}^{l\times l}$, $D = (D^{jj'})_{1\leq j,j'\leq m} \in \mathbb{R}^{m\times m}$, $F = (F^{ij})_{1\leq i\leq l, 1\leq j\leq m} \in \mathbb{R}^{l\times m}$ are defined as

$$\begin{aligned}
\Phi^{ii'} &= \frac{1}{b-a}\int_{\mathcal{T}} \phi_i(t)\phi_{i'}(t)\,dt,\\
D^{jj'} &= \int_{\mathcal{X}} \partial_x\psi_j(x)\partial_x\psi_{j'}(x)\,dx,\\
F^{ij} &= \frac{1}{b-a}\int_{\mathcal{T}} \langle f(t,.), \phi_i\otimes\psi_j(t,.)\rangle_{H^{-1}(\mathcal{X}),H_0^1(\mathcal{X})}\,dt.
\end{aligned}$$

In practice, to simplify the computation of the last term in $\mathcal{E}(v)$, we use the simple numerical integration formula:

$$\int_{\mathcal{T}\times\mathcal{X}} w(t,x)\,dt\,dx \approx \frac{1}{l}\frac{1}{m+2}\sum_{i=1}^{l}\sum_{j=0}^{m+1} w(t_i,x_j),$$

with $w(t,x) = \left[\sum_{i=1}^{l}\sum_{j=1}^{m} V^{ij}\phi_i(t)\otimes\psi_j(x) - g(t,x)\right]_+^2$. For this, we need to assume that g is a continuous function on $\overline{\Omega}$.

In the end, introducing $\widetilde{\rho} = \frac{\rho}{l(m+2)}$ and $G = (g(t_i,x_j))_{1\leq i\leq l,\, 1\leq j\leq m}$, the discretized version of problem (6.42) is given by:

$$\begin{cases} \text{Find } U\in\mathbb{R}^{l\times m} \text{ such that}\\ U = \operatorname{argmin}_{V\in\mathbb{R}^{l\times m}} \frac{1}{2}\Phi V D : V - F : V + \frac{\widetilde{\rho}}{2}[G-V]_+ : [G-V]_+, \end{cases} \tag{6.47}$$

where for $A, B\in\mathbb{R}^{l\times m}$,

$$A : B = \operatorname{Tr}(AB^T) = \sum_{1\leq i\leq l}\sum_{1\leq j\leq m} A_{ij}B_{ij}.$$

This problem is equivalent to:

$$\text{Find } U\in\mathbb{R}^{l\times m} \text{ such that } \Phi U D = F + \widetilde{\rho}[G-U]_+.$$

For each function $r\in\widetilde{V}_t$ and $s\in\widetilde{V}_x$, we denote by $R\in\mathbb{R}^l$ and $S\in\mathbb{R}^m$, the vectors which are defined by

$$\forall 1\leq i\leq l,\ R^i = r(t_i),$$

and

$$\forall 1\leq j\leq m,\ S^j = s(x_j).$$

Approximations of the functions $r(t)$ and $s(x)$ can then be expanded respectively in the bases $(\phi_i)_{1\leq i\leq l}$ and $(\psi_j)_{1\leq j\leq m}$ as

$$\begin{aligned}
r(t) &= \sum_{i=1}^{l} R^i\phi_i(t),\\
s(x) &= \sum_{j=1}^{l} S^j\psi_j(x).
\end{aligned}$$

The greedy algorithm can then be rewritten as:
Choose a threshold $\varepsilon > 0$ and set $F_0 = F$, $G_0 = G$. At iteration $n\geq 1$:

1. find $R_n = (R_n^i)_{1\leq i\leq l}$ and $S_n = (S_n^j)_{1\leq j\leq m}$ two vectors respectively in $\mathbb{R}^l$ and $\mathbb{R}^m$ such that:

$$(R_n, S_n) \in \underset{(R,S)\in\mathbb{R}^l\times\mathbb{R}^m}{\operatorname{argmin}} \widetilde{\mathcal{E}}_n(R,S),$$

with

$$\begin{aligned}\widetilde{\mathcal{E}}_n(R,S) &= \frac{1}{2}\Phi(RS^T)D : (RS^T) - F_{n-1} : (RS^T) \\ &\quad + \frac{\widetilde{\rho}}{2}[G_{n-1} - RS^T]_+ : [G_{n-1} - RS^T]_+.\end{aligned}$$

2. set $F_n = F_{n-1} - (R_n S_n^T)D$, and $G_n = G_{n-1} - R_n S_n^T$.

3. if $\|R_n S_n^T\| \geq \varepsilon$, proceed to iteration $n+1$. Otherwise, stop.

The remaining question is: how can we compute (R_n, S_n) at step 1? This critical step is described in the following section.

6.6.2 Computing (R_n, S_n)

Fixed-point procedure

Let us first describe a method which has been proposed by Nouy [157] and Chinesta [5], that is the fixed-point procedure and which we use in our final numerical implementation (see Section 6.6.2). We present this algorithm in a particular case. Let us consider $V_t = \mathbb{R}^l$, $V_x = \mathbb{R}^m$ and $V = \mathbb{R}^{l\times m}$ endowed with the Frobenius norms defined by (6.23). We fix a given matrix $M \in \mathbb{R}^{l\times m}$. Let us define the energy functional as $\widetilde{\mathcal{E}}(W) = \|M - W\|_V^2$ for $W \in \mathbb{R}^{l\times m}$. In this particular case, applying the greedy algorithm described above consists in computing the Singular Value Decomposition of the matrix M.

In this particular case, the greedy algorithm can be rewritten in the following form.

Choose a threshold $\varepsilon > 0$ and set $M_0 = M$. At iteration $n \geq 1$,

1. find two vectors R_n and S_n respectively in $\mathbb{R}^l$ and $\mathbb{R}^m$ such that

$$(R_n, S_n) \in \underset{(R,S)\in\mathbb{R}^l\times\mathbb{R}^m}{\operatorname{argmin}} \left\|M_{n-1} - RS^T\right\|_V^2. \tag{6.48}$$

2. set $M_n = M_{n-1} - R_n S_n^T$.

3. if $\|R_n S_n^T\|_V \geq \varepsilon$, proceed to iteration $n+1$. Otherwise, stop.

The Euler equation associated to this problem can be rewritten as

$$\left\{\begin{array}{rcl}\|S_n\|_{V_x}^2 R_n &=& M_{n-1}S_n, \\ \|R_n\|_{V_t}^2 S_n &=& (M_{n-1})^T R_n.\end{array}\right.$$

The method which is generally used [130] to solve these Euler equation is a fixed-point algorithm, which simply reads (for a fixed n): at iteration $q \geq 0$, compute two vectors $(R_n^{(q+1)}, S_n^{(q+1)}) \in \mathbb{R}^k \times \mathbb{R}^l$ such that

$$\left\{\begin{array}{rcl}\|S_n^{(q)}\|_{V_x}^2 R_n^{(q+1)} &=& M_{n-1}S_n^{(q)}, \\ \|R_n^{(q+1)}\|_{V_t}^2 S_n^{(q+1)} &=& (M_{n-1})^T R_n^{(q+1)}.\end{array}\right. \tag{6.49}$$

One can check [130] that this procedure is similar to the power method to compute the largest eigenvalue (and associated eigenvector) of the matrix $(M_{n-1})^T M_{n-1}$.

One could think of applying this fixed-point procedure to the case of the obstacle problem we consider in this article. In our case, the Euler equation

$$\left\{\begin{array}{rcl}(\Phi R_n : R_n) D S_n &=& F_{n-1}^T R_n + \widetilde{\rho}[G_{n-1} - R_n S_n^T]_+^T R_n,\\ (DS_n : S_n)\Phi R_n &=& F_{n-1} S_n + \widetilde{\rho}[G_{n-1} - R_n S_n^T]_+ S_n.\end{array}\right.$$

could be solved *a priori* with a fixed point algorithm, which, at iteration q, might be written as

$$\left\{\begin{array}{rcl}(\Phi R_n^{(q)} : R_n^{(q)}) D S_n^{(q+1)} &=& F_{n-1}^T R_n^{(q)} + \widetilde{\rho}\left[G_{n-1} - R_n^{(q)}\left(S_n^{(q)}\right)^T\right]_+^T R_n^{(q)},\\ (DS_n^{(q+1)} : S_n^{(q+1)})\Phi R_n^{(q+1)} &=& F_{n-1} S_n^{(q+1)} + \widetilde{\rho}\left[G_{n-1} - R_n^{(q)}\left(S_n^{(q+1)}\right)^T\right]_+ S_n^{(q+1)}.\end{array}\right.$$

Unfortunately, we were not able to make this fully-explicit fixed point algorithm converge for large values of the parameter ρ. We therefore decided to resort to a minimization procedure.

Minimization procedure

The approach we adopt then is the following. We choose an initial pair $(R_n^0, S_n^0) \in \mathbb{R}^l \times \mathbb{R}^m$ and then perform a quasi-newton algorithm to find a local minimum of the function

$$\frac{1}{2}\Phi\left(RS^T\right) D : \left(RS^T\right) - F_{n-1} : \left(RS^T\right) + \frac{\rho}{2}\left[G_{n-1} - RS^T\right]_+ : \left[G_{n-1} - RS^T\right]_+.$$

The main difficulty is to find a proper initial pair $(R_n^{(0)}, S_n^{(0)})$ such that

$$\frac{1}{2}\Phi\left(R_n^{(0)} S_n^{(0)T}\right) D : \left(R_n^{(0)} S_n^{(0)T}\right) - F_{n-1} : \left(R_n^{(0)} S_n^{(0)T}\right) + \frac{\widetilde{\rho}}{2}\left[G_{n-1} - R_n^{(0)} S_n^{(0)T}\right]_+ : \left[G_{n-1} - R_n^{(0)} S_n^{(0)T}\right]_+$$

$$< \frac{\widetilde{\rho}}{2}[G_{n-1}]_+ : [G_{n-1}]_+,$$

to ensure that the energy decreases (see (6.33)).

Let us describe our approach in the continuous setting with the notation used in Section 6.4. It consists in finding a pair $\left(r_n^{(0)}, s_n^{(0)}\right) \in V_t \times V_x$ such that

$$\mathcal{E}\left(u_{n-1} + r_n^{(0)} \otimes s_n^{(0)}\right) < \mathcal{E}\left(u_{n-1}\right).$$

We notice that for $(r, s) \in V_t \times V_x$, and $\eta > 0$, we have

$$\mathcal{E}\left(u_{n-1} + \eta r \otimes s\right) - \mathcal{E}\left(u_{n-1}\right) = \eta\left\langle \mathcal{E}'\left(u_{n-1}\right), r \otimes s\right\rangle + o(\eta),$$

for η small enough.

The idea is then to find a pair $(r, s) \in V_t \times V_x$ such that $\langle \mathcal{E}'(u_{n-1}), r \otimes s\rangle < 0$, so that there exists $\eta > 0$ small enough for which $\mathcal{E}\left(u_{n-1} + \eta r \otimes s\right) - \mathcal{E}\left(u_{n-1}\right) < 0$. Then, $r_n^{(0)} \otimes s_n^{(0)} = \eta r \otimes s$ is a good initial guess.

Let us first consider the pair $\left(\overline{r_n^{(0)}}, \overline{s_n^{(0)}}\right) \in V_t \times V_x$ such that

$$\left(\overline{r_n^{(0)}}, \overline{s_n^{(0)}}\right) \in \underset{(r,s)\in V_t\times V_x}{\operatorname{argmin}} \frac{1}{2} \left\|\mathcal{E}'\left(u_{n-1}\right) - r \otimes s\right\|_V^2.$$

In other words, we consider $\left(\overline{r_n^{(0)}}, \overline{s_n^{(0)}}\right)$ the first term of the singular value decomposition of $\mathcal{E}'\left(u_{n-1}\right)$ in V. The Euler equations then imply

$$-\left\langle \mathcal{E}'(u_{n-1}) - \overline{r_n^{(0)}} \otimes \overline{s_n^{(0)}}, \overline{r_n^{(0)}} \otimes \overline{s_n^{(0)}} \right\rangle = 0,$$

and therefore,

$$\left\langle \mathcal{E}'(u_{n-1}), \overline{r_n^{(0)}} \otimes \overline{s_n^{(0)}} \right\rangle = \left\| \overline{r_n^{(0)}} \otimes \overline{s_n^{(0)}} \right\|_V^2 > 0.$$

By taking $r_n^{(0)} \otimes s_n^{(0)} = -\eta \overline{r_n^{(0)}} \otimes \overline{s_n^{(0)}}$, there exists then $\eta > 0$ small enough such that

$$\mathcal{E}\left(u_{n-1} + r_n^{(0)} \otimes s_n^{(0)}\right) - \mathcal{E}\left(u_{n-1}\right) < 0.$$

In the discrete case associated to problem (6.14), $\left(\overline{R_n^{(0)}}, \overline{S_n^{(0)}}\right)$ is obtained by taking the first term of the singular value decomposition of the matrix $F_{n-1} + \widetilde{\rho}[G_{n-1}]_+$. This can be done with a fixed point procedure similar to (6.49).

Once we have this initial guess $\left(R_n^{(0)}, S_n^{(0)}\right)$, we perform a quasi-newton algorithm to minimize the energy. The computations are done with the software Scilab [2] and the quasi-Newton procedure is performed via the *optim* procedure of Scilab.

Let us point out that this procedure is intrusive in general.

6.6.3 One-dimensional membrane problem

In this section, we present the results we obtained with this algorithm on the following membrane problem.

We suppose $\mathcal{X} = \mathcal{T} = (0,1)$. We consider a random variable T following a uniform law of probability on the interval $(0,1)$. We wish to study problem (6.42) with the following values for f and g,

$$\forall (t,x) \in (0,1)^2\ ,\ f(t,x) = -1 \text{ and } g(t,x) = t[\sin(3\pi x)]_+ + (t-1)[\sin(3\pi x)]_-.$$

The negative part of $a \in \mathbb{R}$, *i.e.* $[a]_- = 0$ if $a \geq 0$, and $[a]_- = -a$ if $a \leq 0$, is denoted by $[a]_-$.

The above problem models a rope attached at $x = 0$ and $x = 1$ subjected to gravity and resting upon obstacles whose altitudes are given by $g(t,x)$. The quantity $u(t,x)$ then represents the altitude of the rope at abscissa x when $T = t$.

This problem is approximated by problem (6.47) with parameter $\widetilde{\rho} = 2500$. The problem is discretized with a regular mesh and $\mathbb{P}_1$ finite elements in each direction. Discretization parameters are chosen as $l = m = 40$.

Fig. 6.2 represents the altitude of the obstacles given by $g(t,x)$ for $(t,x) \in [0,1]^2$.

The algorithm described in the previous sections is then applied with the following stopping criterion: $\|R_n S_n^T\|_V < 5.10^{-5}$ with $\|A\|_V = \sqrt{\mathrm{Tr}(AA^T)} = \sqrt{\sum_{i=1}^k \sum_{j=1}^l A_{ij}^2}$ for $A \in \mathbb{R}^{k\times l}$.

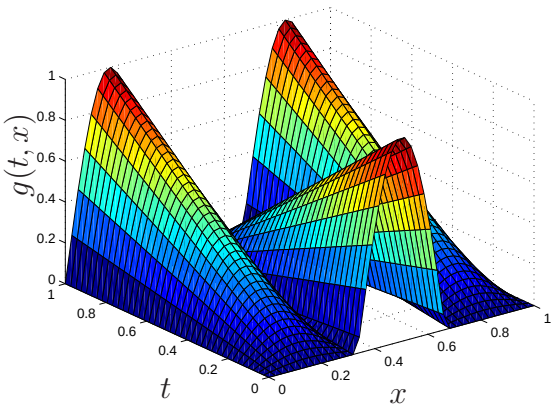

Figure 6.2: Altitude of the obstacles as a function of t and x.

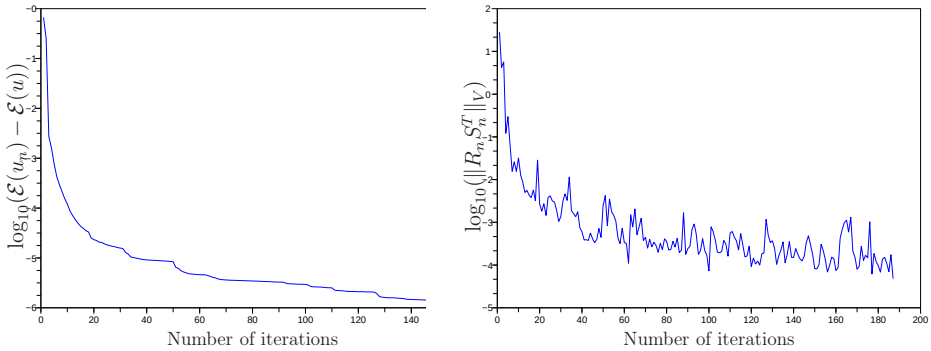

Figure 6.3: Evolution of $\log_{10}(\mathcal{E}(u_n)-\mathcal{E}(u))$ (left) and of $\log_{10}(\|R_n S_n^T\|_V)$ (right) as a function of n.

Fig. 6.3 represents the evolution of $\log_{10}(\mathcal{E}(u_n)-\mathcal{E}(u))$ and of $\log_{10}(\|R_n S_n^T\|_V)$.

We can see that the greedy algorithm captures very quickly the main modes of the solution.

Fig. 6.4 represents the results obtained for the solution $u(t,x)$. Fig. 6.5 and FIg. 6.6 represent $u(t,x)$ and $g(t,x)$ for some special values of T.

As we can observe, the solution does not exactly satisfies the constraint $u(t,x) \geq g(t,x)$. This is due to the fact that we approximate a solution $u_{\widetilde{\rho}}$ of the penalized problem (6.14) for $\widetilde{\rho} = 2500$. This is the main drawback of our method: we do not approximate directly the solution of the initial obstacle problem but the solution of a close regularized problem. Indeed, if we try to further increase the parameter ρ, we face the main drawback of penalization methods, that is the ill-conditioning of the resulting matrices.

6.7 Conclusion

In this article, we presented a greedy algorithm based on variable decomposition aiming at computing the global minimum of a strongly convex energy functional. We proved that, provided that the gradient of the energy is Lipschitz on bounded sets, and that the Hilbert

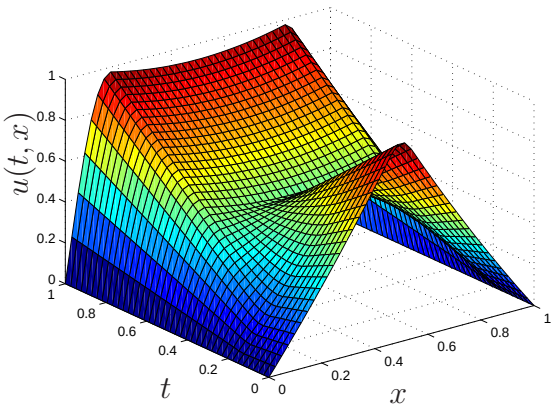

Figure 6.4: Altitude of the rope as a function of t and x.

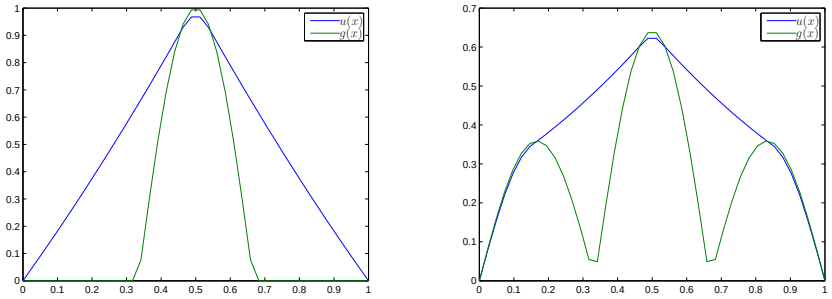

Figure 6.5: Profile of u and g for $T = 0$ (left) and $T = 0.375$ (right).

spaces considered satisfy assumptions $(A1)$ and $(A2)$, then the approximation given by our algorithm strongly converges towards the desired result. One of the main advantage of the algorithm is that it can deal with highly nonlinear problems. We also proved that in finite dimension, this algorithm converges exponentially fast.

We applied this algorithm in the context of uncertainty quantification on obstacle problems. In this frame, we considered regularizations of this kind of problems by penalization methods. Indeed, the obstacle problem can be approximated by a global minimization problem defined on the entire Hilbert space of some strongly convex energy functional where the constraints of the initial problem are replaced by penalization terms in the expression of the functional. Our algorithm gives a good approximation of the solutions of the regularized problem. However, the problem of ill-conditioned matrices, which is inherent to penalization methods, limits the accuracy with which we can approach the solution of the initial obstacle problem.

One way to circumvent this problem is to use augmented Lagrangian methods (see [90, 94, 101]) instead of penalization methods. Indeed, the former algorithms converge towards the true solution of the initial obstacle problems. The adaptation of our algorithm to such methods is work in progress.

Another extension of our work would be to consider other problems than obstacle

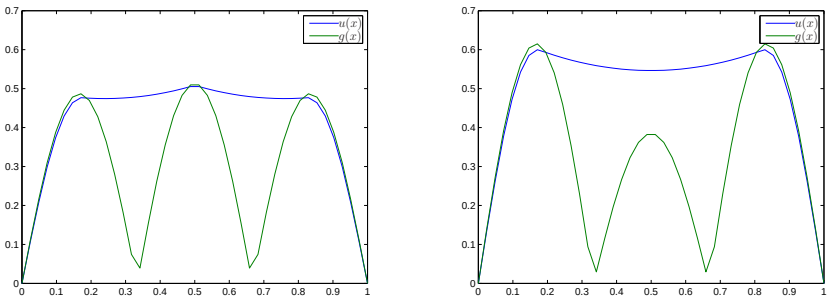

Figure 6.6: Profile of u and g for $T = 0.5$ (left) and $T = 0.625$ (right).

problems. In [157], a similar algorithm based on Proper Generalized Decomposition is used to study uncertainty quantification upon a Burger type equation. We believe that it could be possible to extend our proof of convergence in the case of such hyperbolic systems.

Acknowledgment

We would like to thank the Michelin company for financial support.

Bibliography

[1] Freefem++ finite element software. *http://www.freefem.org/*.

[2] Scilab software. *http://www.scilab.org/*.

[3] R.J. Adler. *The geometry of random fields*. Wiley, Chichester, 1981.

[4] A. Alvino, G. Trombetti, and P.-L. Lions. On optimization problems with prescribed rearrangements. *Nonlinear Analysis*, 13:185–220, 1989.

[5] A. Ammar, B. Mokdad, F. Chinesta, and R. Keunings. A new family of solvers for some classes of multidimensional partial differential equations encountered in kinetic theory modeling of complex fluids. *Journal of Non-Newtonian Fluid Mechanics*, 139:153–176, 2006.

[6] A. Anantharaman and E. Cancès. Existence of minimizers for Kohn-Sham models in quantum chemistry. *Annales de l'Institut Henri Poincaré, Analyse non linéaire*, 26:2425–2455, 2009.

[7] D. Arnold. Differential complexes and numerical stability. *Proceedings of the ICM 2002*, 1:137–157, 2002.

[8] H. Avci and A.T. Gürkanli. Multipliers and tensor products of $l(p, q)$ Lorentz spaces. *Acta Mathematica Scientia*, 27:107–116, 2007.

[9] I. Babuška and J. Osborn. Eigenvalue problems. *Handbook of Numerical Analysis*, 2:641–787, 1991.

[10] I. Babuška and P. Chatzipantelidis. On solving elliptic stochastic partial differential equations. *Computational Methods in Applied Mechanics and Engineering*, 191:4093–4122, 2002.

[11] I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 45:1005–1034, 2007.

[12] A.R. Barron, A. Cohen, W. Dahmen, and R. DeVore. Approximation and learning by greedy algorithms. *Annals of Statistics*, 36:64–94, 2008.

[13] A.D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, 38:3098–3100, 1988.

[14] R.E. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[15] J. Bergh and J. Löfström. *Interpolation spaces*. Springer-Verlag, 1976.

[16] G. Beylkin and M.J. Mohlenkamp. Algorithms for numerical analysis in high dimensions. *SIAM Journal on Scientific Computing*, 26:2133, 2005.

[17] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtasczczyk. Convergence Rates for Greedy Algorithms in Reduced Basis Methods. *SIAM Journal on Mathematical Analysis*, 43:1457–1472, 2011.

[18] X. Blanc, C. Le Bris, and P.-L. Lions. A definition of the ground state energy for systems composed of infinitely many particles. *Communications in Partial Differential Equations*, 28:439–475, 2003.

[19] F. Bloch. Uber die Quantenmechanik der Elektronen in Kristallgittern. *Zeitschrift für Physik*, 52:555–560, 1928.

[20] D. Boffi, F. Brezzi, and L. Gastaldi. On the problem of spurious eigenvalues in the approximation of linear elliptic problems in mixed form. *Mathematics of Computation*, 69:121–140, 1999.

[21] L. Boulton. Non-variational approximation of discrete eigenvalues of self-adjoint operators. *IMA Journal on Numerical Analysis*, 27:102–121, 2007.

[22] L. Boulton and N. Boussaid. Non-variational computation of the eigenstates of Dirac operators with radially symmetric potentials. *LMS Journal of Computation and Mathematics*, 13:10–32, 2010.

[23] L. Boulton, N. Boussaid, and M. Lewin. Generalised Weyl theorems and spectral pollution in the Galerkin method. *Journal of Spectral Theory, in press*, 2012.

[24] L. Boulton and M. Levitin. On the approximation of the eigenvalues of perturbed periodic Schrödinger operators. *Journal of Physics A*, 40:9319–9329, 2007.

[25] S. Boyaval, C. Le Bris, T. Lelièvre, Y. Maday, N.C. Nguyen, and A.T. Patera. Reduced basis techniques for stochastic problems. *Archives of Computational methods in Engineering*, 17:435–454, 2010.

[26] S. Boyaval, C. Le Bris, Y. Maday, N.C. Nguyen, and A.T. Patera. A reduced basis approach for variational problems with stochastic parameters: Application to heat conduction with variable Robin coefficient. *Computer Methods in Applied Mechanics and Engineering*, 198:3187–3206, 2009.

[27] L. Breiman. *Probability*. Classics in Applied Mathematics, 1992.

[28] H. Brézis, R. Benguria, and E.H. Lieb. The Thomas-Fermi-von Weiszäcker theory of atoms and molecules. *Communications in Mathematical Physics*, 79:167–180, 1981.

[29] A. Brezzi and M. Fortin. *Mixed and Hybrid finite element methods*. Springer-Verlag, 1991.

[30] A. Buffa, Y. Maday, A.T. Patera, C. Prud'homme, and G. Turinici. A priori convergence of the greedy algorithm for the parametrized reduced basis. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46:595–603, 2012.

[31] H. Bungartz. An adaptative Poisson solver using hierarchical bases and sparse grids. *Iterative Methods in Linear Algebra*, pages 293–310, 1992.

[32] H. Bungartz and M. Griebel. Sparse grids. *Acta Numerica*, 13:147–269, 2004.

[33] R.E. Caflisch. Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, 7:1–49, 1998.

[34] R.H. Cameron and W.T. Martin. The orthogonal development of non-linear functionals in series of Fourier-Hermite functionals. *Annals of Mathematics*, 48:385–392, 1947.

[35] 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*, 46:341–388, 2012.

[36] E. Cancès, A. Deleurence, and M. Lewin. A new approach to the modeling of local defects in crystals: the reduced Hartree-Fock case. *Communications on Mathematical Physics*, 281:129–177, 2008.

[37] E. Cancès, A. Deleurence, and M. Lewin. Non-perturbative embedding of local defects in crystalline materials. *Journal of Physics: Condensed Matter*, 20:294213, 2008.

[38] E. Cancès and V. Ehrlacher. Local defects are always neutral in the Thomas-Fermi-von Weiszäcker theory of crystals. *Archive for Rational Mechanics and Analysis*, 202:933–973, 2011.

[39] E Cancès, V. Ehrlacher, and T. Lelièvre. Convergence of a greedy algorithm for high-dimensional convex problems. *Mathematical Models and Methods in Applied Sciences*, 21:2433–2467, 2011.

[40] E. Cancès, V. Ehrlacher, and Y. Maday. Periodic Schrödinger operators with local defects and spectral pollution. *Submitted to SIAM Journal on Numerical Analysis*, 2011.

[41] E. Cancès, V. Ehrlacher, and Y. Maday. Non-consistent approximations of self-adjoint eigenproblems: Application to the supercell method. *Submitted to Numerische Mathematik*, 2012.

[42] E. Cancès, C. Le Bris, and Y. Maday. *Méthodes mathématiques en chimie quantique: une introdution*. Springer, 2006.

[43] E. Cancès and M. Lewin. The dielectric permittivity of crystals in the reduced Hartree-Fock approximation. *Archive for Rational Mechanics and Analysis*, 197:139–177, 2010.

[44] E. Cancès and G. Stoltz. A mathematical formulation of the random phase approximation for crystals. *Preprint arXiv 1109.2416*, 2011.

[45] E. Candès, J. Romberg, and T. Tao. Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. *IEEE Transactions on Information Theory*, 52:489–509, 2006.

[46] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Commications on Pure and Applied Mathematics*, 59:1207–1223, 2006.

[47] E. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215, 2005.

[48] E.J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus de l'Acdémie des Sciences de Paris*, 346:589–592, 2008.

[49] I. Catto, C. Le Bris, and P.-L. Lions. Limite thermodynamique pour des modèles de type Thomas-Fermi. *Notes aux Comptes Rendus de l'Académie des Sciences, Série I*, 322:357–364, 1996.

[50] I. Catto, C. Le Bris, and P.-L. Lions. *Mathematical theory of thermodynamic limits: Thomas-Fermi type models*. Oxford University Press, 1998.

[51] I. Catto, C. Le Bris, and P.-L. Lions. Sur la limite thermodynamique pour des modèles de type Hartree et Hartree-Fock. *Notes aux Comptes Rendus de l'Académie des Sciences, Série I*, 327:259–266, 1998.

[52] I. Catto, C. Le Bris, and P.-L. Lions. Recent mathematical results on the quantum modelling of crystals. *Lecture Notes in Chemistry*, 74:95–119, 2000.

[53] I. Catto, C. Le Bris, and P.-L. Lions. On the thermodynamic limit for Hartree-Fock type models. *Annales de l'Institut Henri Poincaré, Analyse non linéaire*, 18:687–760, 2001.

[54] I. Catto, C. Le Bris, and P.-L. Lions. On some periodic Hartree-type models for crystals. *Annales de l'Institut Henri Poincaré, Analyse non linéaire*, 19:143–190, 2002.

[55] P. Chaix and D. Iracane. From quantum electrodynamics to mean-field theory: I. The Bogoliubov-Dirac-Fock formalism. *Journal of Physics B*, 22:3791–3814, 1989.

[56] F. Chatelin. *Spectral Approximation of Linear Operators*. Academic Press, 1983.

[57] A. Chkifa, A. Cohen, R. DeVore, and C. Schwab. Sparse Adaptive Taylor Approximation Algorithms for Parametric and Stochastic Elliptic PDEs. *Submitted*, 2011.

[58] A. Cohen, R. DeVore, and C. Schwab. Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs. *Foundations of Computational Mathematics*, 10:615–646, 2010.

[59] A. Cohen, R. DeVore, and C. Schwab. Analytic Regularity and Polynomial Approximation of Parametric Stochastic Elliptic PDEs. *Analysis and Applications*, 9:11–47, 2011.

[60] M.L. Cohen and T.K. Bergstresser. Band structures and Pseudopotential Form Factors for Fourteen Semiconductors of the Diamond and Zinc-blende Structures. *Physical Review*, 141:789–796, 1966.

[61] L. Conlon. *Differentiable Manifolds: A First Course*. Birkhauser, 1993.

[62] C. Cotar, G. Friesecke, and C. Kluppelberg. Density functional theory and optimal transportation with Coulomb cost. *arXiv: 1104.0603*, 2011.

[63] W. Dai and O. Milenkovich. Subspace pursuit for compressed sensing: Closing the gap between performance and complexity. *IEEE Journal of Selected Topics in Signal Processing*, 4:310–316, 2010.

[64] M. Dauge and M. Suri. Numerical approximation of the spectra of non-compact operators arising in buckling problems. *Journal of Numerical Mathematics*, 10:193–219, 2002.

[65] E. Davies and M. Plum. Spectral pollution. *IMA Journal on Numerical Analysis*, 24:417–438, 2004.

[66] E.B. Davis and M. Plum. Spectral pollution. *IMA Journal on Numerical Analysis*, 24:417–438, 2004.

[67] L. De Lathauwer, B. De Moor, and J. Vandewalle. A Multilinear Singular Value Decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21:1253–1278, 2000.

[68] L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1 and rank-$(r_1, r_2, \cdots, r_n)$ Approximation and applications of Higher-Order Tensors. *SIAM Journal on Matrix Analysis and Applications*, 21:1324–1342, 2000.

[69] V. de Silva and L.H. Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications: Special Issue on Tensor Decompositions and Applications*, 30:1084–1127, 2008.

[70] A. Deleurence. *Modélisation mathématique et simulation numérique de la structure électronique de cristaux en présence de défauts ponctuels.* thèse de l'Ecole Nationale des Ponts et Chaussées, 2008.

[71] J. Descloux. Essential numerical range of an operator with respect to a coercive form and the approximation of its spectrum by the Galerkin method. *SIAM Journal on Numerical Analysis*, 18:1128–1133, 1981.

[72] J. Descloux, N. Nassif, and J. Rappaz. On spectral approximation. Part 1: The problem of convergence. *RAIRO Analyse numérique*, 12:97–112, 1978.

[73] J. Descloux, N. Nassif, and J. Rappaz. On spectral approximation. Part 2: Error estimates for the Galerkin method. *RAIRO Analyse numérique*, 12:113–119, 1978.

[74] R. DeVore, R. Howard, and C.A. Micchelli. Optimal non-linear approximation. *Manuscripta Mathematica*, 63:469–478, 1989.

[75] R.A. DeVore and V.N. Temlyakov. Some remarks on greedy algorithms. *Advances in Computational Mathematics*, 5:173–187, 1996.

[76] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:5406–5425, 2006.

[77] D.L. Donoho, Y. Tsaig, and J.-L. Starck. Sparse solution of undetermined linear equations by stagewise orthogonal matching pursuit. *Technical Report*, 2006.

[78] A. Doostan, R. Ghanem, and J. Red-Horse. Stochastic model reductions for chaos representations. *Computer Methods in Applied Mechanics and Engineering*, 196:3951–3966, 2007.

[79] A. Doostan and H. Owhadi. A non-adapted sparse approximation of PDEs with stochastic inputs. *Journal on Computational Physics*, 230:3015–3034, 2011.

[80] R.M. Dreizler and E.K.U. Gross. *Density Functional Theory.* Springer Berlin, 1990.

[81] W. E, W. Ren, and E. Van den Eijnden. A string method for the study of rare events. *Physical Review B*, 66:052301, 2002.

[82] L. Eldén and B. Savas. A Newton-Grassmann Method for Computing the Best Multilinear Rank-(r_1, r_2, r_3) Approximation of a Tensor. *SIAM Journal on Matrix Analysis and Applications*, 31:248–271, 2009.

[83] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements.* Volume 159 of Applied Mathematical Series, Springer, 2004.

[84] C. Fefferman. The thermodynamic limit for a crystal. *Communications in Mathematical Physics*, 98:289–311, 1985.

[85] E. Fermi. Un Metodo Statistico per la Determinazione di alcune Prioprietà dell'Atomo. *Rendiconti Accademia Nazionale Lincei*, 6:602–607, 1927.

[86] L. Figueroa and E. Suli. Greedy Approximation of High-Dimensional Ornstein-Uhlenbeck Operators with Unbounded Drift. *arXiv:1103.0726*, 2011.

[87] S. Fliss and P. Joly. Exact boundary conditions for time-harmonic wave propagation in locally perturbed periodic media. *Applied Numerical Mathematics*, 59:2155–2178, 2009.

[88] S. Fliss and P. Joly. Wave propagation in locally perturbed periodic media (case with absorption): Numerical aspects. *Journal of Computational Physics*, 231:1244–1271, 2012.

[89] G. Floquet. Sur les équations différentielles linéaires à coefficients périodiques. *Annales Scientifiques de l'Ecole Normale Supérieure*, 3:47–89, 1883.

[90] M. Fortin and R. Glowinski. *Méthodes de Lagrangien augmenté - Application à la résolution numérique de problèmes aux limites.* Dunod, 1982.

[91] G. Friesecke. The Multiconfiguration Equations for Atoms and Molecules: Charge Quantization and Existence of Solutions. *Archive for Rational Mechanics and Analysis*, 169:35–71, 2003.

[92] R. Ghanem and P. Spanos. *Stochastic finite elements: a spectral approach.* Springer, Berlin, 1991.

[93] M. Ghimenti and M. Lewin. Properties of the periodic Hartree-Fock minimizer. *Calculus of Variations and Partial Differential Equations*, 35:39–56, 2009.

[94] R. Glowinski and P. Le Tallec. *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics.* Society for Industrial and Applied Mathematics, 1989.

[95] R. Glowinski, J.-L. Lions, and R. Trémolières. *Analyse numérique des inéquations variationnelles - Théorie générale et premières applications.* Dunod, 1976.

[96] L. Grasedyck. Hierarchical Singular Value Decomposition of Tensors. *SIAM Journal on Matrix Analysis and Applications*, 31:2029–2054, 2010.

[97] U. Grenander and G. Szego. *Toeplitz forms and their applications.* University of California Press, Berkeley-Los Angeles, 1958.

[98] M. Griebel and S. Knapek. Optimized tensor-product approximation spaces. *Constructive Approximation*, 16:525–540, 2000.

[99] M. Griesemer and F. Hantsch. Unique Solutions to Hartree-Fock Equations for Closed Shell Atoms. *to appear in Archive for Rational Mechanics and Analysis*, 2011.

[100] D.J. Griffiths. *Introduction to Quantum Mechanics (2nd edition)*. Addison Wesley, 2004.

[101] C. Grossmann, H.-G. Roos, and M. Stynes. *Numerical Treatment of Partial Differential Equations*. Springer, 2007.

[102] D. Guilbarg and N.S. Trudinger. *Elliptic partial differential equations of second order*. Springer Berlin, second edition, 1983.

[103] W. Hackbusch. *Tensor Spaces and Numerical Tensor Calculus*. Springer, 2012.

[104] W. Hackbusch and S. Kuhn. A new scheme for the tensor representation. *Journal of Fourier Analysis and Applications*, 15:706–722, 2009.

[105] C. Hainzl, M. Lewin, and E. Séré. Existence of atoms and molecules in the mean-field approximation of no-photon quantum electrodynamics. *Archive for Rational Mechanics and Analysis*, 192:453–499, 2009.

[106] C. Hainzl, M. Lewin, E. Séré, and J.P. Solovej. A minimization method for relativistic electrons in a mean-field approximation of quantum electrodynamics. *Physical Review A*, 76:052104, 2007.

[107] C. Hainzl, M. Lewin, and J.P. Solovej. The mean-field approximation in electrodynamics: the no-photon case. *Communications on Pure and Applied Mathematics*, 60:546–596, 2007.

[108] C. Hainzl, M. Lewin, and J.P. Solovej. The thermodynamic limit of quantum Coulomb systems. Part I: General Theory. *Advances in Mathematics*, 221:454–487, 2009.

[109] C. Hainzl, M. Lewin, and J.P. Solovej. The thermodynamic limit of quantum Coulomb systems. Part II: Applications. *Advances in Mathematics*, 221:488–546, 2009.

[110] A.C. Hansen. On the approximation of spectra of linear operators on Hilbert spaces. *Journal on Functional Analysis*, 254:2092–2126, 2008.

[111] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Physical Reviews*, 136:B864–B871, 1964.

[112] S. Holtz, T. Rohwedder, and R. Schneider. The Alternating Linear Scheme for Tensor Optimization in the TT format. *To appear in SIAM Journal on Scientific Computing*, 2011.

[113] S. Holtz, T. Rohwedder, and R. Schneider. On manifolds of tensors of fixed TT-rank. *to appear in Numerische Mathematik*, 2011.

[114] R.O. Jones and O. Gunnarson. The density functional formalism, its application and prospects. *Review of Modern Physics*, 61:689–746, 1989.

[115] K. Karhunen. Zur Spektraltheorie stochasticher Prozesse. *Annales Academiae Scientiarum Fennicae*, 34, 1946.

[116] C. Kittel. *Quantum theory of solids.* John Wiley & Sons, 2nd edition, 1987.

[117] M. Kleiber and T.D. Hien. *The stochastic finite element method. Basic perturbation technique and computer implementation.* Wiley, Chichester, 1992.

[118] O. Koch and C. Lubich. Dynamical Low-rank Approximation of Tensors. *SIAM Journal on Matrix Analysis and Applications*, 31:2360, 2010.

[119] W. Kohn and L.J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review A*, 140:1133–1138, 1965.

[120] T.G. Kolda and B.W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 51:455–500, 2009.

[121] A.S. Kompaneets and E.S. Pavlvoskii. *Soviet Physics - Journal of Experimental and Theoretical Physics Letters*, 4:328–336, 1957.

[122] E. Korotyaev. Lattice Dislocations in a 1-Dimensional Model. *Communications in Mathematical Physics*, 213:471–489, 2000.

[123] P. Ladevèze. *Nonlinear computational structural mechanics: new approaches and non-incremental methods of calculation.* Springer, Berlin, 1999.

[124] S. Lahbabi. *in preparation.* Thèse de l'Université de Cergy-Pontoise, 2013.

[125] D.C. Langreth and J.P. Perdew. Theory of nonuniform electronic systems. I. Analysis of the gradient approximation and a generalization that works. *Physical Review B*, 21:5469–5493, 1980.

[126] B. Langwallner, C. Ortner, and E. Suli. Existence and convergence results for the Galerkin approximation of an electronic density functional. *OxMOS Technical Report*, 21, 2009.

[127] C. Le Bris. *Quelques problèmes mathématiques en chimie quantique moléculaire.* Thèse de l'Ecole Polytechnique, 1993.

[128] C. Le Bris. Some results on the Thomas-Fermi-Dirac-von Weiszäcker model. *Differential and Integral Equations*, 6:337–353, 1993.

[129] C. Le Bris. A general approach for multiconfiguration methods in quantum molecular chemsitry. *Annales de l'Institut Henri Poincaré*, 11:441–484, 1994.

[130] C. Le Bris, T. Lelièvre, and Y. Maday. Results and Questions on a Nonlinear Approximation Approach for Solving High-dimensional Partial Differential Equations. *Construsctive Approximation*, 30:621–651, 2009.

[131] M. Levitin and E. Shargorodsky. Spectral pollution and second order relative spectra for self-adjoint operators. *IMA Journal on Numerical Analysis*, 24:393–416, 2004.

[132] A. Levy and J. Rubinstein. Some properties of smoothed principal component analysis for functional data. *Journal of the Optical Society of America*, 16:28–35, 1999.

[133] M. Lewin. Solutions of the multiconfigurational equations in quantum chemistry. *Archives for Rational Mechanics and Analysis*, 171:83–114, 2004.

[134] M. Lewin and E. Séré. Spectral pollution and how to avoid it (with applications to Dirac and periodic Schrödinger operators). *Proceedings of the London Mathematical Society*, 100:864–900, 2010.

[135] E.H. Lieb. Thomas-Fermi and related theories of atoms and molecules. *Reviews of Modern Physics*, 53:603–641, 1981.

[136] E.H. Lieb. Variational Principle for Many-Fermion Systems. *Physical Review Letters*, 46:457–459, 1981.

[137] E.H. Lieb. Density Functional for Coulomb systems. *International Journal of Quantum Chemistry*, 24:143–277, 1983.

[138] E.H. Lieb and M. Loss. *Analysis*. American Mathematical Society, second edition, 2001.

[139] E.H. Lieb and B. Simon. The Hartree-Fock theory for Coulomb systems. *Communications in Mathematical Physics*, 53:185–194, 1977.

[140] E.H. Lieb and B. Simon. The Thomas-Fermi theory of atoms, molecules and solids. *Advances in Mathematics*, 23:22–116, 1977.

[141] P.-L. Lions. The concentration-compactness principle in the calculus of variations. The locally compact case, part 1 and 2. *Annales de l'Institut Henri Poincaré*, 1:109–145 and 223–283, 1984.

[142] P.-L. Lions. Solutions of Hartree-Fock equations for Coulomb systems. *Communications in Mathematical Physics*, 109:33–97, 1987.

[143] M. Loève. Fonctions aléatoires du second ordre. *Comptes Rendus de l'Académie des Sciences de Paris*, 220, 1945.

[144] C. Lubich. From Quantum to Classical Molecular Dynamics: Reduced Methods and Numerical Analysis. *Zurich Lectures in advanced mathematics, EMS*, 2008.

[145] Y. Maday, N.C. Nguyen, A.T. Patera, and G.S.H. Pau. A General Multipurpose Interpolation Procedure: The Magic Points. *Communications on Pure and Applied Analysis*, 8:383–404, 2009.

[146] G.D. Mahan. *Many Particle Physics (Physics of Solids and Liquids)*. Springer, 2000.

[147] M. Mantoiu and R. Purice. A priori decay for eigenfunctions of perturbed periodic Schrödinger operators. *Annales de l'Institut Henri Poincaré*, 2:525–521, 2001.

[148] M.D. McKay, R.J. Beckman, and W.J. Conover. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, 21:239–245, 1979.

[149] W.H. Mills. Optimal error estimates for the finite element spectral approximation of noncompact operators. *SIAM Journal on Numerical Analysis*, 16:704–718, 1979.

[150] W.H. Mills. The resolvent stability condition for spectra convergence with application to the finite element approximation of noncompact operators. *SIAM Journal on Numerical Analysis*, 16:695–703, 1979.

[151] D. Needell and J.A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26:301–321, 2009.

[152] D. Needell and R. Vershynin. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *Preprint arXiv*, 2007.

[153] R.B. Nelsen. *An Introduction to Copulas*. Springer, 1999.

[154] A. Nouy. Recent developments in spectral stochastic methods for the numerical solution of stochastic partial differential equations. *Archives of Computational Methods in Engineering*, 16:251–285, 2009.

[155] A. Nouy. A priori tensor approximations for the numerical solution of high dimensional problems: alternative definitions. *Preprint*, 2012.

[156] A. Nouy and A. Falco. Proper Generalized Decomposition for Nonlinear Convex Problems in Tensor Banach Spaces. *Submitted to Numerische Mathematik*, 2011.

[157] A. Nouy and O. Le Maitre. Generalized Spectral Decomposition Method for Stochastic Non Linear Problems. *Journal of Computational Physics*, 228:202–235, 2009.

[158] I.V. Oseledets. Tensor-Train Decomposition. *SIAM Journal on Scientific Computing*, 33:2295–2317, 2011.

[159] J.P. Perdew, K. Burke, and M. Ernzerhof. Generalized Gradient Approximation Made Simple. *Physical Review Letters*, 77:3865–3868, 1996.

[160] J.P. Perdew and Y. Wang. Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation. *Physical Review B*, 33:8800–8802, 1986.

[161] C. Pisani. Quantum-mechanical treatment of the energetics of local defects in crystals: A few answers and many open questions. *Phase Transitions*, 52:123–136, 1994.

[162] J. Rappaz, J. Sanchez Hubert, J. Sanchez Palencia, and D. Vasiliev. On spectral pollution in the finite element approximation of thin elastic 'membrane' shells. *Journal of Numerical Mathematics*, 75:473–500, 1997.

[163] M. Reed and B. Simon. *Methods of Modern Mathematical Physics IV: Analysis of Operators*. Academic Press, 1978.

[164] R. Schneider, T. Rohwedder, and O. Legeza. Tensor methods in quantum chemistry. *Submitted to Archive for Rational Mechanics and Analysis*, 2012.

[165] L. Schwartz. *Théorie des distributions*. Hermann, 1966.

[166] V. Shabaev, I.I. Tupitsyn, V.A. Yerokhin, G. Plunien, and G. Soff. Dual kinetic balance approach to basis-set expansions for the Dirac equation. *Physical Review Letters*, 93:130405, 2004.

[167] E. Shargorodsky. Geometry of higher order relative spectra and projection methods. *Journal of Operator Theory*, 44:43–62, 2000.

[168] B. Simon. Schrödinger semigroups. *Bulletin of the American Mathematical Society*, 7:447–526, 1982.

[169] S. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Soviet Mathematics Doklady*, 3:240–243, 1963.

[170] J.P. Solovej. Universality in the Thomas-Fermi-von Weizsäcker model of atoms and molecules. *Communications in Mathematical Physics*, 129:561–598, 1990.

[171] J.P. Solovej. Proof of the ionization conjecture in a reduced Hartree-Fock model. *Inventiones Mathematicae*, 104:291–311, 1991.

[172] S. Soussi. Convergence of the supercell method for defect modes calculations in photonic crystals. *SIAM Journal on Numerical Analysis*, 43:1175–1201, 2005.

[173] A.M. Stoneham. *Theory of defects in solids: electronic structure of defects in insulators and semiconductors*. Oxford University Press, 2001.

[174] V.N. Temlyakov. Greedy Approximation. *Acta Numerica*, 17:235–409, 2008.

[175] L.H. Thomas. The calculation of atomic fields. *Mathematical Proceedings of the Cambridge Philosophical Society*, 23:542–548, 1927.

[176] J. Tropp and A. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53:4655–4666, 2007.

[177] N.S. Trudinger. Linear elliptic operators with measurable coefficients. *Annali della Scuola Normale Superiore di Pisa, Classe di Scienze*, 27:265–308, 1973.

[178] J. Tryoen. *Adaptive stochastic Galerkin methods for parametric uncertainty propagation in hyperbolic systems*. Thèse de l'Université Paris-Est, 2012.

[179] T. von Petersdorff and C. Schwab. Numerical solution of parabolic equations in high dimensions. *M2AN Mathematical Modelling and Numerical Analysis*, 38:93–127, 2004.

[180] C.G. Webster, F. Nobile, and R. Tempone. A sparse grid collocation method for partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 46:2309–2345, 2007.

[181] C.F. Weiszäcker. Zur Theorie der Kernmassen. *Zeitschrif für Physik*, 96:431–458, 1935.

[182] N. Wiener. The homogeneous chaos. *American Journal of Mathematics*, 60:897–936, 1938.

[183] D.B. Xiu and G.E. Karniadakis. The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24:619–644, 2002.

[184] C. Zenger. Sparse grids. *Parallel Algorithms for Partial Differential Equations, ed. W. Hackbusch, Vieweg*, pages 241–251.

[185] V. Zheludev. The spectrum of Schrödinger operator, with a periodic potential, defined on the half-axis. *Works of Department of Mathematical Analysis of Kaliningrad State University (in Russian)*, pages 18–37, 1969.

[186] A. Zhou. Finite dimensional approximations for the electronic ground state solution of a molecular system. *Mathematical Methods in the Applied Sciences*, 30:429–447, 1990.